Lower Bounds and a Near-Optimal Shrinkage Estimator for Least Squares using Random Projections

Srivatsan Sridhar, Mert Pilanci, and Ayfer Özgür

Stanford University

{svatsan, pilanci, aozgur}@stanford.edu

Abstract

In this work, we consider the deterministic optimization using random projections as a statistical estimation problem, where the squared distance between the predictions from the estimator and the true solution is the error metric. In approximately solving a large scale least squares problem using Gaussian sketches, we show that the sketched solution has a conditional Gaussian distribution with the true solution as its mean. Firstly, tight worst case error lower bounds with explicit constants are derived for any estimator using the Gaussian sketch, and the classical sketching is shown to be the optimal unbiased estimator. For biased estimators, the lower bound also incorporates prior knowledge about the true solution. Secondly, we use the James-Stein estimator to derive an improved estimator for the least squares solution using the Gaussian sketch. An upper bound on the expected error of this estimator is derived, which is smaller than the error of the classical Gaussian sketch solution for any given data. The upper and lower bounds match when the SNR of the true solution is known to be small and the data matrix is well conditioned. Empirically, this estimator achieves smaller error on simulated and real datasets, and works for other common sketching methods as well.

1 Introduction

One of the simplest and most ubiquitous of optimization problems is the least squares problem. The least squares problem applies to several common models such as linear regression, maximum likelihood estimation, and loss functions in neural networks. The unconstrained least squares problem can be defined as

$$x_{LS} = \arg\min_{x \in \mathbb{R}^d} \|Ax - y\|_2^2$$

The data matrix $A \in \mathbb{R}^{n \times d}$ and the data vector $y \in \mathbb{R}^n$ are fixed and known, and x_{LS} is the least squares solution. In this work, we consider $n > d$. This problem has a unique analytical solution $x_{LS} = A^\dagger y$ where $A^\dagger = (A^T A)^{-1} A^T$ is the pseudoinverse of A, whenever $rank(A) = d$. This analytical solution can be computed efficiently using singular value decomposition, or Cholesky or QR decomposition, all of which require $\mathcal{O}(nd^2)$ time.

Often in modern datasets, we deal with very large scale data, i.e. n and d are very large, making the computation of the true least squares solution very expensive. We can approximate the least squares solution by projecting A and y onto a lower dimension $m < n$, and then solving the least squares problem using the projected data. This method is known as sketching, and can be defined as

$$\widehat{x} = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sy\|_2^2$$

where $S \in \mathbb{R}^{m \times n}$ is a random projection matrix. m is called the sketch size. This sketched problem can be solved in $\mathcal{O}(md^2)$ time, plus the time to compute the sketch SA. Apart from reducing computation time, sketching can also be useful to provide differential privacy [1, 2, 3]. A reasonable sketch size must be larger than d, and grows proportionally to d.

The error of the sketched estimator can be measured as the distance between the predictions due to the estimator and the true solution $\|A(\widehat{x} - x_{LS})\|_2^2$ (prediction error), which

we will consider in this work. It is known that this approximately grows as d/m for several sketching methods [4, 5, 6].

There are several common sketching methods, discussed in [7, 8, 9, 10]. One simple method is forming SA and Sy by sampling m rows of A and y [7, 8, 11, 12]. One can instead use a random matrix S from a data-oblivious distribution. $S_{ij} \overset{\text{i.i.d.}}{\sim} \frac{1}{\sqrt{m}}\mathcal{N}(0,1)$, called the Gaussian sketch, is the canonical example [8]. Many other sketching methods, such as the Rademacher sketch, and the Fast Johnson-Lindenstrauss (JL) sketch, tend to behave very similar to the Gaussian sketch.

The Gaussian sketch has many desirable properties. Firstly, the expected error can be exactly computed. Secondly, fundamental to our results, the sketched solution $\widehat{x}$ has a Gaussian distribution with mean x_{LS}, when conditioned on SA. The latter property allows us to view the deterministic optimization problem as a task of statistical estimation - estimating the mean x_{LS} from $\widehat{x}$. To our knowledge, previous studies have not considered the estimation aspect of sketching.

