Debiasing Sketched Ridge Regression: A Functional Estimation Perspective

Yucong Liu
School of Mathematics
Georgia Institute of Technology
Atlanta, GA 30332
yucongliu@gatech.edu

Florian Schäfer
Courant Institute of Mathematical Sciences
New York University
New York, NY 10012
florian.schaefer@nyu.edu

Abstract

We study sketched ridge regression through the lens of functional estimation. The sketched solution is a plug-in estimator of a nonlinear function of a second-moment/covariance matrix. Within this framework, we analyze two classical bias-reduction principles—iterative bootstrap and linear aggregation across multiple estimators. We show that these resampling-based procedures cancel low-order bias terms and yield higher-order bias decay in the sketch size without increasing the order of the computational complexity. Concretely, let $\mathbf{A} \in \mathbb{R}^{n\times r}$ have full column rank r and let the sketch size be s. For Gaussian sketching, a k-step iterative bootstrap estimator achieves bias of order $(\sqrt{r/s})^{k+2}$. For linear aggregation over m plug-in estimators with different sketch sizes, we obtain bias of order $(\sqrt{r/s})^{m+1}$ for Gaussian sketching and corresponding result for common discrete sketches, including uniform and leverage score sampling.

1 Introduction

Solving linear systems $\mathbf{Ax} = \mathbf{b}$ and related linear problems, including regression—particularly ridge regression—is fundamental to a wide range of scientific and engineering applications. Randomized Numerical Linear Algebra (RNLA) [Mahoney, 2011, Woodruff, 2014, Drineas and Mahoney, 2016, Kannan and Vempala, 2017, Martinsson and Tropp, 2020, Murray et al., 2023] provides an alternative approach via multiplication with a random matrix $\mathbf{S}$. Such methods enable randomized dimension reduction and consequently reduce computational cost. At the same time, this randomness incurs additional error and bias as the price paid for computational efficiency.

More precisely, a standard approach in sketching is the following: for a tall matrix $\mathbf{A} \in \mathbb{R}^{n\times r}$, one reduces the computational cost of ridge regression by replacing the original problem $\min_{\mathbf{x}\in\mathbb{R}^r} \|\mathbf{Ax}-\mathbf{b}\|_2^2 + \gamma\|\mathbf{x}\|_2^2$ with the sketched problem $\min_{\mathbf{x}\in\mathbb{R}^r} \|\mathbf{SAx}-\mathbf{Sb}\|_2^2 + \gamma\|\mathbf{x}\|_2^2$. This random sketch matrix $\mathbf{S} \in \mathbb{R}^{s\times n}$ projects the data into a lower-dimensional space and reduces the problem size from n to s. While this compression can substantially improve computational efficiency, it also introduces approximation error. Controlling this error is therefore essential for ensuring that sketching methods are both efficient and robust.

This work studies the bias of sketched solution of ridge regression via a functional estimation framework. In particular, the sketched solution can be viewed as a plug-in estimator of a certain function of the covariance matrix. We provide the bias bound for the sketched solution and show the bias reduction phenomenon by iterative bootstrap method and linear aggregation method. Both methods cancel out low-order bias term and achieve a high-order bias with respect to the sketch size. Moreover, with a fixed iteration number, both methods have same computational complexity as one sketched solution.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1.1 Motivation

Given the optimal solution $\mathbf{x}^*$ and a sketched estimator $\hat{\mathbf{x}}$ of a ridge regression problem, the bias $\mathbb{E}\,\hat{\mathbf{x}} - \mathbf{x}^*$ is a persistent distortion in the estimator. While variance measures the spread of $\hat{\mathbf{x}}$ around its mean, the bias captures whether this mean is systematically shifted away from the true solution $\mathbf{x}^*$. Therefore, we aim to understand the bias with respect to sketch size.

Existing theoretical results provide only high-probability error bounds, valid when the sketch size is sufficiently large as a function of the probability tolerance δ and the error tolerance ε. More precisely, Wang et al. [2018] showed that

$$\left(\|\mathbf{A}\hat{\mathbf{x}} - \mathbf{b}\|_2^2 + \gamma\|\hat{\mathbf{x}}\|_2^2\right) - \left(\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2^2 + \gamma\|\mathbf{x}^*\|_2^2\right) \leq \varepsilon C\left(\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2^2 + \gamma\|\mathbf{x}^*\|_2^2\right),$$

with probability at least $1 - \delta$, provided that the sketch size exceeds $s(\varepsilon, \delta)$. Here, C is a constant depending on the matrix $\mathbf{A}$ and the parameter γ. However, this requirement on the sketch size s makes it technically difficult to study the bias $\mathbb{E}\,\hat{\mathbf{x}} - \mathbf{x}^*$ directly as a function of s. In particular, such high-probability bounds do not yield an explicit bias decay rate of the form $\|\,\mathbb{E}\,\hat{\mathbf{x}} - \mathbf{x}^*\| = O(s^{-\alpha})$ for some $\alpha > 0$.

To overcome this technical difficulty, we develop an alternative perspective on the sketched solution by viewing it as a plug-in estimator for a functional estimation problem. Within this framework, we establish a bias decay rate of order $O(s^{-1})$. Moreover, both the iterative bootstrap and the linear aggregation method achieve faster bias decay rates while maintaining the same time complexity.

Another advantage of these two resampling-based iterative methods is that they offer an alternative mechanism for bias reduction through additional iterations. Although increasing the sketch size can also reduce the bias, doing so requires more memory and may be impractical in large-scale or distributed settings. In particular, under distributed computation, the memory available at the central server is often limited, so it may not be feasible to construct or process an overly large sketch in a single step. In contrast, iterative resampling allows us to work with sketches of moderate size while progressively reducing the bias across iterations. Therefore, repeating computations with a reasonable sketch size provides a practical and memory-efficient strategy for improving statistical accuracy.

Moreover, the matrix $(\mathbf{A}^\intercal\mathbf{A})^{-1}$ plays a central role in ordinary least squares and its statistical analysis. When it is approximated by the sketch-based quantity $[(\mathbf{SA})^\intercal(\mathbf{SA})]^{-1}$, one encounters an inversion bias, since in general $\mathbb{E}\left[((\mathbf{SA})^\intercal(\mathbf{SA}))^{-1}\right] \neq (\mathbf{A}^\intercal\mathbf{A})^{-1}$. Our work focuses on the ridge regression setting and provides an analogous perspective for the regularized inversion $(\mathbf{A}^\intercal\mathbf{A} + \gamma\mathbf{I})^{-1}$.

1.2 Related work

Wang et al. [2018] provided a comprehensive study of sketched ridge regression, proposing a model averaging approach and establishing several statistical guarantees. Their method averages multiple sketched solutions constructed with the same sketch size. While this averaging strategy can effectively reduce variance, it does not affect the bias and therefore does not correct the systematic distortion introduced by sketching. Dobriban and Liu [2019], Liu and Dobriban [2020] analyzed the asymptotic properties of sketched estimators. Conditional central limit theorem has been studied by Ahfock et al. [2021]. bootstrap-based methods for error estimation in randomized matrix multiplication and sketched SVD were further developed in Lopes et al. [2019, 2020]. Bias has been studied from related perspectives. Derezinski and Warmuth [2017], Dereziński et al. [2022] developed unbiased estimation methods for linear regression under volume sampling and random-design settings, while Derezinski et al. [2021], Niu et al. [2025] investigated the inversion bias of $(\mathbf{A}^\intercal\mathbf{A})^{-1}$ via a (ε, δ)-unbiased estimator.

Functional estimation is a statistical problem with a long history. The question is how to estimate $f(\theta)$ for some function f and parameter θ of a high-dimensional or even infinite-dimensional model [Donoho and Liu, 1991, Birgé and Massart, 1995, Laurent, 1996, Lepski et al., 1999, Robins et al., 2008]. Recently, Jiao and Han [2020] studied the estimation of one dimensional smooth function $f(\theta)$ based on binomial observations. They summarized three different methods: bootstrap, linear aggregation via several estimators and using Taylor expansion. Bootstrap, introduced by Efron [1979], then generalized by Hall and Martin [1988], is a classic resample technique, aiming to iteratively reduce bias. The linear aggregation method is a general version of Jackknife. The original

Jackknife was initially introduced by Quenouille [1949, 1956] and Tukey [1958]. This method uses delete-one-sample estimator, inspiring using several different estimators to cancel bias.

Our work is inspired by recent developments about estimation for functionals of covariance operators. This line of research originated with a sharp characterization of covariance operator estimation in high-dimensional and infinite-dimensional settings [Koltchinskii and Lounici, 2017]. These techniques were subsequently extended to the estimation of bilinear forms of an unknown covariance operator [Koltchinskii and Lounici, 2016] and linear functionals of its eigenvectors [Koltchinskii et al., 2020]. More recently, the estimation of real-valued Hölder-smooth functionals of covariance operators has been systematically studied [Koltchinskii, 2018, 2021, Koltchinskii and Zhilova, 2021, Koltchinskii, 2025]. In addition, Koltchinskii [2026] investigated the trace functionals of covariance operators, an important class of orthogonally invariant functionals.

1.3 Notation

Matrices and vectors. We denote the cone of $r \times r$ symmetric positive semi-definite matrices as $\mathbb{S}_+^r$ and the space of positive definite matrices as $\mathbb{S}_{++}^r$. For a matrix $\mathbf{X} \in \mathbb{S}_+^r$, $\|\mathbf{X}\|$ denotes its operator norm. For a vector $\mathbf{v} = (v_1, v_2, \cdots, v_n) \in \mathbb{R}^n$, we use $\|\mathbf{v}\|_2 = (\sum_{i=1}^n v_i^2)^{1/2}$ for the Euclidean norm and $\|\mathbf{v}\|_\infty = \max_{1\le i\le n} |v_i|$ for the infinity norm. We write $\mathbf{e}_i$ for the i-th standard basis vector with the i-th entry equal to 1 and all others equal to 0. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, $\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{x}^\intercal \mathbf{y}$. For $\mathbf{A}, \mathbf{B} \in \mathbb{S}_+^r$, $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{tr}(\mathbf{A}^\intercal \mathbf{B})$.

Asymptotic notation. For nonnegative A, B, we write $A \lesssim B$ if $A \le CB$ for a constant $C > 0$. If the constant C depends on a parameter m, we write $A \lesssim_m B$. We write $A \gtrsim B$ if $B \lesssim A$, and $A \asymp B$ if both $A \lesssim B$ and $A \gtrsim B$.

Orlicz norms. Let $\psi : \mathbb{R} \to \mathbb{R}_+$ be convex with $\psi(0) = 0$ and increasing on $\mathbb{R}_+$. For a random variable ξ, we define its Orlicz norm with respect to ψ as

$$\|\xi\|_\psi = \inf\left\{c > 0 : \mathbb{E}\,\psi\left(\frac{|\xi|}{c}\right) \le 1\right\}.$$

In particular, if $\psi(u) = |u|^p$ for $p \ge 1$, then $\|\xi\|_\psi = \|\xi\|_{L_p}$. Denote $\psi_\alpha(u) = e^{u^\alpha} - 1$, $\alpha \ge 1$. Then the ψ_α describes various type of decay of the tails of random variables. For example, $\|\xi\|_{\psi_2}$ is the sub-Gaussian norm. If ξ is a random vector, $\|\xi\|_{\psi_2} := \sup_{u:\|u\|_2=1} \|\langle \xi, u\rangle\|_{\psi_2}$

1.4 Outline and contributions

1. We introduce the sketched ridge regression in Section 2. We provide a new insight of sketched ridge regression by formulating it as a functional estimation problem in Section 2.2. In particular, both the classic and Hessian sketch solutions can be viewed as plug-in estimators of nonlinear functions of a second-moment (covariance) matrix. Under this framework, we show a bias bound in Section 2.3.
2. We study two principled debiasing strategies: iterative bootstrap in Section 3 and linear aggregation method in Section 4. These procedures cancel low-order bias terms. At the same time, the overall computational complexity remains of the same order as a single sketched solve. We show that, in the Gaussian case, a k-step bootstrap-corrected estimator achieves bias of order $(\sqrt{r/s})^{k+2}$. For linear aggregation over m plug-in estimators, we obtain bias of order $(\sqrt{r/s})^{m+1}$ even beyond the Gaussian setting.
3. We complement the theoretical analysis with experiments on three real-world datasets, evaluating bias reduction, relative objective gaps, and runtime across sketching schemes. Experimental results, implementation details, and proofs are provided in the appendix.

2 Sketching as a Functional Estimation Problem

2.1 Sketched Ridge Regression

Consider a ridge regression problem, i.e. a least square problem with ℓ_2 regularization term:

$$\min_{\mathbf{x}\in\mathbb{R}^r} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \gamma\|\mathbf{x}\|_2^2, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n\times r}$ with full column rank and $\mathbf{b} \in \mathbb{R}^n$, $\gamma > 0$.

The optimal solution of the ridge regression problem (1) is given by

$$\mathbf{x}^* = (\mathbf{A}^\intercal\mathbf{A} + \gamma\mathbf{I})^{-1}\mathbf{A}^\intercal\mathbf{b} \tag{2}$$

Given a sketch matrix $\mathbf{S} \in \mathbb{R}^{s\times n}$, the corresponding sketching problem is

$$\min_{\mathbf{x}\in\mathbb{R}^r} \|\mathbf{SAx} - \mathbf{Sb}\|_2^2 + \gamma\|\mathbf{x}\|_2^2, \tag{3}$$

with classic sketch solution

$$\hat{\mathbf{x}}_c = (\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{SA} + \gamma\mathbf{I})^{-1}\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{Sb}. \tag{4}$$

There is an alternative sketch solution called Hessian sketch solution, defined as

$$\hat{\mathbf{x}}_h = (\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{SA} + \gamma\mathbf{I})^{-1}\mathbf{A}^\intercal\mathbf{b}. \tag{5}$$

Sketch solutions achieve faster computation with time complexity $O(sr^2) + T_s(\mathbf{A}) + T_s(\mathbf{b})$ for classic sketch solution (4) and $O(sr^2 + nr) + T_s(\mathbf{A})$ for Hessian sketch solution (5), compared with $O(nr^2)$ for the optimal solution (2). Here $T_s(\mathbf{A})$ denotes the time complexity to form $\mathbf{SA}$ and $T_s(\mathbf{b})$ denotes the time complexity to form $\mathbf{Sb}$ as in Equation (3). One main difference between the classic solution (4) and the Hessian sketch solution (5) is that the Hessian sketch only sketches the Hessian matrix, with no effect on the target vector $\mathbf{b}$. This inspires an iterative Hessian sketch method to iteratively decrease the residual [Pilanci and Wainwright, 2016, Wang et al., 2017].

2.2 Sketched solution is a plug-in estimator

We now re-formulate the sketch solution as an estimation problem.

Consider the compact SVD decomposition of $\mathbf{A} = \mathbf{U\Lambda V}^\intercal$ with $\mathbf{U} \in \mathbb{R}^{n\times r}$, $\mathbf{\Lambda}, \mathbf{V} \in \mathbb{R}^{r\times r}$. Denote the singular value of $\mathbf{A}$ as $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_r > 0$. To construct a random sketch matrix $\mathbf{S} \in \mathbb{R}^{s\times n}$, we view its rows $\mathbf{s}_1^\intercal, \ldots, \mathbf{s}_s^\intercal \in \mathbb{R}^n$ as i.i.d. samples from a distribution P on $\mathbb{R}^n$. Define the compressed sample $\mathbf{z}_i = \sqrt{s}\mathbf{U}^\intercal\mathbf{s}_i \in \mathbb{R}^r$ for each index i. One can directly get

$$\mathbf{U}^\intercal\mathbf{S}^\intercal\mathbf{SU} = \frac{1}{s}\sum_{i=1}^{s}\mathbf{z}_i\mathbf{z}_i^\intercal.$$

We emphasize that the compressed sample provides a dimension reduction from n to r, which captures the intrinsic dimension. The sample $\mathbf{z}_i$ is only used for analytical purpose, we do not explicitly compute it to get sketch solution or use it in the iterative debiasing steps.

Equivalently, we can view sketching as generating i.i.d. "compressed" random variables $\mathbf{z}_i \in \mathbb{R}^r$ from a random vector $\mathbf{z}$. We call $\mathbf{z}$ an isotropic random vector if its second-moment matrix satisfies

$$\mathbf{\Sigma} := \mathbb{E}\,\mathbf{z}\mathbf{z}^\intercal = \mathbf{I}.$$

We say $\mathbf{S}$ is an isotropic sketch if it satisfies the above condition. An estimator of $\mathbf{\Sigma}$ is the sample second-moment matrix

$$\hat{\mathbf{\Sigma}}_s := \frac{1}{s}\sum_{i=1}^{s}\mathbf{z}_i\mathbf{z}_i^\intercal = \mathbf{U}^\intercal\mathbf{S}^\intercal\mathbf{SU}.$$

If $\mathbb{E}\,\mathbf{z} = 0$, then $\mathbf{\Sigma}$ coincides with the covariance matrix and $\hat{\mathbf{\Sigma}}_s$ is called the sample covariance matrix. It is clear that $\hat{\mathbf{\Sigma}}_s$ is an unbiased estimator of $\mathbf{\Sigma}$, namely, $\mathbb{E}\,\hat{\mathbf{\Sigma}}_s = \mathbf{\Sigma}$.

We begin with the following theorem, which quantifies the closeness of $\hat{\mathbf{\Sigma}}_s$ to $\mathbf{\Sigma}$.

Theorem 2.1. *If $\mathbf{S}$ is an isotropic sketch matrix with sketch size s, then for every $p \geq 1$,*

$$\mathbb{E}^{1/p}\|\hat{\mathbf{\Sigma}}_s - \mathbf{\Sigma}\|^p \lesssim_p \kappa\left(\sqrt{\frac{r}{s}} \vee \frac{r}{s}\right) \tag{6}$$

Here $\kappa = \|\mathbf{z}\|_{\psi_2}^2$ depends on the distribution.

We first connect the matrix $\mathbf{\Sigma}$ and its estimator $\hat{\mathbf{\Sigma}}_s$ with the optimal solution x^* and the Hessian sketch solution $\hat{x}_h$. Define a vector valued function from $\mathbb{S}_+^r$ to $\mathbb{R}^r$ with $\gamma > 0$:

$$f_h(\mathbf{X}) = (\mathbf{V}\mathbf{\Lambda}^\intercal\mathbf{X}\mathbf{\Lambda}\mathbf{V}^\intercal + \gamma\mathbf{I}_r)^{-1}\mathbf{A}^\intercal\mathbf{b}. \tag{7}$$

Note that under the isotropic condition, we have $\mathbf{\Sigma} = \mathbb{E}\,\mathbf{z}\mathbf{z}^\intercal = \mathbf{I}_r$. Consequently, by Equation 2 and the SVD of $\mathbf{A}$, the optimal solution x^* can be expressed as a function of the second-order moment matrix $\mathbf{\Sigma}$:

$$\mathbf{x}^* = (\mathbf{V}\mathbf{\Lambda}^\intercal\mathbf{U}^\intercal\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal + \gamma\mathbf{I}_r)^{-1}\mathbf{A}^\intercal\mathbf{b} = f_h(\mathbf{I}_r) = f_h(\mathbf{\Sigma}) = f_h(\mathbb{E}\,\mathbf{z}\mathbf{z}^\intercal).$$

Moreover, the Hessian sketch solution $\hat{x}_h$ can be expressed as the same function of the $\hat{\mathbf{\Sigma}}_s$:

$$\hat{\mathbf{x}}_h = (\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{A} + \gamma\mathbf{I}_r)^{-1}\mathbf{A}^\intercal\mathbf{b} = \left[\mathbf{V}\mathbf{\Lambda}^\intercal(\mathbf{U}^\intercal\mathbf{S}^\intercal)(\mathbf{S}\mathbf{U})\mathbf{\Lambda}\mathbf{V}^\intercal + \gamma\mathbf{I}\right]^{-1}\mathbf{A}^\intercal\mathbf{b} = f_h(\hat{\mathbf{\Sigma}}_s).$$

As a result, the Hessian sketch solution is a plug-in estimator $f_h(\hat{\mathbf{\Sigma}}_s)$ of $f_h(\mathbf{\Sigma})$ given unbiased estimator $\hat{\mathbf{\Sigma}}_s$ of $\mathbf{\Sigma}$.

Similarly, the classic sketch solution can also be formulated as a functional estimation problem involving an additional term $\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{b}$. Consider the SVD of the augmented matrix $[\mathbf{A}\ \mathbf{b}] = \mathbf{U}_+\mathbf{\Lambda}_+\mathbf{V}_+^\intercal$. Here $\mathbf{U}_+ \in \mathbb{R}^{n\times r_+}$, $\mathbf{\Lambda}_+ \in \mathbb{R}^{r_+\times r_+}$, $\mathbf{V}_+ \in \mathbb{R}^{(r+1)\times r_+}$, where $r_+ = r$ if the vector $\mathbf{b}$ lies in the column space of $\mathbf{A}$, and $r_+ = r+1$ otherwise.

Define the compressed samples $\mathbf{z}_{+,i} := \sqrt{s}\mathbf{U}_+^\intercal\mathbf{s}_i \in \mathbb{R}^{r_+}$. For the classic sketch, we require isotropy on this augmented space, namely,

$$\mathbf{\Sigma}_+ = \mathbb{E}\,\mathbf{z}_{+,i}\mathbf{z}_{+,i}^\intercal = \mathbf{I}_{r_+}.$$

The corresponding sample second-moment matrix is

$$\hat{\mathbf{\Sigma}}_{+,s} = \frac{1}{s}\sum_{i=1}^{s}\mathbf{z}_{+,i}\mathbf{z}_{+,i}^\intercal = \mathbf{U}_+^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{U}_+.$$

To express the classic sketch solution, define the vector-valued function $f_c : \mathbb{S}_+^{r_+} \to \mathbb{R}^r$ by

$$f_c(\mathbf{X}) = (\mathbf{V}_+\mathbf{\Lambda}_+^\intercal\mathbf{X}\mathbf{\Lambda}_+\mathbf{V}_+^\intercal + \gamma\mathbf{I})^{-1}\mathbf{V}_+\mathbf{\Lambda}_+^\intercal\mathbf{X}\mathbf{U}_+^\intercal\mathbf{b}.$$

Since $\mathbf{b}$ lies in the column space of $\mathbf{U}_+$, we obtain a representation of the optimal solution:

$$\mathbf{x}^* = (\mathbf{A}^\intercal\mathbf{A} + \gamma\mathbf{I}_r)^{-1}\mathbf{A}^\intercal\mathbf{b} = f_c(\mathbf{I}_{r_+}) = f_c(\mathbf{\Sigma}_+).$$

Moreover, the classic sketch solution satisfies

$$\hat{\mathbf{x}}_c = (\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{A} + \gamma\mathbf{I}_r)^{-1}\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{b} = f_c(\hat{\mathbf{\Sigma}}_{+,s}).$$

Thus, the classic sketch solution is a plug-in estimator of $f_c(\mathbf{\Sigma}_+) = \mathbf{x}^*$, based on the unbiased estimator $\hat{\mathbf{\Sigma}}_{+,s}$ of $\mathbf{\Sigma}_+$.

Note here, the isotropic condition is common in sketch problems. In Appendix B, we introduce uniform sampling, leverage score sampling and Gaussian scheme, all of them satisfy this condition.

For notational convenience, we henceforth suppress the subscript $+$ on the compressed samples and sample second-moment matrices. Accordingly, $\mathbf{\Sigma}$ and $\hat{\mathbf{\Sigma}}_s$ denote $r \times r$ matrices when used with f_h, and $r_+ \times r_+$ matrices when used with f_c.

2.3 Bias of plug-in estimator without tail probability

The plug-in estimator causes bias in functional estimation problem in general, such that

$$\mathbb{E}\,f(\hat{\mathbf{\Sigma}}_s) \neq f(\mathbf{\Sigma}),$$

for functional $f : \mathbb{S}_+^r \to \mathbb{R}$. One well-known example is the log-determinant of the covariance matrix in a Gaussian model, for which the plug-in estimator has an explicit bias term [Cai et al., 2015]. In this work, we aim to study the bias in sketched ridge regression, i.e the difference between approximated solution $\hat{\mathbf{x}}_c, \hat{\mathbf{x}}_h$ and the optimal solution $\mathbf{x}^*$ under expectation:

$$\mathbb{E}\,f_c(\hat{\mathbf{\Sigma}}_s) - f_c(\mathbf{\Sigma}) \quad \text{and} \quad \mathbb{E}\,f_h(\hat{\mathbf{\Sigma}}_s) - f_h(\mathbf{\Sigma}).$$

This bias is intrinsic due to the non-linearity of f_c and f_h. Unlike variance, which can often be reduced by repeating the same randomized experiment and averaging, bias cannot in general be removed simply by averaging multiple plug-in estimators. Given k independent estimators $f_c(\hat{\mathbf{\Sigma}}_{(i)})$ or $f_h(\hat{\mathbf{\Sigma}}_{(i)})$, $1 \le i \le k$, with same sample size, the bias remains unchanged after averaging:

$$\mathbb{E}\, f_c(\hat{\mathbf{\Sigma}}) - f_c(\mathbf{\Sigma}) = \mathbb{E}\, \frac{1}{k}\sum_{i=1}^{k} f_c(\hat{\mathbf{\Sigma}}_{(i)}) - f_c(\mathbf{\Sigma}), \quad \mathbb{E}\, f_h(\hat{\mathbf{\Sigma}}) - f_h(\mathbf{\Sigma}) = \mathbb{E}\, \frac{1}{k}\sum_{i=1}^{k} f_h(\hat{\mathbf{\Sigma}}_{(i)}) - f_h(\mathbf{\Sigma}).$$

This limitation motivates a direct analysis of the bias itself. We state the corresponding bias bound below.

Theorem 2.2. *If $\mathbf{S}$ is an isotropic sketch matrix with sketch size s, then for a Hessian sketch solution, we have*

$$\left\| \mathbb{E}\, f_h(\hat{\mathbf{\Sigma}}_s) - f_h(\mathbf{\Sigma}) \right\|_\infty \lesssim_{\sigma_1,\gamma,\|\mathbf{b}\|} \kappa^2 \left[\frac{r}{s} \vee \left(\frac{r}{s}\right)^2 \right],$$

Similarly, for the classic sketch solution, we have

$$\left\| \mathbb{E}\, f_c(\hat{\mathbf{\Sigma}}_s) - f_c(\mathbf{\Sigma}) \right\|_\infty \lesssim_{\sigma_1,\gamma,\|\mathbf{b}\|} \kappa^2 \left[\frac{r_+}{s} \vee \left(\frac{r_+}{s}\right)^2 \right].$$

In particular, if $\mathbf{S}$ is a Gaussian sketch matrix, the classic sketch bias admits the sharper bound:

$$\left\| \mathbb{E}\, f_c(\hat{\mathbf{\Sigma}}_s) - f_c(\mathbf{\Sigma}) \right\|_\infty \lesssim_{\sigma_1,\gamma,\|\mathbf{b}\|} \left[\frac{r}{s} \vee \left(\frac{r}{s}\right)^2 \right].$$

Here $\kappa = \|\mathbf{z}\|_{\psi_2}^2$ depends on the distribution.

The sharper dependence on r under Gaussian sketching arises from independence along orthogonal directions. Write

$$\mathbf{b} = \mathbf{U}\mathbf{U}^\intercal\mathbf{b} + \mathbf{b}_\perp, \qquad \mathbf{b}_\perp := (\mathbf{I}_n - \mathbf{U}\mathbf{U}^\intercal)\mathbf{b}.$$

The classic sketch solution then decomposes as

$$\hat{\mathbf{x}}_c = (\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{A} + \gamma\mathbf{I}_r)^{-1}\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{U}\mathbf{U}^\intercal\mathbf{b} + (\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{A} + \gamma\mathbf{I}_r)^{-1}\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{b}_\perp.$$

For a Gaussian sketch, $\mathbf{S}\mathbf{A}$ and $\mathbf{S}\mathbf{b}_\perp$ are independent, and $\mathbf{S}\mathbf{b}_\perp$ has mean zero. Then the second term has zero expectation. The first term is precisely the classic sketch solution with response $\mathbf{U}\mathbf{U}^\intercal\mathbf{b}$. Consequently, the bias coincides with that of the projected-response problem, whose analysis involves only r dimensions. The orthogonal component $\mathbf{b}_\perp$ increase only variance without contributing to the bias. For general sketches, isotropy alone does not guarantee this cancellation, and our general bound uses the augmented dimension r_+.

When the sketch size s satisfies $s \gtrsim r$, Theorem 2.2 indicates a bias decay of rate $O(s^{-1})$. Both covariance estimation error bound and sketch bias bound rely on the sub-Gaussian norm, which indicates the tail decay. In the Gaussian case, $\kappa \asymp 1$ and in the sub-Gaussian case, $\kappa \lesssim 1$. It's known that if an isotropic random vector supported in a finite set has constant sub-gaussian norm, then the cardinality of the support set must be exponentially large; see [Vershynin, 2018]. Therefore, discrete sampling schemes necessarily have relatively large sub-Gaussian norms. In Appendix B.1 we provide upper bounds for the uniform sketch and leverage score sampling. Therein, we also discuss when the upper bound is attained.

3 Iterative Bootstrap

We study the bootstrap in the Gaussian case. Let $\hat{\mathbf{\Sigma}}$ be the estimator based on observations $\mathbf{z}_1, \ldots, \mathbf{z}_s \sim \mathcal{N}(0, \mathbf{\Sigma}), \mathbf{\Sigma} \in \mathbb{S}_{++}^r$ and $f(\hat{\mathbf{\Sigma}})$ be the plug-in estimator of $f(\mathbf{\Sigma})$ for some functional f. Then the first order bias can be described as

$$\mathbb{E}\, f(\hat{\mathbf{\Sigma}}) - f(\mathbf{\Sigma}) := (\mathcal{B}f)(\mathbf{\Sigma}).$$

The first order bias correction yields a new estimator $f(\hat{\mathbf{\Sigma}}) - (\mathcal{B}f)(\hat{\mathbf{\Sigma}})$, with a second order bias

$$\mathbb{E}\, f(\hat{\mathbf{\Sigma}}) - \mathbb{E}(\mathcal{B}f)(\hat{\mathbf{\Sigma}}) - f(\mathbf{\Sigma}) = (\mathcal{B}f)(\mathbf{\Sigma}) - \mathbb{E}(\mathcal{B}f)(\hat{\mathbf{\Sigma}}) = -(\mathcal{B}^2 f)(\mathbf{\Sigma}).$$

Then the second order bias correction yields an estimator

$$f(\hat{\mathbf{\Sigma}}) - (\mathcal{B}f)(\hat{\mathbf{\Sigma}}) + (\mathcal{B}^2 f)(\hat{\mathbf{\Sigma}}),$$

and higher-order corrections can be obtained in a similar fashion:

$$\sum_{j=0}^{k}(-1)^j \left(\mathcal{B}^j f\right)(\mathbf{\Sigma}) = f(\mathbf{\Sigma}) + \sum_{j=1}^{k}(-1)^j \left(\mathcal{B}^j f\right)(\mathbf{\Sigma}).$$

This procedure yields the following estimator

$$f_k(\hat{\mathbf{\Sigma}}) = \sum_{i=1}^{k+1}(-1)^{i+1}\binom{k+1}{i} f(\hat{\mathbf{\Sigma}}_{(i)}) \tag{8}$$

Here $\hat{\mathbf{\Sigma}}_{(i)}$ is the covariance estimator based on an i.i.d. sample $\mathbf{z}_1^{(i-1)}, \ldots, \mathbf{z}_s^{(i-1)}$ drawn from $\mathcal{N}(0, \hat{\mathbf{\Sigma}}_{(i-1)})$. We initialize the recursion with $\hat{\mathbf{\Sigma}}_{(1)} = \hat{\mathbf{\Sigma}}$, $\mathbf{z}_i^{(0)} = \mathbf{z}_i$.

In practice, the iterative bootstrap is implemented in an alternative way by repeatedly resampling from existing samples, i.e resample s samples from $\mathbf{z}_1^{(i-1)}, \cdots, \mathbf{z}_s^{(i-1)}$ with replacement to construct $\mathbf{z}_1^{(i)}, \cdots, \mathbf{z}_s^{(i)}$. As mentioned, we don't calculate covariance estimator and plug into target function f_c and f_h. One can resample from the observed sketch, i.e uniformly sampling s rows with replacement from $\mathbf{S}$. This step has a computational cost of $O(s)$, which can be ignored in the total computational cost. Also the re-computing plug-in estimators via solving the linear system remain the same complexity as the original sketch solution. As a result, for a fixed iteration number, the iterative bootstrap will not increase the computational complexity.

It's known that, if a function f is k-th time continuously differentiable, its k-th order forward difference yields the following property: as $h \to 0$,

$$\Delta_h^k f(x) = \sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i} f(x+ih) = f^{(k)}h^k + o(h^k).$$

Denote $\hat{\mathbf{\Sigma}}_{(0)} = \mathbf{\Sigma}$, the error of iterative bootstrap (8) has a similar structure:

$$\begin{aligned} f(\mathbf{\Sigma}) - f_k(\hat{\mathbf{\Sigma}}) &= \sum_{i=0}^{k+1}(-1)^{i}\binom{k+1}{i} f(\hat{\mathbf{\Sigma}}_{(i)}) \\ &= \sum_{i=0}^{k+1}(-1)^{i}\binom{k+1}{i} f\left(\hat{\mathbf{\Sigma}}_{(0)} + \sum_{j=1}^{i}\left[\hat{\mathbf{\Sigma}}_{(j)} - \hat{\mathbf{\Sigma}}_{(j-1)}\right]\right). \end{aligned}$$

Koltchinskii and Zhilova [2021] developed a technique called random homotopy, to control the step $\hat{\mathbf{\Sigma}}_{(j)} - \hat{\mathbf{\Sigma}}_{(j-1)}$, which is the difference between one covariance and its estimator. Equipped with that, one can derive similar bias bound as the forward difference of continuously differentiable functions.