Despite its simplicity and effectiveness, there are very few results about error lower bounds for the Gaussian sketch, and these lower bounds involve constants that are hard to compute [5, 6, 13]. Thus there are no results about the optimality of the sketching methods. In this paper, we provide two error lower bounds for the Gaussian sketch. The first one which holds for unbiased estimators only, shows that the classical sketching method is the optimal unbiased estimator. The second lower bound holds for general estimators in the presence of prior knowledge about the solution, and is in terms of explicit constants. Further, much of the previous work assumes a statistical observation model on the data, such as $y = Ax^* + e$, where e is Gaussian noise, which is restrictive and may not be valid in many cases. Instead, our lower bounds assume no such model. Interestingly, the statistical nature of the problem arises because of the randomness in the sketching only, and not from the data model.

We also note that standard sketching methods are based on a sketch-and-solve approach (i.e. solving the original problem with the sketched data), which lead to unbiased estimators

in the Gaussian sketch. However, this may not be the best approach, e.g., iterative sketching methods perform better [5]). For estimating the unknown mean of a high-dimensional Gaussian vector, it is known that a "shrinkage estimator" (e.g. the James-Stein estimator) has lower variance than the maximum likelihood (ML) estimator. We use this fact to derive an improved estimator for x_{LS}, based on the Gaussian sketch. This estimator is biased, however, has lower error than the classical Gaussian sketch, more so when the data is noisy. Furthermore, this estimator achieves our lower bound when the SNR of the true solution is known to be small and the data matrix is well conditioned. Thus, our analysis introduces a bias-variance trade-off in sketching methods to improve the classical sketch-and-solve paradigm.

1.1 Our Contributions

We first show two worst case lower bounds on the prediction error of any estimator that can be obtained from the Gaussian sketch. The first one holds for unbiased estimators of x_{LS}, and shows that the classical sketching method is the minimum variance unbiased estimator (MVUE) of x_{LS}. Although MVUE is a well known concept for statistical estimation under additive Gaussian noise, this is a novel result for sketching the deterministic least squares optimization problem where the only source of randomness is the sketching matrix. The second lower bound holds for biased estimators as well. The second lower bound deals with constrained least squares problems in general, and this allows us to incorporate the effect of prior knowledge about the least squares solution (such as constraints on the norm or SNR). Both the lower bounds are easy to evaluate in terms of properties of the data (and the constraints). To derive the lower bounds, we use the Fisher information, along with the Cramér-Rao inequality for the unbiased case, and the van Trees inequality [14] for the general case.

Our second result is an improved estimator for the Gaussian sketch, based on the James-Stein estimator. We provide an approximate form of the estimator which can be evaluated from the Gaussian sketch solution without much additional computation. We derive an

upper bound on the prediction error of this estimator, and show that the error is lesser than that of the classical Gaussian sketch for any given data, even more so when the signal-to-noise ratio (SNR) of the data is low. Finally we show that the lower and upper bounds match for large m upto constants in the additive term, when the SNR of the data is known to be small and A is well-conditioned.

Empirically, we show that our estimator achieves lower prediction error for synthetic as well as real data, especially when the SNR is small. We also show that our estimator works equally well for other sketching methods such as sampling and Fast JL sketch. We also show extensions of our results to matrix regression problems.

1.2 Previous Works

There are many different sketching methods in literature [7, 8, 9, 10]. Among sampling strategies, a simple uniform sampling can perform poorly in the worst case, but sampling based on leverage scores is known to be better [11]. Other than Gaussian, common sketching matrices are the Rademacher sketch ($S_{ij} \sim \frac{1}{\sqrt{m}}\text{Unif}\{+1, -1\}$), and the Fast Johnson-Lindenstrauss (JL) sketch [12, 7, 15]. Multiplying A with a dense matrix S can require $\mathcal{O}(mnd)$ time. But the Fast JL sketch introduced in [16] (based on a randomized Hadamard transform) can be computed efficiently in $\mathcal{O}(nd \log n)$ time. The Clarkson-Woodruff (CW) sketch [17] can be computed in just $\mathcal{O}(nd)$ time and has performance similar to the Fast JL sketch [10, 18].

The statistical perspective of sketching is not new, and has been considered in [18] for Gaussian, Fast JL and CW sketches, in [19] for sparse Rademacher sketches, and in [20] for leverage score sampling. [4] and [6] analyze asymptotic properties of various sampling and sketching methods. While for Gaussian sketches, it is possible to analyze statistical properties for finite n, d, the other sketches are analyzed only in asymptotic regimes. In fact, [18] and [19] show that the solutions from Fast JL, CW and sparse Rademacher sketches asymptotically converge to the Gaussian sketch estimator in distribution. Although our analysis of upper and lower bounds is based on the Gaussian sketch, using this result we

can argue that our results apply to a broader class of sketching matrices in the asymptotic regime ($n \to \infty$).

Error lower bounds on sketching for convex constrained least squares are shown in [5, 6, 13]. [13] gives lower bounds on the sketch size m for the residual error to be within $(1+\delta)$ times the residual error of the true solution. [5] provides a worst case prediction error lower bound using tools from hypothesis testing and Fano's inequality. [6] verifies that this lower bound is tight up to constants for Gaussian, Fast JL sketches and certain sampling methods. However that lower bound involves properties of the convex constraint set that are hard to evaluate, and constants that are not tight. In this work, we provide a lower bound for Gaussian sketches, which can be exactly evaluated. This is done by using a different technique for minimax lower bounds using the Fisher information, which was shown in [21] to give lower bounds with simpler analysis and precise constants. Another difference is that while the lower bound in [5] assumes a statistical model on the data, our lower bound is a worst case bound for any given data.

1.3 Organization

Section 2 provides the background about properties of the Gaussian sketch and the James-Stein estimator that will be useful to understand our results. Section 3 gives the two lower bound results. Section 4 gives the improved estimator, and an upper bound on its error, and analyzes the tightness of the upper and lower bounds. Section 5 contains empirical results on simulated and real data, on other sketching methods, and on alternate forms of the proposed estimator. Section 6 gives the conclusion and interesting directions opened by this work. Proofs of all theoretical results are given in Section 7.

2 Background

2.1 Gaussian Sketches for Least Squares Optimization

Consider an unconstrained least squares optimization problem

$$x_{LS} = \arg\min_{x \in \mathbb{R}^d} \|Ax - y\|_2^2 \tag{1}$$

where the data matrix $A \in \mathbb{R}^{n \times d}$ and the data vector $y \in \mathbb{R}^n$ are fixed and known. It is assumed that $n > d$. Throughout this project, we will assume that $\text{rank}(A) = d$, so that $A^T A$ is invertible. The classical Gaussian sketch is defined as

$$\widehat{x} = \arg\min_{x \in \mathbb{R}^d} \|SAx - Sy\|_2^2 \tag{2}$$

where A and y are replaced by low dimensional projections SA and Sy, where $S \in \mathbb{R}^{m \times n}$ has i.i.d. entries $S_{ij} \sim \frac{1}{\sqrt{m}}\mathcal{N}(0,1)$. Note that the sketching matrix S satisfies $\mathbb{E}[S^T S] = I_n$.

Lemma 1 *The classical Gaussian sketch $\widehat{x}$ satisfies*

$$\widehat{x} = x_{LS} + (SA)^\dagger S y^\perp \tag{3}$$

where $(SA)^\dagger = (A^T S^T SA)^{-1} A^T S^T$ is the pseudoinverse of SA and $y^\perp = y - Ax_{LS}$ such that $A^T y^\perp = 0$. $\left\|y^\perp\right\|_2^2$ is called the residual error. Further, $Sy^\perp$, and the columns of SA are independent multivariate Gaussians, and

$$\widehat{x} \sim \mathcal{N}\left(x_{LS}, \frac{1}{m}\left\|y^\perp\right\|_2^2 (A^T S^T SA)^{-1}\right) \mid SA \tag{4}$$

Here the notation $\mid SA$ means conditioning on the random variable SA. Also Lemma 1 assumes that $A^T S^T SA$ is invertible, which is true with probability 1, if $A^T A$ is invertible [22].

This property allows us to look at the least squares optimization problem as a statistical estimation problem of estimating the mean of a multivariate Gaussian from its sample, conditioned on the matrix SA. Note that this property holds irrespective of the distribution of the data, and is a property of the Gaussian sketch method. In particular this means that $\widehat{x}$ is an unbiased estimator of x_{LS} since $\mathbb{E}[\widehat{x}] = x_{LS}$.

The expected error of the classical Gaussian sketch can be evaluated exactly. This is known from earlier works as well [18], and is stated below again.