Let $f_{c,k}(\hat{\mathbf{\Sigma}})$ be the k-step bootstrap estimator of $f_c(\mathbf{\Sigma})$ and $f_{h,k}(\hat{\mathbf{\Sigma}})$ be the k-step bootstrap estimator of $f_h(\mathbf{\Sigma})$. We now state our main bias reduction result for sketched ridge regression under Gaussian sketching. As discussed above, sketching methods produce plug-in estimators whose bias decays at rate of $1/s$. By exploiting higher-order smoothness of the underlying functionals and applying iterated bias correction, we show that one can substantially accelerate the decay of the bias with respect to the sketch dimension. In particular, when the sketching matrix is Gaussian, the proposed k-th order bias-corrected estimators achieve a bias of order $(\sqrt{r/s})^{k+2}$, improving upon the baseline r/s rate. When $k = 0$, these estimators coincide with the classic sketch solution (4) and Hessian sketch solution (5), staying the same bias order.

Theorem 3.1. *Let $\mathbf{S}$ be a Gaussian sketch matrix and $s \geq Cr$ for some constant C, For the k-step parametric Gaussian bootstrap estimators and arbitrary $\mathbf{b} \in \mathbb{R}^n$, the Hessian sketch bias satisfies*

$$\left\|\mathbb{E}\, f_{h,k}(\hat{\mathbf{\Sigma}}_s) - f_h(\mathbf{\Sigma})\right\|_\infty \lesssim_{k,\sigma_1,\gamma,\|\mathbf{b}\|} \left(\sqrt{\frac{r}{s}}\right)^{k+2}.$$

Similarly, the classic sketch bias satisfies

$$\left\|\mathbb{E}\, f_{c,k}(\hat{\mathbf{\Sigma}}_s) - f_c(\mathbf{\Sigma})\right\|_\infty \lesssim_{k,\sigma_1,\gamma,\|\mathbf{b}\|} \left(\sqrt{\frac{r}{s}}\right)^{k+2}.$$

While the preceding theorem establishes improved bias decay for the proposed bias-corrected estimators, controlling the bias alone is not sufficient to characterize statistical performance. In particular, a useful estimator must balance bias reduction with variance inflation induced by the correction procedure. We therefore complement the bias analysis with L_p-risk bounds, which capture both variance and bias in a unified manner. The following theorem shows that, for Gaussian sketching, each coordinate of the k-step bootstrap-corrected estimator admits an L_p error bound with two components: a plug-in estimator variance term of order $1/\sqrt{s}$ and a bias term of order $(\sqrt{r/s})^{k+2}$.

Theorem 3.2. *Let $\mathbf{S} \in \mathbb{R}^{s\times n}$ be a Gaussian sketch matrix, $s \geq Cr$ for some constant C and consider the k-step parametric Gaussian bootstrap estimators. For arbitrary $\mathbf{b} \in \mathbb{R}^n$, every $p \geq 1$, and every $1 \leq i \leq r$, the Hessian sketch estimator satisfies*

$$\left\|\left\langle f_{h,k}(\hat{\mathbf{\Sigma}}_s) - f_h(\mathbf{\Sigma}), \mathbf{e}_i\right\rangle\right\|_{L_p} \lesssim_{k,\sigma_1,\gamma,\|\mathbf{b}\|,p} \frac{1}{\sqrt{s}} + \left(\sqrt{\frac{r}{s}}\right)^{k+2}.$$

For the classic sketch, we have

$$\left\|\left\langle f_{c,k}(\hat{\mathbf{\Sigma}}_s) - f_c(\mathbf{\Sigma}), \mathbf{e}_i\right\rangle\right\|_{L_p} \lesssim_{k,\sigma_1,\gamma,\|\mathbf{b}\|,p} \frac{1}{\sqrt{s}} + \left(\sqrt{\frac{r_+}{s}}\right)^{k+2}.$$

The distinction between the two risk bounds reflects the different ways in which the response enters the sketched estimators. For the Hessian sketch, the response appears only through $\mathbf{A}^\intercal\mathbf{b}$, and $\mathbf{A}^\intercal\mathbf{b}_\perp = 0$. Thus, the orthogonal response component has no effect on the estimator. In contrast, the classic sketch contains the additional term

$$(\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{A} + \gamma\mathbf{I}_r)^{-1}\mathbf{A}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{b}_\perp.$$

Under Gaussian sketching, this term has zero expectation, but generally has nonzero variance. It therefore does not contribute to the bias, while its fluctuations must be accounted for in the L_p-risk analysis, including after bootstrap correction. To account for this additional component, we use the r_+-dimensional augmented representation in the classic sketch risk analysis, whereas the Hessian sketch involves only the r-dimensional representation. Nevertheless, the Gaussian bias bound retains the sharper dependence on r.

Asymptotically, when the model dimension grows with the sketch size as $r \asymp s^\alpha$ for $\alpha \in (1/2, 1)$, bias will dominant in the previous L_p risk bound. To be precise, the L_p risk bound is of order $r/s \asymp s^{\alpha-1}$. However, with sufficient iteration via bootstrap, one may achieve error of order $s^{-1/2} < s^{\alpha-1}$. We state it as the following corollary.

Corollary 3.2.1. *If $\mathbf{S} \in \mathbb{R}^{s\times n}$ is chosen as a Gaussian sketch matrix and $s \geq Cr$ for some constant C and $s^\alpha \asymp r$, for $\alpha \in (1/2, 1)$. Choose $k \geq \lceil 1/(1-\alpha)\rceil - 2$, then for every $1 \leq i \leq r$,*

$$\left\|\langle f_{h,k}(\hat{\mathbf{\Sigma}}_s) - f_h(\mathbf{\Sigma}), \mathbf{e}_i\rangle\right\|_{L_p} \lesssim_{k,\sigma_1,\gamma,\|\mathbf{b}\|,p} s^{-1/2}.$$

and

$$\left\|\langle f_{c,k}(\hat{\mathbf{\Sigma}}_s) - f_c(\mathbf{\Sigma}), \mathbf{e}_i\rangle\right\|_{L_p} \lesssim_{k,\sigma_1,\gamma,\|\mathbf{b}\|,p} s^{-1/2}.$$

The next section develops debiasing strategy based on linear aggregation across multiple estimators, which cancels leading bias terms via an explicit choice of coefficients.

4 Careful Bias Reduction via Linear Aggregation

In this section, we will introduce the bias reduction via linear aggregation over multiple plug-in estimators. Let $m \geq 2$ be the number of estimators and $s/c \leq s_1 < \cdots < s_m = s$ for some constant

$c > 1$. Then, the estimators for the classic sketch solution (4) and Hessian sketch solution (5) are designed as

$$\hat{T}_{c,m} = \sum_{j=1}^{m} C_j f_c(\hat{\mathbf{\Sigma}}_{s_j}), \quad \hat{T}_{h,m} = \sum_{j=1}^{m} C_j f_h(\hat{\mathbf{\Sigma}}_{s_j}),$$

where the coefficient C_j satisfies the following linear system:

$$\sum_{j=1}^{m} C_j = 1, \qquad \sum_{j=1}^{m} \frac{C_j}{s_j^\ell} = 0, \tag{9}$$

for all integer $1 \le \ell \le m-1$. This linear system (9) has explicit solution such that

$$C_j = \prod_{i \neq j} \frac{s_j}{s_j - s_i}.$$

In this study, we only require the samples used to construct each $\hat{\mathbf{\Sigma}}_{s_j}$ to be i.i.d, which means that one may re-use same sample to construct $\hat{\mathbf{\Sigma}}_{s_j}$ for different j. This helps control the total sample size. Even if sample reuse is not allowed, the total number of samples remains at most $O(ms)$. Throughout the remainder of this work, we make the following assumption, which ensures that the aggregation coefficients do not blow up. This assumption can be easily achieved if the sample size is chosen as $s_j = \rho^{j-m} s$ for some fixed $\rho > 1$. It is also worth emphasizing that, under this choice of s_j, computing $f_c(\hat{\mathbf{\Sigma}}_{s_j})$ and $f_h(\hat{\mathbf{\Sigma}}_{s_j})$ has the same order of computational complexity as computing $f_c(\hat{\mathbf{\Sigma}}_s)$ and $f_h(\hat{\mathbf{\Sigma}}_s)$ for each j,. Consequently, for a fixed iteration count m, there is no increase in the order of the total computational complexity.

Assumption 1. *Suppose that* $\sum_{j=1}^{m} |C_j| \lesssim 1$.

The bias term admits a Taylor expansion with respect to the sketch size and the choice of coefficients 9 cancels out bias terms of order $s^{-\ell}$ explicitly for $1 \le \ell \le m-1$ as:

$$\mathbb{E}\, f(\hat{\mathbf{\Sigma}}_s) - f(\mathbf{\Sigma}) \approx \sum_{\ell=1}^{m-1} \frac{\beta_{\ell,m}(\mathbf{\Sigma})}{s^\ell} \text{ and } \sum_{j=1}^{m} C_j \frac{\beta_{\ell,m}(\mathbf{\Sigma})}{s_j^\ell} = \beta_{\ell,m}(\mathbf{\Sigma}) \sum_{j=1}^{m} \frac{C_j}{s_j^\ell} = 0,$$

with β independent to the sketch size. For details, we refer to Appendix F. We now specialize our bias bound in Theorem 4.1 and L_p risk bound in Theorem 4.2. The L_p bound only work for Gaussian case, due to a technical limitation via Gaussian concentration property, while the bias bound works in general.

Theorem 4.1. *Under the preceding assumptions on the sketch sizes and aggregation coefficients, let* $\mathbf{S} \in \mathbb{R}^{s \times n}$ *be an isotropic sketch matrix. For arbitrary* $\mathbf{b} \in \mathbb{R}^n$*, the Hessian sketch estimator satisfies*

$$\left\| \mathbb{E}\, \hat{T}_{h,m} - f_h(\mathbf{\Sigma}) \right\|_\infty \lesssim_{m,c,\gamma,\sigma_1,\|\mathbf{b}\|} \left[\kappa \left(\sqrt{\frac{r}{s}} \vee \frac{r}{s} \right) \right]^{m+1}. \tag{10}$$

Similarly, the classic sketch estimator satisfies

$$\left\| \mathbb{E}\, \hat{T}_{c,m} - f_c(\mathbf{\Sigma}) \right\|_\infty \lesssim_{m,c,\gamma,\sigma_1,\|\mathbf{b}\|} \left[\kappa \left(\sqrt{\frac{r_+}{s}} \vee \frac{r_+}{s} \right) \right]^{m+1}. \tag{11}$$

Here $\kappa = \|\mathbf{z}\|_{\psi_2}^2$ *depends on the distribution. Under Gaussian sketching,* r_+ *in the classic sketch bias bound can be replaced by* r*.*

Theorem 4.2. *Under the preceding assumptions on the sketch sizes and aggregation coefficients, suppose that the sketches are Gaussian. For arbitrary* $\mathbf{b} \in \mathbb{R}^n$*, every* $1 \le p < \infty$*, and every* $1 \le i \le r$*, the Hessian sketch estimator satisfies*

$$\left\| \langle \hat{T}_{h,m} - f_h(\mathbf{\Sigma}), \mathbf{e}_i \rangle \right\|_{L_p} \lesssim_{m,\gamma,\sigma_1,\|\mathbf{b}\|} \left(\sqrt{\frac{p}{s}} \left(\frac{r}{s} \vee 1 \right) + \frac{p}{s} \right) + \left(\sqrt{\frac{r}{s}} \vee \frac{r}{s} \right)^{m+1}. \tag{12}$$

Similarly, the classic sketch satisfies that

$$\left\| \langle \hat{T}_{c,m} - f_c(\mathbf{\Sigma}), \mathbf{e}_i \rangle \right\|_{L_p} \lesssim_{m,\gamma,\sigma_1,\|\mathbf{b}\|} \left(\sqrt{\frac{p}{s}} \left(\frac{r_+}{s} \vee 1 \right) + \frac{p}{s} \right) + \left(\sqrt{\frac{r}{s}} \vee \frac{r}{s} \right)^{m+1}. \tag{13}$$

The two terms in each bound control the random error around the estimator's expectation and its bias, respectively. For the classic sketch, $\mathbf{b}_\perp$ only contribute random error but no bias, under Gaussian setting. As a consequence, we account for this additional source of random error using the augmented representation of dimension r_+ while the bias part within dimension r. The Hessian sketch is not affected and both terms are expressed using r.

Similar to the bootstrap, the asymptotic conclusion also holds for the linear aggregation estimator. If $r \asymp s^\alpha$, increase m makes the bias decay faster, achieving a final order $s^{-1/2}$. However, without debiasing, original sketch methods only achieve order of $s^{\alpha-1}$.

Corollary 4.2.1. *If $\mathbf{S} \in \mathbb{R}^{s\times n}$ is chosen as a Gaussian sketch matrix and $s^\alpha \asymp r$, for $\alpha \in (1/2, 1)$. Choose $m \geq \lceil 1/(1-\alpha)\rceil - 1$, then for every $1 \leq i \leq r$,*

$$\left\|\langle \hat{T}_{h,m} - f_h(\mathbf{\Sigma}), \mathbf{e}_i\rangle\right\|_{L_p} \lesssim_{m,\sigma_1,\gamma,\|\mathbf{b}\|,p} s^{-1/2}, \quad \left\|\langle \hat{T}_{c,m} - f_c(\mathbf{\Sigma}), \mathbf{e}_i\rangle\right\|_{L_p} \lesssim_{m,\sigma_1,\gamma,\|\mathbf{b}\|,p} s^{-1/2}.$$

The bootstrap and linear aggregation are similar in spirit, but different. One difference here is the index m for the linear aggregation method is the number of plug-in estimators, starting at 1, while the index k for bootstrap method is the times of resampling, starting at 0. When $m = 1$ and $k = 0$, both are the single plug-in estimator, i.e the original sketch solution. Furthermore if $s > r$, and we pick $m = k + 1$, both methods yield same bias bound of order $(\sqrt{r/s})^{m+1}$ and same L_p risk of order $(\sqrt{1/s} + (\sqrt{r/s})^{m+1})$ in the Gaussian case. In particular, both approaches are linear combinations of $m = k + 1$ different estimators.

5 Conclusions and Open Questions

This work established theoretical guarantees for two resampling-based debiasing methods, iterative bootstrap and linear aggregation, for sketched ridge regression. We showed that these methods achieve higher-order bias reduction, and for Gaussian sketches we further derived L_p risk bounds.

Several questions remain open. In particular, L_p risk bounds for non-Gaussian sketches are still unknown, and a principled analysis of bootstrap debiasing under discrete sketching distributions is not yet available. In addition, our results provide only upper bounds, while matching lower bounds remain open. More broadly, it would be of interest to extend the analysis beyond Gaussian models and obtain sharper distribution-dependent risk guarantees.

Acknowledgments and Disclosure of Funding

The authors gratefully acknowledge support from the Office of Naval Research under award number N00014-23-1-2545 (Untangling Computation). FS was also supported by the Alfred P. Sloan Foundation via a Sloan Research Fellowship in Mathematics.

References

D. C. Ahfock, W. J. Astle, and S. Richardson. Statistical properties of sketching algorithms. *Biometrika*, 108(2):283–297, 2021. ISSN 0006-3444,1464-3510.

L. Birgé and P. Massart. Estimation of integral functionals of a density. *Ann. Statist.*, 23(1):11–29, 1995.

T. T. Cai, T. Liang, and H. H. Zhou. Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions. *J. Multivariate Anal.*, 137:161–172, 2015.

M. Derezinski and M. K. Warmuth. Unbiased estimates for linear regression via volume sampling. *Advances in Neural Information Processing Systems*, 30, 2017.

M. Derezinski, Z. Liao, E. Dobriban, and M. Mahoney. Sparse sketches with small inversion bias. In *Conference on Learning Theory*, pages 1467–1510. PMLR, 2021.

M. Dereziński, M. K. Warmuth, and D. Hsu. Unbiased estimators for random design regression. *J. Mach. Learn. Res.*, 23:Paper No. [167], 46, 2022.

E. Dobriban and S. Liu. Asymptotics for sketching in least squares regression. *Advances in Neural Information Processing Systems*, 32, 2019.

D. L. Donoho and R. C. Liu. Geometrizing rates of convergence. II, III. *Ann. Statist.*, 19(2):633–667, 668–701, 1991.

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

P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, 13(1):3475–3506, 2012.

B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979. ISSN 0090-5364,2168-8966.

P. Hall and M. A. Martin. On bootstrap resampling and iteration. *Biometrika*, 75(4):661–671, 1988.

J. Jiao and Y. Han. Bias correction with jackknife, bootstrap, and Taylor series. *IEEE Trans. Inform. Theory*, 66(7):4392–4418, 2020.

W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in modern analysis and probability (New Haven, Conn., 1982)*, volume 26 of *Contemp. Math.*, pages 189–206. Amer. Math. Soc., Providence, RI, 1984.

R. Kannan and S. Vempala. Randomized algorithms in numerical linear algebra. *Acta Numer.*, 26: 95–135, 2017.

V. Koltchinskii. Asymptotic efficiency in high-dimensional covariance estimation. In *Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. IV. Invited lectures*, pages 2903–2923. World Sci. Publ., Hackensack, NJ, 2018.

V. Koltchinskii. Asymptotically efficient estimation of smooth functionals of covariance operators. *J. Eur. Math. Soc. (JEMS)*, 23(3):765–843, 2021.