Lemma 2 *The prediction error of the classical Gaussian sketch $\widehat{x}$ is:*

$$\mathbb{E}\left[\|A(\widehat{x} - x_{LS})\|_2^2\right] = \frac{d}{m-d-1}\left\|y^{\perp}\right\|_2^2 \tag{5}$$

2.2 James-Stein Shrinkage Estimator

Suppose we are given $X \sim \mathcal{N}(\theta, I_d)$ where $\theta \in \mathbb{R}^d$ is an unknown mean vector. We want to design an estimator $\widehat{\theta}(X)$ such that it minimizes the expected squared error $R(\widehat{\theta}, \theta) = \mathbb{E}\|\widehat{\theta} - \theta\|_2^2$. The maximum likelihood estimator for θ is $\widehat{\theta}(X) = X$ which is an unbiased estimator, i.e. $\mathbb{E}\,\widehat{\theta} = \theta$. However, from [23] and [24], we have that the estimator given by

$$\widehat{\theta}(X) = \left(1 - \frac{d-2}{\|X\|_2^2}\right) X \tag{6}$$

results in a lower variance than the maximum likelihood estimator for $d > 2$ for any value of θ. In statistical language, $\widehat{\theta}(X) = X$ is an inadmissible estimator and (6) strictly dominates the former. This is a "shrinkage estimator" (as it shrinks the estimate towards 0), now known as the James-Stein estimator. This estimator has been very useful in statistics, see [25] for example. A simple proof and explanation of this result can be found in [26] and [27]. There are also several improvements and generalizations of this estimator [28, 29, 30, 31] to other shrinkage estimators. Particularly, we look at the generalization to the case where $X \sim \mathcal{N}(\theta, \Sigma)$.

Lemma 3 *Suppose $X \sim \mathcal{N}(\theta, \Sigma)$ where $\theta \in \mathbb{R}^d$ is the unknown mean with $d > 2$. Σ and a single sample X are known. Then consider the estimator for θ as*

$$\widehat{\theta}(X) = \left(1 - \frac{d-2}{X^T \Sigma^{-1} X}\right) X \tag{7}$$

Then, its expected normalized squared error is given by

$$\mathbb{E}[(\widehat{\theta} - \theta)^T \Sigma^{-1} (\widehat{\theta} - \theta)] = d - (d-2)^2 \mathbb{E}\left[\frac{1}{X^T \Sigma^{-1} X}\right] \tag{8}$$

If we used the estimator $\widehat{\theta}(X) = X$, then we would have $\mathbb{E}[(X - \theta)^T \Sigma^{-1} (X - \theta)] = d$ and the estimator shown above achieves strictly lesser expected error for any value of θ when $d > 2$. However, this is a biased estimator and the expected squared error will be lower when the true value of θ is close to zero, as compared to θ far from zero. More generally, we can also shrink the estimator towards any other point θ_0 and derive an estimator $\theta_0 + \left(1 - (d-2)/X^T \Sigma^{-1} X\right)(X - \theta_0)$, so that the error is smaller when θ is close to θ_0.

The above section motivates us to use the shrinkage estimator to estimate the true least squares solution x_{LS} from the solution $\widehat{x}$ of one sketched least squares problem.

3 Error Lower Bounds for Gaussian Sketch

This section presents lower bounds for estimators derived from the Gaussian sketch. The following two theorems give prediction error lower bounds for any estimator of x_{LS} that uses only the Gaussian sketched data SA and Sy. Theorem 4 is a lower bound for unbiased estimators only while Theorem 5 holds for biased estimators as well. All the proofs are given in Section 7.

3.1 Lower Bound on Error for Unbiased Estimators

Theorem 4 *For any unbiased estimator $\widetilde{x}$ of x_{LS}, (i.e. $\mathbb{E}[\widetilde{x}] = x_{LS}$) obtained from the Gaussian sketched data SA and Sy,*

$$\mathbb{E}\left[\|A(\widetilde{x} - x_{LS})\|_2^2\right] \geq \frac{d}{m-d-1}\left\|y^{\perp}\right\|_2^2 \tag{9}$$

From Lemma 2, the classical Gaussian sketch achieves the lower bound for unbiased estimators and is the minimum variance unbiased estimator of x_{LS} using only the Gaussian sketched data SA and Sy. Thus we have the tight lower bound

$$\inf_{\widetilde{x}(SA,Sy):\mathbb{E}[\widetilde{x}]=x_{LS}} \mathbb{E}\left[\|A(\widetilde{x} - x_{LS})\|_2^2\right] = \frac{d}{m-d-1}\left\|y^{\perp}\right\|_2^2, \tag{10}$$

where the infimum ranges over all unbiased estimators of x_{LS} that are functions of (SA, Sy). However, it can be seen ahead that when biased estimators are allowed, the classical Gaussian sketch is sub-optimal.

3.2 A More General Lower Bound

For the purpose of this theorem, consider a constrained least squares optimization problem

$$x_{LS} = \arg\min_{x \in \mathcal{C}} \|Ax - y\|_2^2 \tag{11}$$

where $\mathcal{C}$ is a convex constraint set. For the unconstrained problem, set $\mathcal{C} = \mathbb{R}^d$.

Theorem 5 *Suppose there exists $B > 0$ such that $[-B, B]^d \subseteq \mathcal{C}$. Then for any estimator $\widetilde{x}$ for x_{LS} obtained from the Gaussian sketched data SA and Sy,*

$$\mathbb{E}\left[\|A(\widetilde{x} - x_{LS})\|_2^2\right] \geq \frac{d}{m}\left\|y^{\perp}\right\|_2^2 \left(1 - \frac{\pi^2\left\|y^{\perp}\right\|_2^2}{mB^2\sigma_{min}(A^TA)}\right) \tag{12}$$

where $\sigma_{min}(A^TA)$ is the minimum eigenvalue of A^TA. In particular, for unconstrained least

squares optimization,

$$\mathbb{E}\left[\|A(\widetilde{x} - x_{LS})\|_2^2\right] \geq \frac{d}{m}\left\|y^{\perp}\right\|_2^2 \tag{13}$$

The proof will use a method based on the van Trees inequality which uses the Fisher information to lower bound the error [14]. This method has been inspired from [21] and it provides a neat lower bound in the constrained optimization case too. The lower bound in (13) must be seen as a worst case lower bound (since it holds for all possible values of x_{LS}). In practice, if we wish to impose any regularity conditions on the solution (such as small norm) based on prior knowledge, then we would have a constraint with $B < \infty$ and thus (12) would hold which can give us a lower error lower bound.

4 James-Stein Estimator for Least Squares

4.1 The Improved Estimator

Using the Gaussian distribution in (4) with $\theta = x_{LS}$ and $\Sigma = \frac{1}{m}\left\|y^{\perp}\right\|_2^2 (A^T S^T S A)^{-1}$, and the James-Stein estimator in (7), we get the following estimator for x_{LS}:

$$\widehat{x}_S = \left(1 - \frac{(d-2)\left\|y^{\perp}\right\|_2^2}{m\|SA\widehat{x}\|_2^2}\right)\widehat{x} \tag{14}$$

where $y^{\perp} = y - Ax_{LS}$ is as defined in Lemma 1 and $\widehat{x}$ is the solution of the classical Gaussian sketch (2).

However, $\left\|y^{\perp}\right\|^2 = \|y - Ax_{LS}\|_2^2$ cannot be computed directly without knowing the true solution x_{LS} itself. So we approximate $\left\|y^{\perp}\right\|_2^2$ as follows (an alternate approximation is shown in Section 5).

$$\left\|y^{\perp}\right\|_2^2 \approx \frac{m-d-1}{m-1}\|A\widehat{x} - y\|_2^2 \tag{15}$$

With this approximation, we use the following shrinkage estimator for x_{LS}:

$$\widehat{x}_{shr} = \left(1 - \frac{(d-2)(m-d-1)\|A\widehat{x} - y\|_2^2}{m(m-1)\|SA\widehat{x}\|_2^2}\right)\widehat{x} \tag{16}$$

4.2 Upper Bound on Error

Theorem 6 *(a) The approximation in (15) is an unbiased estimate, i.e.*

$$\mathbb{E}\left[\frac{m-d-1}{m-1}\|A\widehat{x} - y\|_2^2\right] = \|y^{\perp}\|_2^2 \tag{17}$$

(b) For $m > d+3$, the error of the shrinkage estimator (16) in the $\|SA(\cdot)\|_2$ norm is upper bounded as

$$\mathbb{E}\left[\|SA(\widehat{x}_{shr} - x_{LS})\|_2^2\right] \leq \frac{d}{m}\|y^{\perp}\|_2^2\left(1 - \frac{1-\epsilon'(d,m)}{1+(m/d)\|Ax_{LS}\|_2^2/\|y^{\perp}\|_2^2}\right) \tag{18}$$

where $\epsilon'(d,m) = \frac{4(d-1)}{d^2} + \frac{2(d-2)^2}{d(m-1)(m-d-3)}$.