V. Koltchinskii. Estimation of smooth functionals of covariance operators: jackknife bias reduction and bounds in terms of effective rank. *Ann. Inst. Henri Poincaré Probab. Stat.*, 61(1):665–712, 2025.

V. Koltchinskii. Estimation of trace functionals and spectral measures of covariance operators in gaussian models. In *High Dimensional Probability X*, pages 473–550, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-06057-0.

V. Koltchinskii and K. Lounici. Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Ann. Inst. Henri Poincaré Probab. Stat.*, 52(4):1976–2013, 2016.

V. Koltchinskii and K. Lounici. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 2017.

V. Koltchinskii and M. Zhilova. Estimation of smooth functionals in normal models: bias reduction and asymptotic efficiency. *Ann. Statist.*, 49(5):2577–2610, 2021.

V. Koltchinskii, M. Löffler, and R. Nickl. Efficient estimation of linear functionals of principal components. *Ann. Statist.*, 48(1):464–490, 2020.

B. Laurent. Efficient estimation of integral functionals of a density. *Ann. Statist.*, 24(2):659–681, 1996.

M. Ledoux and M. Talagrand. *Probability in Banach spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991. Isoperimetry and processes.

O. Lepski, A. Nemirovski, and V. Spokoiny. On estimation of the L_r norm of a regression function. *Probab. Theory Related Fields*, 113(2):221–253, 1999.

S. Liu and E. Dobriban. Ridge regression: Structure, cross-validation, and sketching. In *International Conference on Learning Representations*, 2020.

M. Lopes, N. B. Erichson, and M. Mahoney. Error estimation for sketched SVD via the bootstrap. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6382–6392. PMLR, 13–18 Jul 2020.

M. E. Lopes, S. Wang, and M. W. Mahoney. A bootstrap method for error estimation in randomized matrix multiplication. *Journal of Machine Learning Research*, 20(39):1–40, 2019.

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

M. W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(2): 123–224, Feb. 2011.

P.-G. Martinsson and J. A. Tropp. Randomized numerical linear algebra: foundations and algorithms. *Acta Numer.*, 29:403–572, 2020.

R. Murray, J. Demmel, M. W. Mahoney, N. B. Erichson, M. Melnichenko, O. A. Malik, L. Grigori, P. Luszczek, M. Dereziński, M. E. Lopes, et al. Randomized numerical linear algebra: A perspective on the field with an eye to software. *arXiv preprint arXiv:2302.11474*, 2023.

C. Niu, Z. Liao, Z. Ling, and M. W. Mahoney. Fundamental bias in inverting random sampling matrices with application to sub-sampled newton. In *Forty-second International Conference on Machine Learning*, 2025.

M. Pilanci and M. J. Wainwright. Iterative Hessian sketch: fast and accurate solution approximation for constrained least-squares. *J. Mach. Learn. Res.*, 17:Paper No. 53, 38, 2016.

M. H. Quenouille. Approximate tests of correlation in time-series. *J. Roy. Statist. Soc. Ser. B*, 11: 68–84, 1949.

M. H. Quenouille. Notes on bias in estimation. *Biometrika*, 43:353–360, 1956.

J. Robins, L. Li, E. Tchetgen, and A. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, volume 2 of *Inst. Math. Stat. (IMS) Collect.*, pages 335–421. Inst. Math. Statist., Beachwood, OH, 2008.

J. Tukey. Bias and confidence in not quite large samples. *Ann. Math. Statist.*, 29:614, 1958.

R. Vershynin. *High-dimensional probability*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2018. ISBN 978-1-108-41519-4. An introduction with applications in data science.

J. Wang, J. Lee, M. Mahdavi, M. Kolar, and N. Srebro. Sketching meets random projection in the dual: A provable recovery algorithm for big and high-dimensional data. In *Artificial Intelligence and Statistics*, pages 1150–1158. PMLR, 2017.

S. Wang, A. Gittens, and M. W. Mahoney. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. *Journal of Machine Learning Research*, 18(218): 1–50, 2018.

D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Found. Trends Theor. Comput. Sci.*, 10(1-2):iv+157, 2014.

A Organization of the appendices

Appendix B introduces additional notation and some well used sketch methods with their sub-Gaussian norm bound. Appendix C collects derivative and smoothness bounds for the target functionals f_c and f_h. Appendix D provides auxiliary results, as well as the proofs of Theorems 2.1 and 2.2. Appendix E contains the proof for the iterative bootstrap method, while Appendix F contains the proof for the linear aggregation method. Additional experiments are presented in Appendix G. We evaluate bias reduction across sketching methods and sampling distributions, compare iterative methods with larger single sketches, contrast empirical and parametric bootstrap, and examine bias decay and the influence of estimator variance.

B Additional Notations and Definitions

A centered random variable X in a Banach space E is called pre-Gaussian iff there exist a centered Gaussian random variable $Y \in E$ with the same covariance operator. Additionally, a centered random variable $X \in E$ is called sub-Gaussian [Ledoux and Talagrand, 1991] iff for all $u \in E^*$,

$$\|\langle X, u\rangle\|_{\psi_2} \lesssim \|\langle X, u\rangle\|_{L_2}.$$

We introduce definitions of functional smoothness on Banach spaces, which will be used only in Appendix E. Let E, F be two Banach spaces and let $U \subset E$. For a function $g : U \mapsto F$, the L_∞ norm and Lipschitz norm are defined as follows

$$\|g\|_{L_\infty(U)} := \sup_{x\in U} \|g(x)\|, \quad \|g\|_{\mathrm{Lip}(U)} := \sup_{x,x'\in U, x\neq x'} \frac{\|g(x) - g(x')\|}{\|x - x'\|},$$

and, for $\rho \in (0, 1]$,

$$\|g\|_{\mathrm{Lip}_\rho(U)} := \sup_{x,x'\in U, x\neq x'} \frac{\|g(x) - g(x')\|}{\|x - x'\|^\rho}.$$

Let $s = m + \rho, m \geq 0, \rho \in (0, 1]$. For an m-times Fréchet differentiable function g from U into F, define

$$\|g\|_{C^s(U;F)} := \max\left(\|g\|_{L_\infty(U)}, \max_{0\leq j\leq m-1} \left\|g^{(j)}\right\|_{\mathrm{Lip}(U)}, \left\|g^{(m)}\right\|_{\mathrm{Lip}_\rho(U)}\right),$$

We will omit F if $F = \mathbb{R}$.

B.1 Some well-used sketch scheme

We introduce three standard sketching distributions that are widely used in randomized least squares and related problems.

Uniform Sampling. Each row of $\mathbf{S}$ selects a single standard basis vector. A probabilistic description is

$$\mathbb{P}\left(\mathbf{s} = \sqrt{\frac{n}{s}}\mathbf{e}_i\right) = \frac{1}{n}, \quad 1 \leq i \leq n.$$

Equivalently, $\mathbf{SA}$ consists of s rows of $\mathbf{A}$ sampled uniformly at random with replacement and scaled by $\sqrt{n/s}$ so that $\mathbb{E}\,\mathbf{S}^\intercal\mathbf{S} = \mathbf{I}_n$ consequently. This choice is attractive when the rows of $\mathbf{A}$ have roughly comparable "importance", and it is computationally cheap since forming $\mathbf{SA}$ reduces to selecting and rescaling rows. However, uniform sampling may require a larger sketch to achieve the same accuracy as other distributions.

Leverage Score Sampling. Leverage score sampling is also a sparse row-sampling sketch, but it adapts to the geometry of $\mathbf{A}$. Let $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal$ be the compact SVD with i-th row in $\mathbf{U}$ to be $\mathbf{u}_i^\intercal \in \mathbb{R}^r$. Define the i-th leverage score of matrix $\mathbf{A}$ as

$$\ell_i := \|\mathbf{u}_i\|_2^2, \qquad 1 \leq i \leq n.$$

The leverage scores satisfy that $\sum_{i=1}^{n} \ell_i = r$. Leverage score sampling chooses row indices with probabilities proportional to ℓ_i, i.e

$$\mathbb{P}\left(\mathbf{s} = \sqrt{\frac{1}{sp_i}}\mathbf{e}_i\right) = p_i, \quad p_i := \frac{\ell_i}{r}.$$

The scaling $\sqrt{1/(sp_i)}$ here again ensures $\mathbb{E}\,\mathbf{S}^\intercal\mathbf{S} = \mathbf{I}_n$. Intuitively, rows with larger leverage score contribute more, so sampling them more frequently. In practice, exact leverage scores require the computation of $\mathbf{U}$; many implementations use approximate leverage scores [Drineas et al., 2012] or shrinkage leverage scores [Ma et al., 2015].

Gaussian Sampling. Each row is an independent Gaussian vector

$$\mathbf{s} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{s}\mathbf{I}_n\right),$$

or equivalently, $\mathbf{S}_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1/s)$. This sketch is rotationally invariant and typically yields clean theoretical guarantees. They are closely related to Johnson–Lindenstrauss embeddings [Johnson and Lindenstrauss, 1984]. Its main drawback is computational. Forming $\mathbf{SA}$ is more expensive than sparse sampling sketches because $\mathbf{S}$ is dense.

We highlight here, all these three sketch scheme satisfy the isotropic condition.

Lemma B.1. *If $\mathbf{S}$ is chosen to be uniform, leverange score or Gaussian sketch matrix, then $\mathbf{S}$ is an isotropic sketch matrix.*

Proof. Since $\mathbf{U}$ has orthonormal columns, it sufficient to check $\mathbb{E}\,\mathbf{s}_i\mathbf{s}_i^\intercal = \mathbf{I}_n/s$, which implies that

$$\mathbb{E}\,\mathbf{z}_i\mathbf{z}_i^\intercal = s\mathbf{U}^\intercal\,\mathbb{E}\left[\mathbf{s}_i\mathbf{s}_i^\intercal\right]\mathbf{U}$$

It's clear here, when $\mathbf{S}$ is uniform sketch,

$$\mathbb{E}\,\mathbf{s}_i\mathbf{s}_i^\intercal = \frac{1}{n}\sum_{i=1}^{n}\frac{n}{s}\mathbf{E}_i = \frac{1}{s}\mathbf{I}_n,$$

when $\mathbf{S}$ is leverage score sketch,

$$\mathbb{E}\,\mathbf{s}_i\mathbf{s}_i^\intercal = \sum_{i=1}^{n}\frac{1}{sp_i}\mathbf{E}_i \cdot p_i = \frac{1}{s}\mathbf{I}_n,$$

and when $\mathbf{S}$ is Gaussian, it's the covariance matrix:

$$\mathbb{E}\,\mathbf{s}_i\mathbf{s}_i^\intercal = \frac{1}{s}\mathbf{I}_n.$$

□

Lemma B.2. *If $\mathbf{S}$ is chosen as uniform sampling, then $\|\mathbf{z}\|_{\psi_2} \lesssim \sqrt{\frac{n}{\log n}}$. If $\mathbf{S}$ is chosen as leverage score sampling, then $\|\mathbf{z}\|_{\psi_2} \lesssim \sqrt{\frac{r}{\log r}}$.*

Proof. For the Uniform case,

$$\|\mathbf{z}\|_{\psi_2} = \sup_{\mathbf{v}\in S^{r-1}} \|\langle\mathbf{z}, \mathbf{v}\rangle\|_{\psi_2}.$$

Given $\mathbf{v} \in S^{r-1}$,

$$\|\langle\mathbf{z}, \mathbf{v}\rangle\|_{\psi_2} = \inf\left\{t > 0 : \frac{1}{n}\sum_{i=1}^{n}\exp\left(\frac{n|\langle\mathbf{u}_i, \mathbf{v}\rangle|^2}{t^2}\right) \le 2\right\}.$$

Notice here $\mathbf{U}^\intercal\mathbf{U} = \mathbf{I}_r$,

$$\sum_{i=1}^{n}|\langle\mathbf{u}_i, \mathbf{v}\rangle|^2 = \|\mathbf{v}\|_2^2 = 1.$$

Then by the convexity of exponential function, we have

$$\frac{1}{n}\sum_{i=1}^{n}\exp\left(\frac{n|\langle \mathbf{u}_i,\mathbf{v}\rangle|^2}{t^2}\right)\le \frac{n-1}{n}+\frac{1}{n}\exp(n/t^2),$$

implying $\|\mathbf{z}\|_{\psi_2}\lesssim\sqrt{\frac{n}{\log n}}$.

Now we move to the leverage score sampling. Similarly,

$$\|\langle \mathbf{z},\mathbf{v}\rangle\|_{\psi_2}=\inf\left\{t>0:\sum_{i=1}^{n}\frac{\ell_i}{r}\exp\left(\frac{r|\langle \mathbf{u}_i,\mathbf{v}\rangle|^2}{\ell_i t^2}\right)\le 2\right\}.$$

For each fixed unit vector $\mathbf{v}\in\mathbb{R}^r$, choose an orthonormal basis $q_1,\dots,q_r$ with $q_1=\mathbf{v}$. Write

$$\mathbf{u}_i=\sum_{j=1}^{r}w_{ij}q_j,\qquad \mathbf{v}=\sum_{j=1}^{r}k_jq_j,\qquad k_1=1,\quad k_j=0\quad (j\ge 2).$$

Then

$$|\langle \mathbf{u}_i,\mathbf{v}\rangle|^2=w_{i1}^2=\sum_{j=1}^{r}w_{ij}^2k_j^2.$$

Moreover, orthonormality and $\mathbf{U}^\intercal\mathbf{U}=\mathbf{I}_r$ imply

$$\ell_i=\sum_{j=1}^{r}w_{ij}^2,\qquad \sum_{i=1}^{n}w_{ij}^2=1\quad\text{for every } j.$$

Rows with $\ell_i=0$ have zero sampling probability and are omitted from the following sums. For every i with $\ell_i>0$ and every $t>0$, convexity of the exponential gives

$$\exp\left(\frac{r|\langle \mathbf{u}_i,\mathbf{v}\rangle|^2}{\ell_i t^2}\right)=\exp\left(\sum_{j=1}^{r}\frac{w_{ij}^2}{\ell_i}\frac{rk_j^2}{t^2}\right)\le\sum_{j=1}^{r}\frac{w_{ij}^2}{\ell_i}\exp\left(\frac{rk_j^2}{t^2}\right).$$

Multiplying by ℓ_i/r and summing over i yields

$$\begin{aligned}\mathbb{E}\exp\left(\frac{|\langle \mathbf{z},\mathbf{v}\rangle|^2}{t^2}\right)&=\sum_{i:\ell_i>0}\frac{\ell_i}{r}\exp\left(\frac{r|\langle \mathbf{u}_i,\mathbf{v}\rangle|^2}{\ell_i t^2}\right)\\&\le\frac{1}{r}\sum_{j=1}^{r}\exp\left(\frac{rk_j^2}{t^2}\right)\\&=\frac{\exp(r/t^2)+r-1}{r}.\end{aligned}$$

Choosing $t^2=r/\log(1+r)$ makes the last expression equal to 2. Therefore, by the definition of the ψ_2 norm,

$$\|\mathbf{z}\|_{\psi_2}=\sup_{\|\mathbf{v}\|_2=1}\|\langle \mathbf{z},\mathbf{v}\rangle\|_{\psi_2}\le\sqrt{\frac{r}{\log(1+r)}}.$$

□

Thu uniform sketch has larger sub-Gaussian norm $\|\mathbf{z}\|_{\psi_2}$ since the random vector has scale $\sqrt{n}$ compared with leverage score sampling. Note here the sub-Gaussian norm here depends on the structure of matrix $\mathbf{U}$ since the rows of $\mathbf{U}$ may reduce the scale of the random vector $\mathbf{z}$ compared with $\mathbf{s}$ in discrete case. For example, when $\mathbf{U}=\begin{bmatrix}\mathbf{I}_r\\0\end{bmatrix}$, then for uniform sampling, $\mathbf{z}$ has probability $1/n$ to be $\sqrt{n}\mathbf{e}_i$, for that case, the upper bound on sub-Gaussian norm is attained. Similarly, for the same $\mathbf{U}$, it has r leverage score to be 1 and $n-r$ to be 0, then leverage score sampling is to draw $\sqrt{r}\mathbf{e}_i$ uniformly, with upper bound attained. But when $\mathbf{U}$ has evenly leverage scores, multiplying $\mathbf{U}$ would scale $\mathbf{z}$ by $\sqrt{r/n}$ from $\mathbf{s}$, achieve smaller sub-Gaussian norm.

C Smoothness of Target Functional

Our theory highly relies on the smoothness of the functional f_c and f_h. Due to regularization term, the derivatives can be upper bounded as below. We remind the reader that, the dimension is of r for Hessian sketch and r_+ for classic sketch.

Lemma C.1. *Let $\mathbf{A} \in \mathbb{R}^{n\times r}$ have full column rank, let $\mathbf{b} \in \mathbb{R}^n$, let $\gamma > 0$, and set $\sigma_1 = \|\mathbf{A}\|$. Write the compact SVD as $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal$. Let $\mathbf{U}_+$ have r_+ orthonormal columns forming a basis for the column space of $[\mathbf{A}\ \mathbf{b}]$. For $1 \le i \le r$, define*

$$g_{h,i}(\mathbf{X}) = \langle f_h(\mathbf{X}), \mathbf{e}_i\rangle, \quad g_{c,i}(\mathbf{X}) = \langle f_c(\mathbf{X}), \mathbf{e}_i\rangle.$$

Here derivative norms are multilinear operator norms with the matrix operator norm on each symmetric argument; L_∞ denotes the supremum of these norms over the indicated domain.

For every integer $j \ge 1$,

$$\|D^j g_{h,i}\|_{L_\infty(\mathbb{S}^r_+)} \le j!\frac{\sigma_1^{2j}}{\gamma^{j+1}}\|\mathbf{A}^\intercal\mathbf{b}\|_2.$$

For every $\mathbf{X} \in \mathbb{S}^{r_+}_+$,

$$\|D^j g_{c,i}(\mathbf{X})\| \le j!\frac{\sigma_1^{2j-1}}{\gamma^j}\|\mathbf{b} - \mathbf{A}f_c(\mathbf{X})\|_2.$$

In particular,

$$\|D^j g_{c,i}(\mathbf{I}_{r_+})\| \le j!\frac{\sigma_1^{2j-1}}{\gamma^j}\|\mathbf{b}\|_2.$$

Proof. Recall the definitions of the two functionals:

$$f_h(\mathbf{X}) = M_h(\mathbf{X})^{-1}\mathbf{A}^\intercal\mathbf{b}, \qquad f_c(\mathbf{X}) = M_c(\mathbf{X})^{-1}\mathbf{A}^\intercal(\mathbf{U}_+\mathbf{X}\mathbf{U}_+^\intercal)\mathbf{b},$$

where

$$M_h(\mathbf{X}) = \mathbf{V}\mathbf{\Lambda}^\intercal\mathbf{X}\mathbf{\Lambda}\mathbf{V}^\intercal + \gamma\mathbf{I}_r,$$
$$M_c(\mathbf{X}) = \mathbf{A}^\intercal(\mathbf{U}_+\mathbf{X}\mathbf{U}_+^\intercal)\mathbf{A} + \gamma\mathbf{I}_r.$$

The data, the bases, and γ are fixed throughout differentiation. For either $q \in \{h, c\}$ and positive-semidefinite $\mathbf{X}$ in the corresponding domain,

$$M_q(\mathbf{X}) \succeq \gamma\mathbf{I}_r, \qquad \|M_q(\mathbf{X})^{-1}\| \le \gamma^{-1}, \qquad \|DM_q(\mathbf{X})[\mathbf{H}]\| \le \sigma_1^2\|\mathbf{H}\|.$$

The same formulas define smooth functions on an open neighborhood of each such $\mathbf{X}$ in the space of symmetric matrices, where M_q remains positive definite.

For the Hessian functional, differentiate $M_h(\mathbf{X})f_h(\mathbf{X}) = \mathbf{A}^\intercal\mathbf{b}$. Since M_h is affine, for arbitrary symmetric directions $\mathbf{H}_1, \dots, \mathbf{H}_j$ and every $j \ge 1$,

$$D^j f_h(\mathbf{X})[\mathbf{H}_1, \dots, \mathbf{H}_j] = -M_h(\mathbf{X})^{-1}\sum_{\ell=1}^{j}\Big(DM_h(\mathbf{X})[\mathbf{H}_\ell]\cdot D^{j-1}f_h(\mathbf{X})[\mathbf{H}_1, \dots, \widehat{\mathbf{H}_\ell}, \dots, \mathbf{H}_j]\Big),$$

where the hat denotes omission and $D^0 f_h(\mathbf{X}) = f_h(\mathbf{X})$. Using the Euclidean norm on vector-valued outputs gives

$$\|D^j f_h(\mathbf{X})\| \le \frac{j\sigma_1^2}{\gamma}\|D^{j-1}f_h(\mathbf{X})\|.$$

Starting from $\|f_h(\mathbf{X})\|_2 \le \gamma^{-1}\|\mathbf{A}^\intercal\mathbf{b}\|_2$, induction yields

$$\|D^j f_h(\mathbf{X})\| \le j!\frac{\sigma_1^{2j}}{\gamma^{j+1}}\|\mathbf{A}^\intercal\mathbf{b}\|_2.$$

For the classic functional, recall

$$M_c(\mathbf{X})f_c(\mathbf{X}) = \mathbf{A}^\intercal\mathbf{U}_+\mathbf{X}\mathbf{U}_+^\intercal\mathbf{b}, \qquad M_c(\mathbf{X}) = \mathbf{A}^\intercal\mathbf{U}_+\mathbf{X}\mathbf{U}_+^\intercal\mathbf{A} + \gamma\mathbf{I}_r.$$

Differentiating in direction $\mathbf{H}$ gives

$$DM_c(\mathbf{X})[\mathbf{H}]\, f_c(\mathbf{X}) + M_c(\mathbf{X}) Df_c(\mathbf{X})[\mathbf{H}] = \mathbf{A}^\intercal \mathbf{U}_+ \mathbf{H} \mathbf{U}_+^\intercal \mathbf{b}.$$

Since

$$DM_c(\mathbf{X})[\mathbf{H}] = \mathbf{A}^\intercal \mathbf{U}_+ \mathbf{H} \mathbf{U}_+^\intercal \mathbf{A},$$

we obtain

$$\begin{aligned} M_c(\mathbf{X}) Df_c(\mathbf{X})[\mathbf{H}] &= \mathbf{A}^\intercal \mathbf{U}_+ \mathbf{H} \mathbf{U}_+^\intercal \mathbf{b} - \mathbf{A}^\intercal \mathbf{U}_+ \mathbf{H} \mathbf{U}_+^\intercal \mathbf{A} f_c(\mathbf{X}) \\ &= \mathbf{A}^\intercal \mathbf{U}_+ \mathbf{H} \mathbf{U}_+^\intercal \big(\mathbf{b} - \mathbf{A} f_c(\mathbf{X})\big). \end{aligned}$$

As a result,

$$Df_c(\mathbf{X})[\mathbf{H}] = M_c(\mathbf{X})^{-1} \mathbf{A}^\intercal \mathbf{U}_+ \mathbf{H} \mathbf{U}_+^\intercal \big(\mathbf{b} - \mathbf{A} f_c(\mathbf{X})\big).$$

Hence

$$\|Df_c(\mathbf{X})\| \le \frac{\sigma_1}{\gamma} \|\mathbf{b} - \mathbf{A} f_c(\mathbf{X})\|_2.$$

The right-hand side of

$$M_c(\mathbf{X}) f_c(\mathbf{X}) = \mathbf{A}^\intercal (\mathbf{U}_+ \mathbf{X} \mathbf{U}_+^\intercal) \mathbf{b}$$

is affine in $\mathbf{X}$. Thus, for $j \ge 2$, the same differentiation argument gives

$$\|D^j f_c(\mathbf{X})\| \le \frac{j \sigma_1^2}{\gamma} \|D^{j-1} f_c(\mathbf{X})\|.$$

Induction at the fixed matrix $\mathbf{X}$, starting from the first-derivative bound, therefore gives

$$\|D^j f_c(\mathbf{X})\| \le j! \frac{\sigma_1^{2j-1}}{\gamma^j} \|\mathbf{b} - \mathbf{A} f_c(\mathbf{X})\|_2.$$

For each coordinate,

$$D^j g_{q,i}(\mathbf{X})[\mathbf{H}_1, \dots, \mathbf{H}_j] = \left\langle D^j f_q(\mathbf{X})[\mathbf{H}_1, \dots, \mathbf{H}_j], \mathbf{e}_i \right\rangle.$$

Since $\|\mathbf{e}_i\|_2 = 1$, the same bounds hold for $D^j g_{q,i}(\mathbf{X})$.

Finally, since $\mathbf{U}_+ \mathbf{U}_+^\intercal \mathbf{A} = \mathbf{A}$ and $\mathbf{U}_+ \mathbf{U}_+^\intercal \mathbf{b} = \mathbf{b}$,

$$f_c(\mathbf{I}_{r_+}) = \mathbf{x}^* = (\mathbf{A}^\intercal \mathbf{A} + \gamma \mathbf{I}_r)^{-1} \mathbf{A}^\intercal \mathbf{b}.$$

Ridge optimality gives

$$\|\mathbf{A}\mathbf{x}^* - \mathbf{b}\|_2^2 + \gamma \|\mathbf{x}^*\|_2^2 \le \|\mathbf{b}\|_2^2.$$

Thus $\|\mathbf{b} - \mathbf{A}\mathbf{x}^*\|_2 \le \|\mathbf{b}\|_2$, proving the population-point bound. □

D Proof of Theorem 2.1 and Theorem 2.2

Lemma D.1. *Let X be weakly square integrable centered random vectors in $\mathbb{R}^d$ with covariance matrix Σ. Assume X is pre-Gaussian and $\|\langle X, u\rangle\|_{\psi_2} \le K \|\langle X, u\rangle\|_{L_2}$ for some K and all unit vector u. Let $X_1, \dots, X_n$ be i.i.d. copies of X, with covariance matrix estimation $\hat{\Sigma}_n = \frac{1}{n} \sum_{i=1}^n X_i X_i^\intercal$ then there exists a constant $C > 0$ such that, for all $t \ge 1$, with probability at least $1 - e^{-t}$,*

$$\|\hat{\Sigma}_n - \Sigma\| \le C K^2 \|\Sigma\| \left(\sqrt{\frac{d}{n}} \vee \frac{d}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n} \right).$$

Proof. The result follows by applying Theorem 8 in [Koltchinskii and Lounici, 2017] and repeating the argument used in the proof of Theorem 4 of [Koltchinskii and Lounici, 2017]. Our lemma can be treated as a finite-dimensional version of previous result with explicit dependence on K. Koltchinskii and Lounici [2017] only focused on the sub-Gaussian case. □

Proposition D.1.1. *Under the same assumption as Lemma D.1,*

$$\mathbb{E}^{1/p} \|\hat{\Sigma} - \Sigma\|^p \lesssim_p K^2 \|\Sigma\| \left(\sqrt{\frac{d}{n}} \vee \frac{d}{n} \right)$$

Proof. It's sufficient to integral the tail probability. Here,

$$\mathbb{E}\,\|\hat{\Sigma}-\Sigma\|^p = p\int_0^\infty u^{p-1}\mathbb{P}(\|\hat{\Sigma}-\Sigma\|>u)du.$$

Set

$$B_n := \sqrt{\frac{d}{n}} \vee \frac{d}{n}.$$

By Lemma E.1, there exists a constant $C>0$ such that for every $t \geq 1$,

$$\mathbb{P}\Big(\|\hat{\Sigma}_n-\Sigma\| > CK^2\|\Sigma\|\Big(B_n \vee \sqrt{\tfrac{t}{n}} \vee \tfrac{t}{n}\Big)\Big) \leq e^{-t}.$$

Since $d \geq 1$, we have $B_n \geq n^{-1/2}$ and $B_n \geq n^{-1}$. Hence, for all $t \geq 1$,

$$B_n \vee \sqrt{\tfrac{t}{n}} \vee \tfrac{t}{n} \leq B_n \vee \sqrt{t}\, B_n \vee tB_n \leq tB_n.$$

Therefore,

$$\mathbb{P}\Big(\|\hat{\Sigma}_n-\Sigma\| > CK^2\|\Sigma\|B_n\, t\Big) \leq e^{-t}, \qquad t \geq 1.$$

Now define

$$Y := \frac{\|\hat{\Sigma}_n-\Sigma\|}{CK^2\|\Sigma\|B_n}.$$

Then $\mathbb{P}(Y>t) \leq e^{-t}$ for all $t \geq 1$. Using the tail integration formula,

$$\mathbb{E}Y^p = p\int_0^\infty u^{p-1}\mathbb{P}(Y>u)\,du,$$

we obtain

$$\mathbb{E}Y^p \leq p\int_0^1 u^{p-1}\,du + p\int_1^\infty u^{p-1}e^{-u}\,du \leq C_p,$$

where $C_p>0$ depends only on p. Taking p-th roots gives

$$\big(\mathbb{E}\|\hat{\Sigma}_n-\Sigma\|^p\big)^{1/p} \leq C_pK^2\|\Sigma\|B_n = C_pK^2\|\Sigma\|\left(\sqrt{\frac{d}{n}} \vee \frac{d}{n}\right).$$

This proves the claim. □

Proof of Theorem 2.1. For non-centered sketch method, it's sufficient to consider a new random vector $\mathbf{s}' = \varepsilon\mathbf{s}$ for scalar valued Rademacher random variable ε which is independent with $\mathbf{s}$. Notice that $\mathbf{s}\mathbf{s}^\intercal$ has the same distribution as $\mathbf{s}'\mathbf{s}'^\intercal$. Due to the isotropic property, $\|\boldsymbol{\Sigma}\| = 1$, which implies that $\|\langle \mathbf{s}, \mathbf{u}\rangle\|_{L_2} = 1$ for unit vector $\mathbf{u}$. As a result, we can directly apply Proposition D.1.1 □

Before proving Theorem 2.2, we establish a useful identity for the classic sketch under Gaussian sketching. Write

$$\mathbf{b} = \mathbf{U}\mathbf{U}^\intercal\mathbf{b} + \mathbf{b}_\perp, \qquad \mathbf{b}_\perp = (\mathbf{I}_n - \mathbf{U}\mathbf{U}^\intercal)\mathbf{b}.$$

Since $\mathbf{A}^\intercal\mathbf{b}_\perp = 0$, the jointly Gaussian random matrix $\mathbf{S}\mathbf{A}$ and vector $\mathbf{S}\mathbf{b}_\perp$ have zero cross-covariances and are therefore independent. Consequently,

$$\mathbb{E}[\mathbf{S}\mathbf{b}_\perp \mid \mathbf{S}\mathbf{A}] = 0.$$

Write $\hat{\mathbf{x}}_c(\mathbf{S};\mathbf{v})$ for the classic sketch estimator (4) with response $\mathbf{v}$. By linearity in the response,

$$\mathbb{E}[\hat{\mathbf{x}}_c(\mathbf{S};\mathbf{b}) - \hat{\mathbf{x}}_c(\mathbf{S};\mathbf{U}\mathbf{U}^\intercal\mathbf{b})\,|\,\mathbf{S}\mathbf{A}] = \left((\mathbf{S}\mathbf{A})^\intercal(\mathbf{S}\mathbf{A}) + \gamma\mathbf{I}_r\right)^{-1}(\mathbf{S}\mathbf{A})^\intercal\,\mathbb{E}[\mathbf{S}\mathbf{b}_\perp \mid \mathbf{S}\mathbf{A}] = 0.$$

These expectations are well defined because $\gamma>0$ and the Gaussian sketch has finite second moments. Taking expectations yields

$$\mathbb{E}\,\hat{\mathbf{x}}_c(\mathbf{S};\mathbf{b}) = \mathbb{E}\,\hat{\mathbf{x}}_c(\mathbf{S};\mathbf{U}\mathbf{U}^\intercal\mathbf{b}). \tag{14}$$

Moreover, $\mathbf{A}^\intercal\mathbf{U}\mathbf{U}^\intercal\mathbf{b} = \mathbf{A}^\intercal\mathbf{b}$, so projecting the response leaves the exact ridge solution unchanged. Thus, the Gaussian classic-sketch bias coincides with that of the projected-response problem. The

latter estimator depends only on $\mathbf{SU}$, allowing its bias to be analyzed using r-dimensional compressed samples rather than the r_+-dimensional augmented samples.

To distinguish the projected-response problem from the augmented formulation, define $\widetilde{f}_c : \mathbb{S}_+^r \to \mathbb{R}^r$ by

$$\widetilde{f}_c(\mathbf{X}) = \left(\mathbf{V}\mathbf{\Lambda}^\intercal\mathbf{X}\mathbf{\Lambda}\mathbf{V}^\intercal + \gamma\mathbf{I}_r\right)^{-1}\mathbf{V}\mathbf{\Lambda}^\intercal\mathbf{X}\mathbf{U}^\intercal\mathbf{b}.$$

Here $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal$ is the compact SVD of $\mathbf{A}$. Then

$$\hat{\mathbf{x}}_c(\mathbf{S}; \mathbf{U}\mathbf{U}^\intercal\mathbf{b}) = \widetilde{f}_c(\mathbf{U}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{U}), \qquad \widetilde{f}_c(\mathbf{I}_r) = \mathbf{x}^*.$$

Under Gaussian sketching, Equation (14) therefore gives

$$\mathbb{E}\, f_c(\mathbf{U}_+^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{U}_+) - f_c(\mathbf{I}_{r_+}) = \mathbb{E}\, \widetilde{f}_c(\mathbf{U}^\intercal\mathbf{S}^\intercal\mathbf{S}\mathbf{U}) - \widetilde{f}_c(\mathbf{I}_r).$$

Thus, the Gaussian bias of the augmented classic functional equals the bias of the r-dimensional projected-response functional $\widetilde{f}_c$.

With a slight abuse of notation, we continue to write f_c for both functionals. In the Gaussian bias analysis, f_c denotes the r-dimensional projected-response functional $\widetilde{f}_c$, using the equality of biases established above. In the L_p-risk analysis and the non-Gaussian bias analysis, f_c denotes the original augmented functional on $\mathbb{S}_+^{r_+}$. The corresponding second-moment matrices are understood to have the appropriate dimensions.