(c) If $m > \mathcal{O}\left((d + \log(1/\delta))/\epsilon^2\right)$, then with probability greater than $1-\delta$, the $\|A(\cdot)\|_2$ norm of the error (prediction error) is upper bounded as

$$\mathbb{E}\left[\|A(\widehat{x}_{shr} - x_{LS})\|_2^2\right] \leq (1+\epsilon)\frac{d}{m}\|y^{\perp}\|_2^2\left(1 - \frac{1-\epsilon'(d,m)}{1+(m/d)\|Ax_{LS}\|_2^2/\|y^{\perp}\|_2^2}\right) \tag{19}$$

Part (c) follows from part (b) since the Gaussian sketching matrix is an Oblivious Subspace Embedding [32].

To examine this upper bound in more detail, we look at the ratio of the prediction errors due to the shrinkage estimator and the classical Gaussian sketch.

$$\frac{\mathbb{E}\left[\|A(\widehat{x}_{shr} - x_{LS})\|_2^2\right]}{\mathbb{E}\left[\|A(\widehat{x} - x_{LS})\|_2^2\right]} \leq (1+\epsilon)\frac{m-d-1}{m}\left(1 - \frac{1-\epsilon'(d,m)}{1+(m/d)\|Ax_{LS}\|_2^2/\|y^{\perp}\|_2^2}\right) = R(A,y) \tag{20}$$

Clearly, $R(A,y) < (1+\epsilon)$. But for large enough m, we expect to have $R(A,y) < 1$ for any A, y. (Through empirical results we will see that this happens for m that is not much larger than d). As m becomes very large, $R(A,y) \to 1$, indicating that the error of the shrinkage estimator approaches the classical Gaussian sketch as m becomes large.

The factor $\rho(A,y) = \|Ax_{LS}\|_2^2 / \|y^\perp\|_2^2 = \|Ax_{LS}\|_2^2 / \|Ax_{LS} - y\|_2^2$ can be interpreted as a signal-to-noise ratio (SNR). $R(A,y)$ is smaller when $\rho(A,y)$ is small, i.e. the original least squares problem is more noisy. The shrinkage factor in (16) also depends on the factor $\|SA\widehat{x}\|_2^2 / \|A\widehat{x} - y\|_2^2$ which can be regarded as an approximation of $\rho(A,y)$. When this ratio is large (with respect to d/m), the shrinkage estimator will be very close to the classical Gaussian sketch.

Note that this upper bound can be smaller than the lower bound (13) because it is a worst case lower bound, and the worst case occurs when $\rho(A,y) \to \infty$. Indeed there cannot be any estimator that beats the lower bound in (13) for all values of A, y.

4.3 Matching the Upper and Lower Bounds

The lower bound in Theorem 5 and the upper bound in Theorem 6 have very similar forms and we show that the shrinkage estimator achieves this lower bound under certain conditions.

Suppose we know that the true solution x_{LS} satisfies the SNR condition $\rho(A,y) \le \eta^2$. If $\|x_{LS}\|_\infty^2 \le \eta^2 \|y^\perp\|_2^2 / d\sigma_{max}(A^TA)$,

$$\rho(A,y) = \frac{\|Ax_{LS}\|_2^2}{\|y^\perp\|_2^2} \le \frac{\sigma_{max}(A^TA)\|x_{LS}\|_2^2}{\|y^\perp\|_2^2} \le \frac{d\sigma_{max}(A^TA)\|x_{LS}\|_\infty^2}{\|y^\perp\|_2^2} \le \eta^2$$

So we can choose $B^2 = \eta^2 \|y^\perp\|_2^2 / d\sigma_{max}(A^TA)$ in the lower bound (12). The lower bound becomes

$$\mathbb{E}\left[\|A(\widetilde{x} - x_{LS})\|_2^2\right] \ge \frac{d}{m}\|y^\perp\|_2^2 \left(1 - \frac{d\pi^2\sigma_{max}(A^TA)}{m\eta^2\sigma_{min}(A^TA)}\right)$$