Proof of Theorem 2.2. It's sufficient to take the Taylor expansion of two solution functions f_c and f_h. For simplicity, we only provide the analysis for $f = \langle f_h, \mathbf{e}_1\rangle$. The proof is exactly the same for $\langle f_c, \mathbf{e}_i\rangle$ and $\langle f_h, \mathbf{e}_i\rangle$ for each i.

Notice that,

$$f(\hat{\mathbf{\Sigma}}_s) = f(\mathbf{\Sigma}) + \langle f'(\mathbf{\Sigma}), \hat{\mathbf{\Sigma}}_s - \mathbf{\Sigma}\rangle + R(\mathbf{\Sigma}; \hat{\mathbf{\Sigma}}_s - \mathbf{\Sigma}),$$

with the remainder term

$$\left|R(\mathbf{\Sigma}; \hat{\mathbf{\Sigma}}_s - \mathbf{\Sigma})\right| \le \frac{1}{2}\left\|f^{(2)}\right\|_{L_\infty}\left\|\hat{\mathbf{\Sigma}}_s - \mathbf{\Sigma}\right\|^2.$$

By taking expectation on both size, we can get a formula of the bias term as

$$\mathbb{E}\, f(\hat{\mathbf{\Sigma}}_s) - f(\mathbf{\Sigma}) = \langle f'(\mathbf{\Sigma}), \mathbb{E}\, \hat{\mathbf{\Sigma}}_s - \mathbf{\Sigma}\rangle + \mathbb{E}\, R(\mathbf{\Sigma}; \hat{\mathbf{\Sigma}}_s - \mathbf{\Sigma}).$$

Since $\hat{\mathbf{\Sigma}}_s$ is unbiased estimator of $\mathbf{\Sigma}$, we have

$$\begin{aligned}\left|\mathbb{E}\, f(\hat{\mathbf{\Sigma}}_s) - f(\mathbf{\Sigma})\right| &= \left|\mathbb{E}\, R(\mathbf{\Sigma}; \hat{\mathbf{\Sigma}}_s - \mathbf{\Sigma})\right| \le \mathbb{E}\left|R(\mathbf{\Sigma}; \hat{\mathbf{\Sigma}}_s - \mathbf{\Sigma})\right| \\ &\lesssim \left\|f^{(2)}\right\|_{L_\infty}\mathbb{E}\left\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\right\|^2 \lesssim_{\sigma_1,\gamma,\|\mathbf{w}\|} \kappa^2\left(\sqrt{\frac{r}{s}} \vee \frac{r}{s}\right)^2.\end{aligned}$$

One can finally conclude the result for every $\langle f_c, \mathbf{e}_i\rangle$. For the Gaussian case, the bias is analyzed with the projected version. □

E Proof of Theorem 3.1 and Theorem 3.2

One can directly use the following Theorem E.1 to show Theorem 3.1 and Theorem 3.2 by choosing $a = 1$ and pick $f(\mathbf{\Sigma}) = \langle f_c(\mathbf{\Sigma}), \mathbf{e}_i\rangle$ or $f(\mathbf{\Sigma}) = \langle f_h(\mathbf{\Sigma}), \mathbf{e}_i\rangle$ for arbitrary i. Note here for the bias analysis, we use the projected functional $\widetilde{f}_c$.

Theorem E.1. *Let $X_1, \ldots, X_n$ be i.i.d. random variables sampled from a Gaussian distribution $\mathcal{N}(0, \Sigma)$ in $\mathbb{R}^d$ with unknown parameter $\theta = \Sigma \in \Theta := \mathbb{S}_{++}^d$. Define*

$$\Theta(a; d) = \{\Sigma \in \mathbb{S}_{++}^d : \sigma(\Sigma) \subset [1/a, a]\}, \quad a \ge 1,$$

where $\sigma(\Sigma)$ is the spectrum of covariance Σ.

Suppose $f \in C^s(\Theta)$ for some $s = k + 1 + \rho, k \ge 1, \rho \in (0, 1]$ and assume that, for some $\epsilon > 0$,

$$a^{s-1+\epsilon}\sqrt{\frac{d}{n}} \le c_s,$$

where $c_s \le 1/2$ is a sufficiently small constant. Then for all $\theta \in \Theta(a;d)$,

$$|\,\mathbb{E}_\theta\, f_k(\hat{\theta}) - f(\theta)| \lesssim_{s,\epsilon} \|f\|_{C^s(\Theta)} a^{(s-1+\epsilon)k+1+\rho} \left(\sqrt{\frac{d}{n}}\right)^s$$

and

$$\sup_{\theta\in\Theta(a;d)} \left\| f_k(\hat{\theta}) - f(\theta) \right\|_{L_p} \lesssim_{s,\epsilon,p} \|f\|_{C^s(\Theta)} \left(\frac{a}{\sqrt{n}} \vee a^{(s-1+\epsilon)k+1+\rho} \left(\sqrt{\frac{d}{n}}\right)^s \right).$$

This theorem can be shown with same technique used for Theorem 5.1 and Proposition 5.2 in [Koltchinskii and Zhilova, 2021]. However, their results do not directly imply the present theorem: Koltchinskii and Zhilova [2021] study functionals f that depend on both the mean and the covariance under Gaussian case, so the relevant covariance estimator is

$$\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\intercal$$

which differs from ours due to the additional uncertainty in the mean.

F Proof of Theorem 4.1 and Theorem 4.2

Lemma F.1. *Let $f : \mathbb{S}_+^r \to \mathbb{R}$ be a coordinate functional of the sketched solution, i.e., $f(\cdot) = \langle f_c(\cdot), \mathbf{e}_i\rangle$ or $f(\cdot) = \langle f_h(\cdot), \mathbf{e}_i\rangle$. Let $\hat{\mathbf{\Sigma}}_s$ denote the estimator of $\mathbf{\Sigma}$ constructed from i.i.d. samples $\mathbf{z}_1, \dots, \mathbf{z}_s$ of a random vector $\mathbf{z}$, where $\mathbf{z}$ is the compressed sample satisfying isotropic condition. Then,*

$$\mathbb{E}\, f(\hat{\mathbf{\Sigma}}_s) - f(\mathbf{\Sigma}) = \sum_{\ell=1}^{m-1} \frac{\beta_{\ell,m}(\mathbf{\Sigma})}{s^\ell} + R_{s,m} \tag{15}$$

with the remainder term

$$|R_{s,m}| \lesssim_m \left\| f^{(m+1)} \right\|_{L_\infty} \mathbb{E} \left\| \hat{\mathbf{\Sigma}}_s - \mathbf{\Sigma} \right\|^{m+1}.$$

Here $\beta_{\ell,m}(\mathbf{\Sigma})$ is a constant coefficient only depends on f and the distribution of $\mathbf{z}$.

The expansion in Lemma F.1 makes the mechanism of linear aggregation transparent: by choosing coefficients that cancel the first $m-1$ terms, one eliminates the leading bias terms. We illustrate this cancellation explicitly in the proof of the following theorem.

Theorem F.2. *Under the assumptions of Lemma F.1, define the linear aggregation estimator*

$$\hat{T} := \sum_{j=1}^m C_j f(\hat{\mathbf{\Sigma}}_{s_j}),$$

where the coefficients $\{C_j\}_{j=1}^m$ satisfy (9)*. Then*

$$\left| \mathbb{E}\,\hat{T} - f(\mathbf{\Sigma}) \right| \lesssim_m \left\| f^{(m+1)} \right\|_{L_\infty} \max_{1\le j\le m} \mathbb{E} \left\| \hat{\mathbf{\Sigma}}_{s_j} - \mathbf{\Sigma} \right\|^{m+1}. \tag{16}$$

Proof. This follows by plugging the expansion in Lemma F.1 into $\mathbb{E}\,\hat{T} - f(\mathbf{\Sigma})$ and using (9) to cancel the terms $s_j^{-\ell}$ for $1 \le \ell \le m-1$.

To be specific, first notice,

$$\mathbb{E}\,\hat{T} - f(\mathbf{\Sigma}) = \mathbb{E}\sum_{j=1}^m C_j f(\hat{\mathbf{\Sigma}}_{s_j}) - f(\mathbf{\Sigma}) = \sum_{j=1}^m C_j \left[\mathbb{E}\, f(\hat{\mathbf{\Sigma}}_{s_j}) - f(\mathbf{\Sigma}) \right],$$

since $\sum_{j=1}^m C_j = 1$. Then by Lemma F.1,

$$\begin{aligned}\mathbb{E}\,\hat{T} - f(\mathbf{\Sigma}) &= \sum_{j=1}^m C_j \left[\sum_{\ell=1}^{m-1} \frac{\beta_{\ell,m}(\mathbf{\Sigma})}{s_j^\ell} + R_{s_j}\right] \\ &= \sum_{\ell=1}^{m-1} \beta_{\ell,m}(\mathbf{\Sigma}) \sum_{j=1}^m \frac{C_j}{s_j^\ell} + \sum_{j=1}^m C_j R_{s_j} \\ &= \sum_{j=1}^m C_j R_{s_j}.\end{aligned}$$

This imples

$$\begin{aligned}\left|\mathbb{E}\,\hat{T} - f(\mathbf{\Sigma})\right| = \left|\sum_{j=1}^m C_j R_{s_j}\right| &\le \left|\sum_{j=1}^m |C_j| \max_{1\le j\le m} |R_{s_j}|\right| \\ &\lesssim_m \left\|f^{(m+1)}\right\|_\infty \max_{1\le j\le m} \mathbb{E}\left\|\hat{\mathbf{\Sigma}}_{s_j} - \mathbf{\Sigma}\right\|^{m+1}.\end{aligned}$$

□

Now we turn to prove Theorem 4.1.

Proof of Theorem 4.1. By Theorem F.2, it's remaining the check $\|f^{(m+1)}\|_{L_\infty}$ and $\mathbb{E}\,\|\hat{\mathbf{\Sigma}}_s - \mathbf{\Sigma}\|^{m+1}$, where f is chosen as $f(\cdot) = \langle f_c(\cdot), \mathbf{e}_i\rangle$ or $f(\cdot) = \langle f_h(\cdot), \mathbf{e}_i\rangle$. The derivative bound is provided by Lemma C.1 and the expectation term is due to Theorem 2.1. □

Proof of Theorem 4.2. Similar as all other proofs, we can define $f(\cdot) = \langle f_h(\cdot), \mathbf{e}_i\rangle$. It's similar for $\langle f_c(\cdot), \mathbf{e}_i\rangle$. It's sufficient to bound the L_p risk for the estimation of f, where the linear aggregation estimator of f is $T_f = \sum_{j=1}^m C_j f(\hat{\mathbf{\Sigma}}_{s_j})$. Notice here,

$$\begin{aligned}\|T_f - f(\mathbf{\Sigma})\|_{L_p} &= \|T_f - \mathbb{E}\,T_f + \mathbb{E}\,T_f - f(\mathbf{\Sigma})\|_{L_p} \\ &\le \sum_{j=1}^m |C_j| \left\|f(\hat{\mathbf{\Sigma}}_{s_j}) - \mathbb{E}\,f(\hat{\mathbf{\Sigma}}_{s_j})\right\|_{L_p} + |\,\mathbb{E}\,T_f - f(\mathbf{\Sigma})| \\ &\lesssim_m \left\|f(\hat{\mathbf{\Sigma}}_s) - \mathbb{E}\,f(\hat{\mathbf{\Sigma}}_s)\right\|_{L_p} + |\,\mathbb{E}\,T_f - f(\mathbf{\Sigma})|\end{aligned}$$

As a result, we only need to control $\|f(\hat{\mathbf{\Sigma}}_s) - \mathbb{E}\,f(\hat{\mathbf{\Sigma}}_s)\|_{L_p}$.

Based on Gaussian concentration inequality with local Lipschitz condition, one can show that

$$\|f(\hat{\mathbf{\Sigma}}_s) - \mathbb{E}\,f(\hat{\mathbf{\Sigma}}_s)\|_{L_p} \lesssim_{\gamma,\sigma_1,\|w\|} \|\mathbf{\Sigma}\| \left(\sqrt{\frac{p}{s}}\left(\frac{r}{s}\vee 1\right) + \frac{p}{s}\right),$$

where $\|\mathbf{\Sigma}\| = 1$ since $\mathbf{\Sigma} = \mathbf{I}_r$. □

Now we go back to the proof of Lemma F.1

Lemma F.1. For simplicity, we write $\mathbb{E}\,Y = \mathbf{\Sigma}$, $Y_i = \mathbf{z}_i\mathbf{z}_i^\intercal$ and $\mathbf{\Sigma}_s = \bar{Y}$. Here $\bar{Y}$ is unbiased estimator with $\mathbb{E}\,\bar{Y} = \mathbb{E}\,Y$.

By taking the Taylor expansion of f at point $\mathbb{E}\,Y$, we have

$$f(\bar{Y}) - f(\mathbb{E}\,Y) = \sum_{j=1}^m \frac{1}{j!} f^{(j)}(\mathbb{E}\,Y)[\bar{Y} - \mathbb{E}\,Y, \cdots, \bar{Y} - \mathbb{E}\,Y] + R,$$

with the remainder term R such that

$$|R| \lesssim_m \|f^{(m+1}\|_{L_\infty} \|\bar{Y} - \mathbb{E}\,Y\|^{m+1}.$$

By taking the expectation on both side, we have

$$\mathbb{E}\, f(\bar{Y}) - f(\mathbb{E}\, Y) = \sum_{j=2}^{m} \frac{\mathbb{E}\, f^{(j)}(\mathbb{E}\, Y)[\bar{Y} - \mathbb{E}\, Y, \cdots, \bar{Y} - \mathbb{E}\, Y]}{j!} + \mathbb{E}\, R$$

with remainder term satisfying that

$$\mathbb{E}\,|R| \lesssim_m \|f^{(m+1)}\|_{L_\infty}\, \mathbb{E}\, \|\bar{Y} - \mathbb{E}\, Y\|^{m+1}.$$

It's sufficient to check the multi-linear form

$$\begin{aligned}
&\mathbb{E}\, f^{(j)}(\mathbb{E}\, Y)[\bar{Y} - \mathbb{E}\, Y, \cdots, \bar{Y} - \mathbb{E}\, Y] \\
&= s^{-j} \sum_{1 \le i_1, \cdots, i_j \le s} \mathbb{E}\, f^{(j)}(\mathbb{E}\, Y)[Y_{i_1} - \mathbb{E}\, Y, \cdots, Y_{i_j} - \mathbb{E}\, Y].
\end{aligned}$$

If at least one of the indexes $i_1, \cdots, i_j$ only appear once, by the independence and $\mathbb{E}\, Y_i - \mathbb{E}\, Y = 0$

$$\mathbb{E}\, f^{(j)}(\mathbb{E}\, Y)[Y_{i_1} - \mathbb{E}\, Y, \cdots, Y_{i_j} - \mathbb{E}\, Y] = 0$$

Then we only need to consider the case where every indexes appear at least twice. Suppose the index i_j appear k_j times. It's sufficient to consider

$$M_{k_1, \cdots, k_m} = \mathbb{E}\, f^{(j)}(\mathbb{E}\, Y)[Y_1 - \mathbb{E}\, Y, \overset{k_1}{\cdots}, Y_1 - \mathbb{E}\, Y, \cdots, Y_m - \mathbb{E}\, Y, \overset{k_m}{\cdots}, Y_m - \mathbb{E}\, Y]$$

since $Y_1, \cdots, Y_s$ are identically distributed. Here $m \le j$ denote the number of distinct indexes. Let

$$B_m = \sum_{k_1 + \cdots, k_m = j, k_t \ge 2, t = 1, \cdots, m} \frac{j!}{k_1! \cdots k_m!} M_{k_1, \cdots, k_m},$$

then

$$\mathbb{E}\, f^{(j)}(\mathbb{E}\, Y)[\bar{Y} - \mathbb{E}\, Y, \cdots, \bar{Y} - \mathbb{E}\, Y] \;=\; \sum_{1 \le m \le \frac{j}{2}} \frac{\binom{s}{m}}{s^j} B_m$$

Now write $\binom{s}{m} = \sum_{i=1}^{m} c_{m,i} s^i$, as a consequence,

$$\mathbb{E}\, f^{(j)}(\mathbb{E}\, Y)[\bar{Y} - \mathbb{E}\, Y, \cdots, \bar{Y} - \mathbb{E}\, Y] \;=\; \sum_{1 \le i \le \frac{j}{2}} s^{i-j} \sum_{i \le m \le \frac{j}{2}} c_{m,i} B_m$$

Note here B_m only depends on f and the distribution and $c_{m,i}$ is a constant, so both are independent to the sketch size. Finally, combine all the multi-linear form, one can conclude the lemma. □

G Experiments

We evaluate bias reduction across sketching methods and sampling distributions, compare iterative methods with larger single sketches, contrast empirical and parametric bootstrap, and examine bias decay as the sketch size increases.

G.1 Experimental setup

Datasets and preprocessing. We conduct experiments on three datasets, Ionosphere[1], Abalone[2], and California Housing[3]. The dataset dimensions and preprocessing steps are summarized in Table 1. For California Housing, the eight features are standardized to have zero mean and unit variance before adding the intercept. The intercept is also ridge-penalized. We perform the experiments on an M2 MacBook.

[1] `https://archive.ics.uci.edu/dataset/52/ionosphere`

[2] `https://archive.ics.uci.edu/dataset/1/abalone`

[3] `https://scikit-learn.org/stable/datasets/real_world.html#california-housing-dataset`

Table 1: Dataset dimensions and preprocessing. Dimensions are reported after preprocessing. All three design matrices have full column rank.

Dataset	n	d	Feature selection	Standardized	Intercept
Ionosphere	351	33	Remove the identically zero second feature	No	No
Abalone	4177	7	Retain the seven numerical features; exclude sex	No	No
California Housing	20640	9	Retain all eight original features	Yes	Yes

Estimators and implementation. We consider three settings: the classic sketch with the original response $\mathbf{b}$, the classic sketch with the projected response $\mathbf{U}\mathbf{U}^\intercal\mathbf{b}$, and the Hessian sketch with the original response. We do not separately test the projected-response Hessian sketch, since projection leaves $\mathbf{A}^\intercal\mathbf{b}$ and hence the estimator unchanged.

Single denotes the regular sketched estimator. PB simulates the Gaussian bootstrap recursion described in Section 3, whereas EB is empirical bootstrap, an empirical row-resampling variant. At each EB transition, we sample s rows uniformly with replacement from the current sketch. For the classic estimator, the sketched features and response are resampled jointly; for the Hessian estimator, the exact right-hand side $\mathbf{A}^\intercal\mathbf{b}$ remains unchanged. Our theoretical analysis does not cover EB. The suffix $k \in \{2, 3\}$ denotes the number of terms. Each simulated bootstrap estimator uses one recursive path with $k - 1$ transitions. Single and bootstrap use sketch size s.

LA-2 and LA-3 denote linear aggregation of 2 and 3 estimators, such that sizes are chosen as $(s, \lfloor s/2 \rfloor)$ and $(s, \lfloor s/2 \rfloor, \lfloor s/3 \rfloor)$, respectively, with coefficients defined by (9).

Evaluation protocol and metrics. We set $\gamma = 0.3$ throughout. Each condition uses 20 independent repetitions. Unless stated otherwise, each repetition contains 10000 estimator draws.

Let $\mathbf{x}^*$ denote the exact ridge-regression solution and $\overline{\mathbf{x}}$ the sample mean estimate of $\mathbb{E}\,\hat{\mathbf{x}}$. The error metrics are evaluated at this sample mean. Define the ridge objective as

$$F_\gamma(\mathbf{x}; \mathbf{b}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \gamma\|\mathbf{x}\|_2^2.$$

We report the relative bias and relative objective gap:

$$\mathrm{RelBias} = \frac{\|\overline{\mathbf{x}} - \mathbf{x}^*\|_\infty}{\|\mathbf{x}^*\|_\infty},$$
$$\mathrm{RelObjGap} = \frac{F_\gamma(\overline{\mathbf{x}}; \mathbf{b}) - F_\gamma(\mathbf{x}^*; \mathbf{b})}{F_\gamma(\mathbf{0}; \mathbf{b}) - F_\gamma(\mathbf{x}^*; \mathbf{b})}.$$

The relative bias measures the discrepancy between the estimated mean and the exact solution, normalized by the size of the exact solution. The relative objective gap measures the excess ridge objective at the estimated mean, normalized by the objective improvement from the zero vector to the exact solution. Runtime is reported in microseconds per estimator as the median across the 20 repetitions. It excludes data preprocessing, response projection.

G.2 Bias reduction across sketching methods

We compare Single, EB-2, EB-3, LA-2, and LA-3 on all three datasets and in all three estimator–response settings, using Gaussian, uniform, and leverage-score sampling. We use $s = 256$ for Ionosphere and Abalone and $s = 512$ for California Housing, keeping these sizes fixed across sampling distributions. This comparison fixes the largest component sketch size rather than the total computational budget.

Results. Tables 2–4 show that correction can substantially reduce both relative bias and relative objective gap. We highlight the smallest reported value among the tested methods in the tables. At the reported sketch sizes, all four iterative estimators improve both relative error and relative objective gap under Gaussian sketching for the projected-response classic sketch estimator and the

Table 2: Debiasing comparison under Gaussian sketching.

Dataset	Metric	Single	EB-2	EB-3	LA-2	LA-3
Classical sketch, original response $\mathbf{b}$						
Ionosphere ($s = 256$)	RelBias	2.060×10^{-3}	$\mathbf{1.424\times10^{-3}}$	2.021×10^{-3}	2.476×10^{-3}	6.516×10^{-3}
	RelObjGap	2.417×10^{-7}	$\mathbf{2.393\times10^{-7}}$	7.283×10^{-7}	6.473×10^{-7}	4.601×10^{-6}
	Time (μs)	186.4	272.7	307.4	310.5	327.0
Abalone ($s = 256$)	RelBias	1.829×10^{-3}	$\mathbf{6.848\times10^{-4}}$	2.748×10^{-3}	1.377×10^{-3}	2.533×10^{-3}
	RelObjGap	3.749×10^{-8}	$\mathbf{7.472\times10^{-9}}$	6.320×10^{-8}	3.055×10^{-8}	5.424×10^{-8}
	Time (μs)	18.1	26.9	35.8	27.5	34.0
California Housing ($s = 512$)	RelBias	1.417×10^{-4}	$\mathbf{1.326\times10^{-4}}$	3.929×10^{-4}	7.417×10^{-4}	1.267×10^{-3}
	RelObjGap	$\mathbf{9.410\times10^{-9}}$	1.403×10^{-8}	6.478×10^{-8}	3.517×10^{-8}	3.075×10^{-7}
	Time (μs)	40.7	55.4	69.9	56.1	69.9
Classical sketch, projected response $\mathbf{UU^\intercal b}$						
Ionosphere ($s = 256$)	RelBias	1.865×10^{-3}	2.912×10^{-4}	$\mathbf{9.414\times10^{-5}}$	6.402×10^{-4}	3.393×10^{-4}
	RelObjGap	1.620×10^{-7}	3.907×10^{-9}	$\mathbf{4.929\times10^{-10}}$	1.935×10^{-8}	7.415×10^{-9}
	Time (μs)	98.2	372.9	665.3	164.1	189.8
Abalone ($s = 256$)	RelBias	1.801×10^{-3}	$\mathbf{3.934\times10^{-5}}$	8.469×10^{-5}	3.964×10^{-5}	1.061×10^{-4}
	RelObjGap	2.829×10^{-8}	$\mathbf{7.857\times10^{-12}}$	3.840×10^{-11}	2.997×10^{-11}	7.764×10^{-11}
	Time (μs)	16.6	31.6	45.9	25.8	32.7
California Housing ($s = 512$)	RelBias	2.185×10^{-6}	$\mathbf{2.514\times10^{-8}}$	1.610×10^{-7}	1.308×10^{-7}	1.913×10^{-7}
	RelObjGap	5.712×10^{-13}	$\mathbf{2.322\times10^{-16}}$	1.410×10^{-15}	2.107×10^{-15}	2.737×10^{-15}
	Time (μs)	37.0	81.7	125.0	55.3	69.1
Hessian sketch, original response $\mathbf{b}$						
Ionosphere ($s = 256$)	RelBias	1.526×10^{-1}	2.482×10^{-2}	$\mathbf{1.205\times10^{-2}}$	5.143×10^{-2}	5.144×10^{-2}
	RelObjGap	2.291×10^{-2}	6.382×10^{-4}	$\mathbf{1.043\times10^{-4}}$	2.863×10^{-3}	1.149×10^{-3}
	Time (μs)	202.4	248.7	309.4	298.7	379.0
Abalone ($s = 256$)	RelBias	3.002×10^{-2}	7.246×10^{-3}	1.128×10^{-2}	$\mathbf{5.722\times10^{-3}}$	2.163×10^{-2}
	RelObjGap	8.613×10^{-4}	$\mathbf{5.743\times10^{-7}}$	7.313×10^{-7}	3.707×10^{-6}	6.886×10^{-6}
	Time (μs)	23.3	37.3	50.9	39.0	49.0
California Housing ($s = 512$)	RelBias	1.951×10^{-2}	$\mathbf{4.759\times10^{-4}}$	1.672×10^{-3}	9.496×10^{-4}	9.174×10^{-3}
	RelObjGap	3.805×10^{-4}	$\mathbf{1.686\times10^{-7}}$	5.171×10^{-7}	7.121×10^{-7}	7.911×10^{-6}
	Time (μs)	78.9	134.1	187.4	131.9	185.5

Hessian estimator on all three datasets. The same improvement is observed for the projected-response classic sketch estimator under leverage-score sampling. For the classic estimator with the original response, improvements under Gaussian and leverage-score sampling are less consistent. Under uniform sampling with the original response, all four iterative estimators reduce the relative objective gap for both classic and Hessian sketches on all three datasets. Relative bias also decreases.

Using three terms does not consistently outperform using two. For example, LA-2 and LA-3 have nearly identical point estimates in the uniform Hessian experiment on Abalone. The smallest reported bias and objective gap can also correspond to different methods: for the uniform Hessian sketch on California Housing, EB-3 gives the smallest relative bias, whereas LA-3 gives the smallest relative objective gap. Overall, the results illustrate setting-dependent benefits of debiasing at additional computational cost.

In summary, our experiments show that iterative correction can effectively reduce sketching bias. Our theoretical results provide upper bounds rather than exact characterizations of the bias, and these bounds need not be tight in a particular practical setting. Although increasing the correction order yields a faster asymptotic decay rate with respect to the sketch size, the implicit constants in our bounds also depend on the correction order. Consequently, at a fixed sketch size, a higher correction order does not necessarily give a smaller upper bound or a smaller actual bias.

Table 3: Debiasing comparison under uniform sampling.

Dataset	Metric	Single	EB-2	EB-3	LA-2	LA-3
Classical sketch, original response $\mathbf{b}$						
Ionosphere ($s = 256$)	RelBias	9.135×10^{-2}	5.580×10^{-2}	$\mathbf{2.539\times10^{-2}}$	1.009×10^{-1}	7.290×10^{-2}
	RelObjGap	1.871×10^{-3}	2.610×10^{-4}	$\mathbf{6.119\times10^{-5}}$	5.220×10^{-4}	5.106×10^{-4}
	Time (μs)	39.4	286.5	534.4	65.3	85.0
Abalone ($s = 256$)	RelBias	4.234×10^{-1}	3.957×10^{-1}	3.831×10^{-1}	3.847×10^{-1}	$\mathbf{3.581\times10^{-1}}$
	RelObjGap	4.188×10^{-4}	3.580×10^{-4}	3.350×10^{-4}	3.379×10^{-4}	$\mathbf{2.933\times10^{-4}}$
	Time (μs)	10.9	26.2	41.3	16.9	20.9
California Housing ($s = 512$)	RelBias	4.041×10^{-2}	2.905×10^{-2}	2.646×10^{-2}	2.656×10^{-2}	$\mathbf{2.403\times10^{-2}}$
	RelObjGap	2.955×10^{-4}	1.292×10^{-4}	1.043×10^{-4}	1.021×10^{-4}	$\mathbf{8.264\times10^{-5}}$
	Time (μs)	19.7	63.7	107.7	31.1	38.1
Classical sketch, projected response $\mathbf{UU^\intercal b}$						
Ionosphere ($s = 256$)	RelBias	5.484×10^{-3}	2.007×10^{-3}	$\mathbf{7.538\times10^{-4}}$	5.818×10^{-3}	3.391×10^{-3}
	RelObjGap	1.537×10^{-6}	2.493×10^{-7}	$\mathbf{5.086\times10^{-8}}$	2.075×10^{-6}	1.071×10^{-6}
	Time (μs)	43.0	289.6	541.5	73.7	91.7
Abalone ($s = 256$)	RelBias	5.151×10^{-2}	4.455×10^{-2}	4.292×10^{-2}	4.312×10^{-2}	$\mathbf{3.983\times10^{-2}}$
	RelObjGap	6.454×10^{-6}	4.544×10^{-6}	4.210×10^{-6}	4.249×10^{-6}	$\mathbf{3.625\times10^{-6}}$
	Time (μs)	9.2	26.8	42.0	20.0	23.3
California Housing ($s = 512$)	RelBias	2.537×10^{-5}	6.314×10^{-6}	4.683×10^{-6}	2.073×10^{-6}	$\mathbf{2.009\times10^{-6}}$
	RelObjGap	2.213×10^{-11}	2.427×10^{-12}	1.623×10^{-12}	8.484×10^{-13}	$\mathbf{6.209\times10^{-13}}$
	Time (μs)	22.4	65.4	108.8	33.0	42.9
Hessian sketch, original response $\mathbf{b}$						
Ionosphere ($s = 256$)	RelBias	1.560×10^{-1}	5.392×10^{-2}	$\mathbf{3.386\times10^{-2}}$	1.199×10^{-1}	1.528×10^{-1}
	RelObjGap	7.979×10^{-3}	8.044×10^{-4}	$\mathbf{2.048\times10^{-4}}$	4.180×10^{-3}	4.110×10^{-3}
	Time (μs)	40.7	284.0	525.7	68.8	89.2
Abalone ($s = 256$)	RelBias	7.765×10^{-1}	6.723×10^{-1}	6.641×10^{-1}	$\mathbf{6.565\times10^{-1}}$	6.596×10^{-1}
	RelObjGap	2.803×10^{-3}	1.036×10^{-3}	1.009×10^{-3}	$\mathbf{1.001\times10^{-3}}$	1.003×10^{-3}
	Time (μs)	15.7	34.7	53.0	26.8	35.8
California Housing ($s = 512$)	RelBias	1.542×10^{-1}	2.218×10^{-2}	$\mathbf{2.099\times10^{-2}}$	3.770×10^{-2}	2.148×10^{-2}
	RelObjGap	1.665×10^{-3}	8.383×10^{-5}	7.257×10^{-5}	9.279×10^{-5}	$\mathbf{6.058\times10^{-5}}$
	Time (μs)	56.8	136.3	214.3	105.8	151.7

G.3 Iterative methods versus larger single sketches

We compare EB-2, EB-3, LA-2, and LA-3 with larger single sketches on Abalone, using classic Gaussian sketching and the projected response $\mathbf{UU^\intercal b}$. The maximum component sketch size is $s \in \{64, 128, 256\}$.

For each iterative method, the corresponding single-sketch size s_{S} equals the sum of its component sizes. For EB, this sum includes rows reused through resampling. Thus, the comparison matches the total number of rows entering the component solves.

Unlike the other experiments, this comparison uses 2000 draws per repetition.

Results. Table 5 shows that iterative bootstrap and linear aggregation achieve smaller relative bias and relative objective gap than their corresponding larger single sketches in all 12 tested configurations. The iterative methods retain only 7.2%–47.4% of the single-sketch relative bias and 0.34%–10.46% of its relative objective gap, while requiring 0.99–1.30 times its measured runtime. Thus, in this setting, using several smaller component sketches for iterative debiasing can be more effective than simply increasing the size of a single sketch, at comparable measured runtime.

Table 4: Debiasing comparison under leverage-score sampling.

Dataset	Metric	Single	EB-2	EB-3	LA-2	LA-3
Classical sketch, original response $\mathbf{b}$						
Ionosphere ($s = 256$)	RelBias	2.423×10^{-3}	$\mathbf{1.971\times10^{-3}}$	3.945×10^{-3}	2.928×10^{-3}	6.100×10^{-3}
	RelObjGap	$\mathbf{4.138\times10^{-7}}$	5.201×10^{-7}	1.046×10^{-6}	1.381×10^{-6}	3.470×10^{-6}
	Time (μs)	75.4	116.4	157.3	107.0	129.1
Abalone ($s = 256$)	RelBias	1.175×10^{-3}	$\mathbf{1.007\times10^{-3}}$	2.508×10^{-3}	2.324×10^{-3}	2.810×10^{-3}
	RelObjGap	1.000×10^{-8}	$\mathbf{7.386\times10^{-9}}$	2.829×10^{-8}	3.658×10^{-8}	1.846×10^{-7}
	Time (μs)	25.1	34.7	43.6	41.5	57.5
California Housing ($s = 512$)	RelBias	3.054×10^{-4}	$\mathbf{1.495\times10^{-4}}$	3.564×10^{-4}	4.352×10^{-4}	1.295×10^{-3}
	RelObjGap	$\mathbf{1.064\times10^{-8}}$	1.865×10^{-8}	4.148×10^{-8}	5.946×10^{-8}	2.924×10^{-7}
	Time (μs)	81.6	96.3	109.5	127.8	173.6
Classical sketch, projected response $\mathbf{UU^\intercal b}$						
Ionosphere ($s = 256$)	RelBias	1.806×10^{-3}	4.011×10^{-4}	$\mathbf{1.191\times10^{-4}}$	8.028×10^{-4}	6.190×10^{-4}
	RelObjGap	1.535×10^{-7}	8.434×10^{-9}	$\mathbf{1.244\times10^{-9}}$	3.303×10^{-8}	2.292×10^{-8}
	Time (μs)	51.5	299.5	548.7	85.6	110.4
Abalone ($s = 256$)	RelBias	1.372×10^{-3}	5.835×10^{-5}	$\mathbf{5.265\times10^{-5}}$	1.006×10^{-4}	1.034×10^{-4}
	RelObjGap	1.615×10^{-8}	3.616×10^{-11}	$\mathbf{1.484\times10^{-11}}$	5.749×10^{-11}	1.873×10^{-10}
	Time (μs)	24.3	39.3	53.9	40.4	53.0
California Housing ($s = 512$)	RelBias	1.778×10^{-6}	7.874×10^{-8}	$\mathbf{2.696\times10^{-8}}$	8.688×10^{-8}	2.166×10^{-7}
	RelObjGap	3.707×10^{-13}	9.512×10^{-16}	6.590×10^{-16}	$\mathbf{4.197\times10^{-16}}$	5.792×10^{-15}
	Time (μs)	77.3	120.6	164.4	127.5	169.4
Hessian sketch, original response $\mathbf{b}$						
Ionosphere ($s = 256$)	RelBias	1.459×10^{-1}	3.773×10^{-2}	$\mathbf{2.061\times10^{-2}}$	6.593×10^{-2}	7.910×10^{-2}
	RelObjGap	2.267×10^{-2}	1.993×10^{-3}	$\mathbf{3.649\times10^{-4}}$	6.977×10^{-3}	9.914×10^{-3}
	Time (μs)	73.5	116.6	177.3	115.1	148.2
Abalone ($s = 256$)	RelBias	1.905×10^{-2}	7.162×10^{-3}	9.384×10^{-3}	$\mathbf{5.245\times10^{-3}}$	2.739×10^{-2}
	RelObjGap	4.819×10^{-4}	$\mathbf{7.070\times10^{-7}}$	1.371×10^{-6}	1.098×10^{-6}	5.990×10^{-6}
	Time (μs)	31.6	45.3	60.0	52.3	71.1
California Housing ($s = 512$)	RelBias	1.589×10^{-2}	$\mathbf{9.219\times10^{-4}}$	1.115×10^{-3}	1.651×10^{-3}	3.609×10^{-3}
	RelObjGap	2.562×10^{-4}	2.778×10^{-7}	$\mathbf{2.776\times10^{-7}}$	6.571×10^{-7}	1.107×10^{-6}
	Time (μs)	119.8	173.6	227.1	206.8	291.9

Table 5: Iterative bootstrap and linear aggregation versus larger single sketches on Abalone. The corresponding single-sketch size s_{S} equals the sum of the component sizes. Bold marks the smaller relative bias and relative objective gap in each row.

Method	Component sizes	Iterative method			Corresponding single sketch			
		RelBias	RelObjGap	Time (μs)	s_{S}	RelBias	RelObjGap	Time (μs)
EB-2	$(64, 64)$	$\mathbf{2.843\times10^{-4}}$	$\mathbf{1.302\times10^{-9}}$	9.17	128	3.794×10^{-3}	1.207×10^{-7}	8.62
EB-3	$(64, 64, 64)$	$\mathbf{1.767\times10^{-4}}$	$\mathbf{1.750\times10^{-10}}$	12.86	192	2.451×10^{-3}	5.173×10^{-8}	12.01
LA-2	$(64, 32)$	$\mathbf{1.468\times10^{-3}}$	$\mathbf{2.230\times10^{-8}}$	8.71	96	5.017×10^{-3}	2.162×10^{-7}	6.71
LA-3	$(64, 32, 21)$	$\mathbf{7.259\times10^{-4}}$	$\mathbf{2.608\times10^{-9}}$	10.76	117	3.954×10^{-3}	1.355×10^{-7}	8.41
EB-2	$(128, 128)$	$\mathbf{1.511\times10^{-4}}$	$\mathbf{1.884\times10^{-10}}$	15.78	256	1.842×10^{-3}	2.833×10^{-8}	15.12
EB-3	$(128, 128, 128)$	$\mathbf{4.662\times10^{-4}}$	$\mathbf{7.913\times10^{-10}}$	22.82	384	1.220×10^{-3}	1.241×10^{-8}	21.83
LA-2	$(128, 64)$	$\mathbf{5.653\times10^{-4}}$	$\mathbf{1.977\times10^{-9}}$	14.22	192	2.451×10^{-3}	5.173×10^{-8}	12.01
LA-3	$(128, 64, 42)$	$\mathbf{7.100\times10^{-4}}$	$\mathbf{3.469\times10^{-9}}$	17.82	234	2.008×10^{-3}	3.317×10^{-8}	14.13
EB-2	$(256, 256)$	$\mathbf{1.196\times10^{-4}}$	$\mathbf{1.241\times10^{-10}}$	30.32	512	9.339×10^{-4}	7.289×10^{-9}	28.19
EB-3	$(256, 256, 256)$	$\mathbf{2.772\times10^{-4}}$	$\mathbf{1.587\times10^{-10}}$	44.82	768	5.845×10^{-4}	3.023×10^{-9}	45.50
LA-2	$(256, 128)$	$\mathbf{1.764\times10^{-4}}$	$\mathbf{9.255\times10^{-11}}$	24.23	384	1.220×10^{-3}	1.241×10^{-8}	21.83
LA-3	$(256, 128, 85)$	$\mathbf{2.472\times10^{-4}}$	$\mathbf{7.785\times10^{-10}}$	30.56	469	1.040×10^{-3}	8.914×10^{-9}	26.18

G.4 Empirical versus parametric bootstrap

We compare EB-2, PB-2, EB-3, and PB-3 with the single-sketch baseline on Abalone, using Gaussian sketches of size $s = 256$ in all three estimator–response settings.

Table 6: Empirical versus parametric bootstrap on Abalone. Bold marks the smallest relative bias and relative objective gap in each row.

Metric	Single	EB-2	PB-2	EB-3	PB-3
Classical sketch, original response $\mathbf{b}$					
RelBias	2.200×10^{-3}	$\mathbf{1.130\times10^{-3}}$	1.497×10^{-3}	1.173×10^{-3}	2.024×10^{-3}
RelObjGap	4.018×10^{-8}	$\mathbf{9.072\times10^{-9}}$	1.288×10^{-8}	1.382×10^{-8}	4.184×10^{-8}
Time (μs)	15.77	30.63	55.57	44.91	95.07
Classical sketch, projected response $\mathbf{UU^\top b}$					
RelBias	1.826×10^{-3}	$\mathbf{3.373\times10^{-5}}$	3.766×10^{-5}	4.102×10^{-5}	5.335×10^{-5}
RelObjGap	2.882×10^{-8}	1.298×10^{-11}	$\mathbf{1.032\times10^{-11}}$	2.353×10^{-11}	1.977×10^{-11}
Time (μs)	18.12	36.19	63.51	52.33	114.27
Hessian sketch, original response $\mathbf{b}$					
RelBias	2.727×10^{-2}	8.812×10^{-3}	$\mathbf{3.297\times10^{-3}}$	1.509×10^{-2}	7.547×10^{-3}
RelObjGap	8.534×10^{-4}	7.583×10^{-7}	1.273×10^{-6}	1.726×10^{-6}	$\mathbf{4.556\times10^{-7}}$
Time (μs)	23.05	43.57	73.75	63.25	122.99

Results. Both bootstrap procedures substantially reduce the reported relative bias and relative objective gap for the projected-response classic sketch and the Hessian sketch. Neither bootstrap procedure uniformly outperforms the other across settings and metrics. PB requires more runtime in the current implementation, which additionally constructs a covariance factor and generates Gaussian samples at each transition. EB is relatively cheaper in computation since it resamples existing rows.

G.5 Bias decay with increasing sketch size

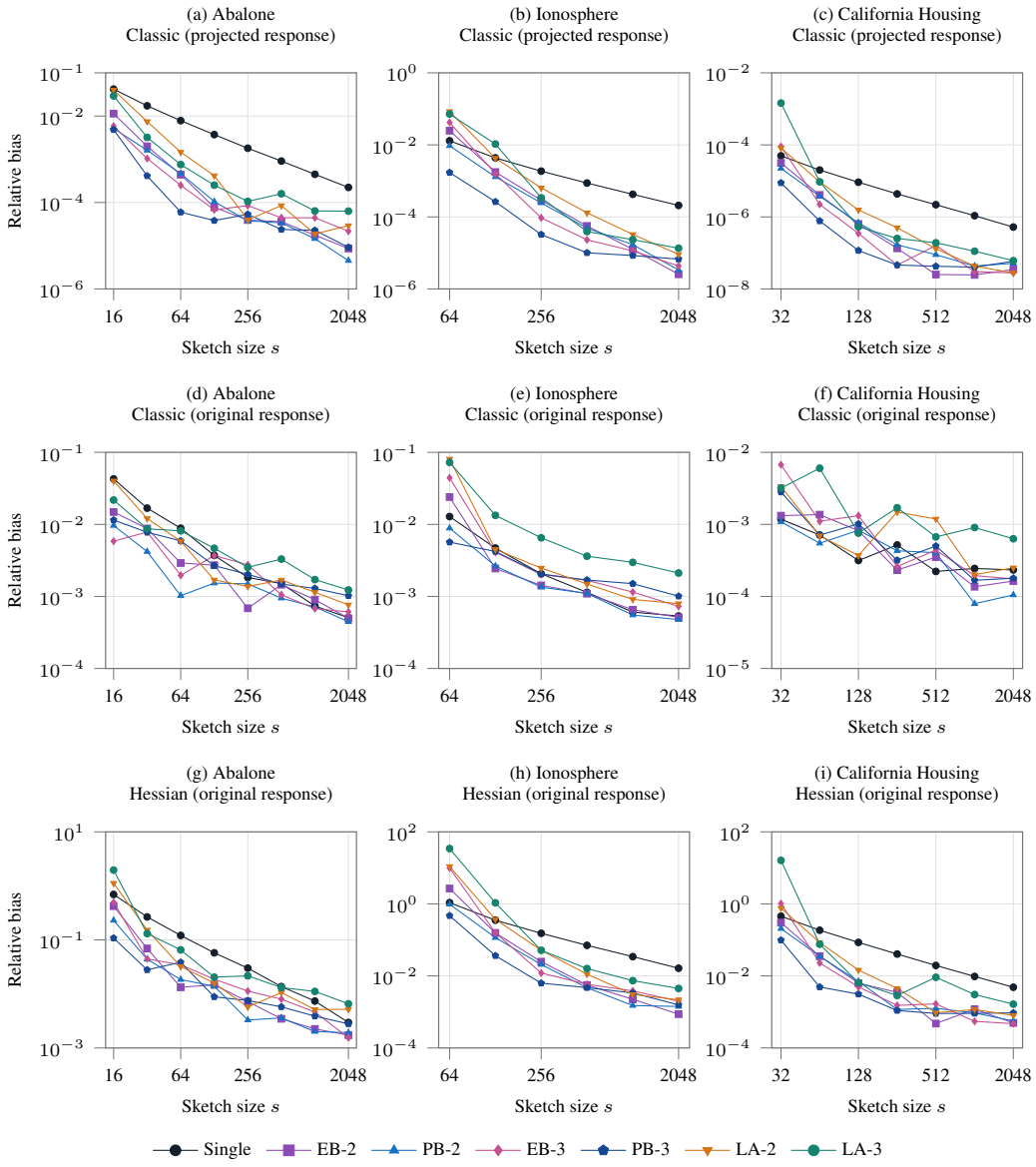

Figure 1: Relative bias versus sketch size for Gaussian sketches across three datasets and three sketch settings.

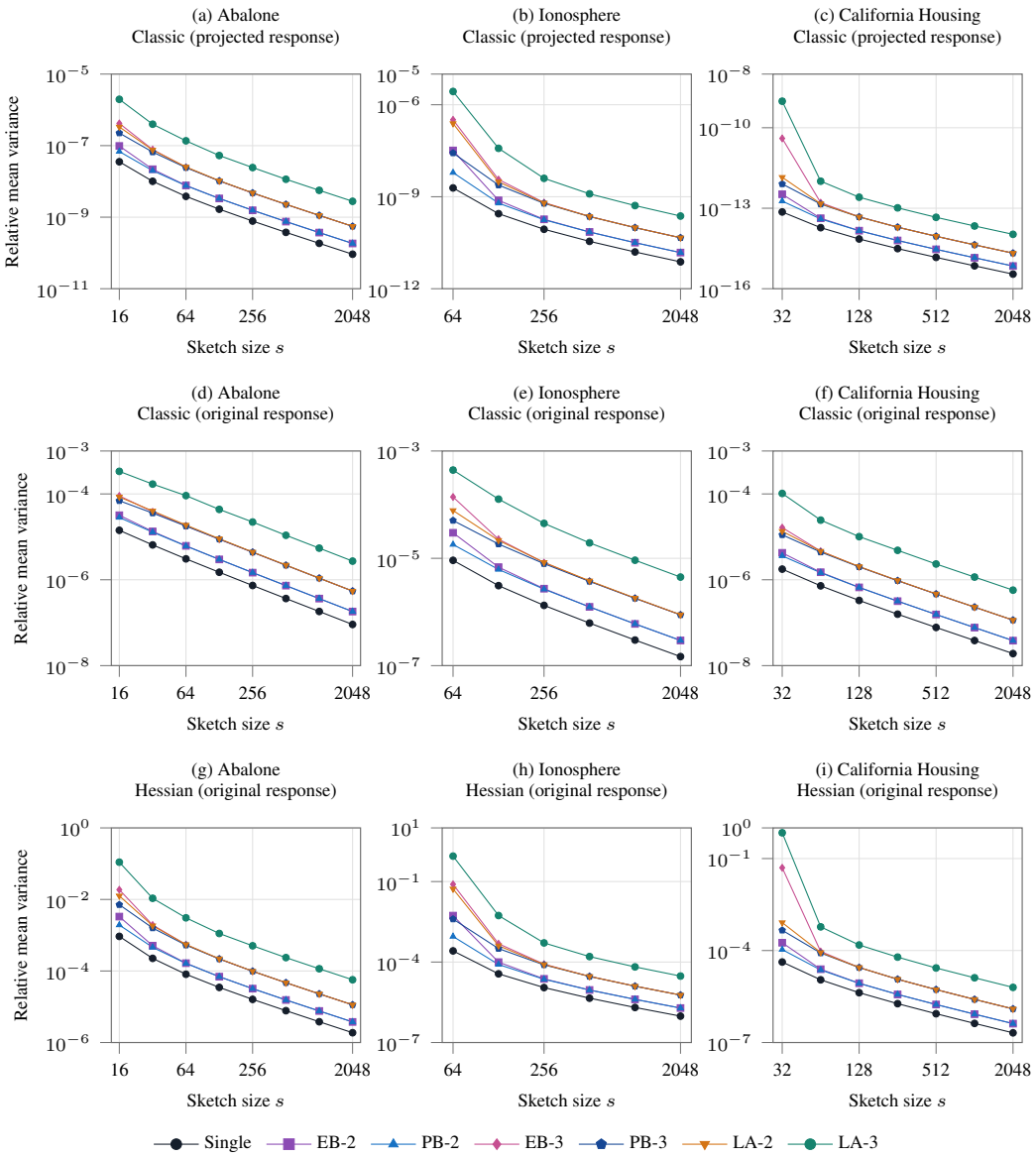

Figure 2: Normalized variance of the Monte Carlo average versus sketch size, with $N = 200{,}000$ draws.

We examine how relative bias changes with sketch size under Gaussian sketching on all three datasets and in all three estimator–response settings. We compare Single, EB-2, PB-2, EB-3, PB-3, LA-2, and LA-3. The sketch sizes double from 16 to 2048 for Abalone, from 64 to 2048 for Ionosphere, and from 32 to 2048 for California Housing. The trace of sample covariance for sketch solution is also reported.

Results. Figure 1 shows that debiasing can substantially reduce bias and produce a faster observed decrease over parts of the tested sketch-size range. This behavior is most apparent for the projected-response classic sketch and the Hessian sketch. For the classic sketch with the original response, the separation between the single-sketch and debiasing methods is less clear.

Some bias decay curves decrease slowly at larger sketch sizes. The high order bias no longer dominant in the error while estimator variance limits the observed rate of decay in this regime.

Effect of estimator variance. Bias reduction can come at the cost of increased estimator variance, measured here by the trace of the covariance matrix. For example, at $s = 2048$, the implemented two-term bootstrap estimators have approximately twice the variance of the single sketch, while three-term bootstrap and two-term linear aggregation have approximately six times its variance, as shown in Figure 2. This increased variance makes small residual biases harder to estimate and can obscure their decay.

For the single classic Gaussian sketch, the contribution from $\mathbf{b}_\perp$ has zero expectation but adds variance relative to the projected-response estimator. Comparing the two response settings illustrates how this additional random error can obscure the estimator performance. The effect can be more pronounced for iterative methods, whose weighted combinations can amplify the random terms involving $\mathbf{b}_\perp$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the study of bias in sketched ridge regression problem. The paper provides theoretical bias reduction guarantee via iterative methods, with empirical validation.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitation and future work is discussed in the Conclusion section

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theorems include clear assumptions and full proof are provided in Appendix D,Appendix E, Appendix F

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix G provides implement details including parameter setting, evaluation metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data is public via UC Irvine Machine Learning Repository (`https://archive.ics.uci.edu`). The code is provided as in the supplemental material.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix G describes the detail of all experiment setting.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the errors across repeated trials and present the results with clear plots and captions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our experiments are not computationally heavy and are run on a commercial laptop. There is no specific resources required to reproduce the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is mainly theoretical and uses only public datasets, and conforms to ethical standards.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is a theoretical work, with no direct societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper cites datasets properly.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This research does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This research does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The he core method development in this research does not involve LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.