If $d \gg 1$ and $m \gg d$, we ignore $\epsilon'(d,m)$ and the 1 in the denominator and the upper

bound (19) now becomes:

$$\mathbb{E}\left[\|A(\widehat{x}_{shr} - x_{LS})\|_2^2\right] \leq (1+\epsilon)\frac{d}{m}\left\|y^{\perp}\right\|_2^2\left(1 - \frac{d}{m\eta^2}\right)$$

Comparing this with the lower bound in (12), it is achieved for large m upto the constant of π^2 in the lower bound, when:

1. we know that $\rho(A, y) \leq \eta^2$ (or the corresponding upper bound on the ℓ^2 or ℓ^∞ norm),
2. d and m are large enough so that d/m and ϵ are small, and
3. the data matrix A is well-conditioned so that $\sigma_{max}(A^TA) \approx \sigma_{min}(A^TA)$.

Condition 1 can be seen as a regularity condition on the solution which is typically desired. As will be seen in the empirical results, Condition 2 can be satisfied for m that is not too large. Thus the only data-dependent condition is Condition 3.

5 Empirical Results

The decrease in prediction error due to the shrinkage estimator will be demonstrated empirically on both simulated and real datasets. The behaviour of the shrinkage estimator is examined for different SNR values of the data, and also compared with the upper and lower bounds. In all the following plots, the (normalized) prediction error refers to $\frac{1}{n}\|A(\widehat{x} - x_{LS})\|_2^2$ (unless otherwise specified).

5.1 Results on Simulated Data

Figure 1 compares the prediction error for the classical Gaussian sketch and the shrinkage estimator on simulated data. For these experiments, we generate each row of A from a multivariate Gaussian distribution with mean $\mu_i = 1$ and covariance $\Sigma_{ij} = (0.5)^{|i-j|}$. We choose x_{LS} as an i.i.d. random normal vector and normalize such that $\|Ax_{LS}\|_2 = 1$. Then

we choose $y = Ax_{LS} + \alpha V w$ where V is a basis for the null space of A^T and $w \sim \mathcal{N}(0, \sigma^2 I_{n-d})$ (Gaussian noise model) with $n = 1024$ and $d = 100$. By choosing $\alpha = (\sqrt{\rho}\|Vw\|_2)^{-1}$, we set the SNR $\rho(A, y) = \rho$ for different values of ρ. This will be referred to as **Gaussian data**.

In Figure 1, the shrinkage estimator achieves much lower error, and the decrease is most noticeable at lower sketch sizes. The decrease in error due to the shrinkage estimator is not much when the SNR is large (Fig. 1c), but it is considerable when the SNR is small (Fig. 1a). These plots also show the lower bound for unbiased estimators (9) and that the classical Gaussian sketch achieves the lower bound. Figure 1d compares the errors $\frac{1}{n}\|A(\widehat{x} - x_{LS})\|_2^2$ and $\frac{1}{n}\|SA(\widehat{x} - x_{LS})\|_2^2$ for the shrinkage estimator, showing that the $(1+\epsilon)$ factor in (19) becomes close to 1 for a sketching dimension that is not too large. The $\frac{1}{n}\|SA(\widehat{x} - x_{LS})\|_2^2$ error is also compared with the upper bound in (18), showing that the upper bound matches the empirical error.

5.2 Shrinkage Estimator on other Sketching Methods

Although the analysis for the shrinkage estimator has been done only for the Gaussian sketch, it is seen to work equally well when applied to the estimate obtained from other types of sketching matrices such as uniform row sampling and the Fast JL transform (Figure 2). In fact, it also works equally well for the CW Sketch, Rademacher sketch and sampling based on row norms or leverage scores (results not shown here).

5.3 Alternate Approximation to (15)

Instead of approximating $\left\|y^{\perp}\right\|_2^2$ using (15), we can also use the following approximation:

$$\left\|y^{\perp}\right\|_2^2 \approx \frac{m}{m-d}\|SA\widehat{x} - Sy\|_2^2 \tag{21}$$

This is also an unbiased estimator (see proof of Theorem 6), and is useful when only the sketched data SA and Sy can be be observed but not the original data A and y. This

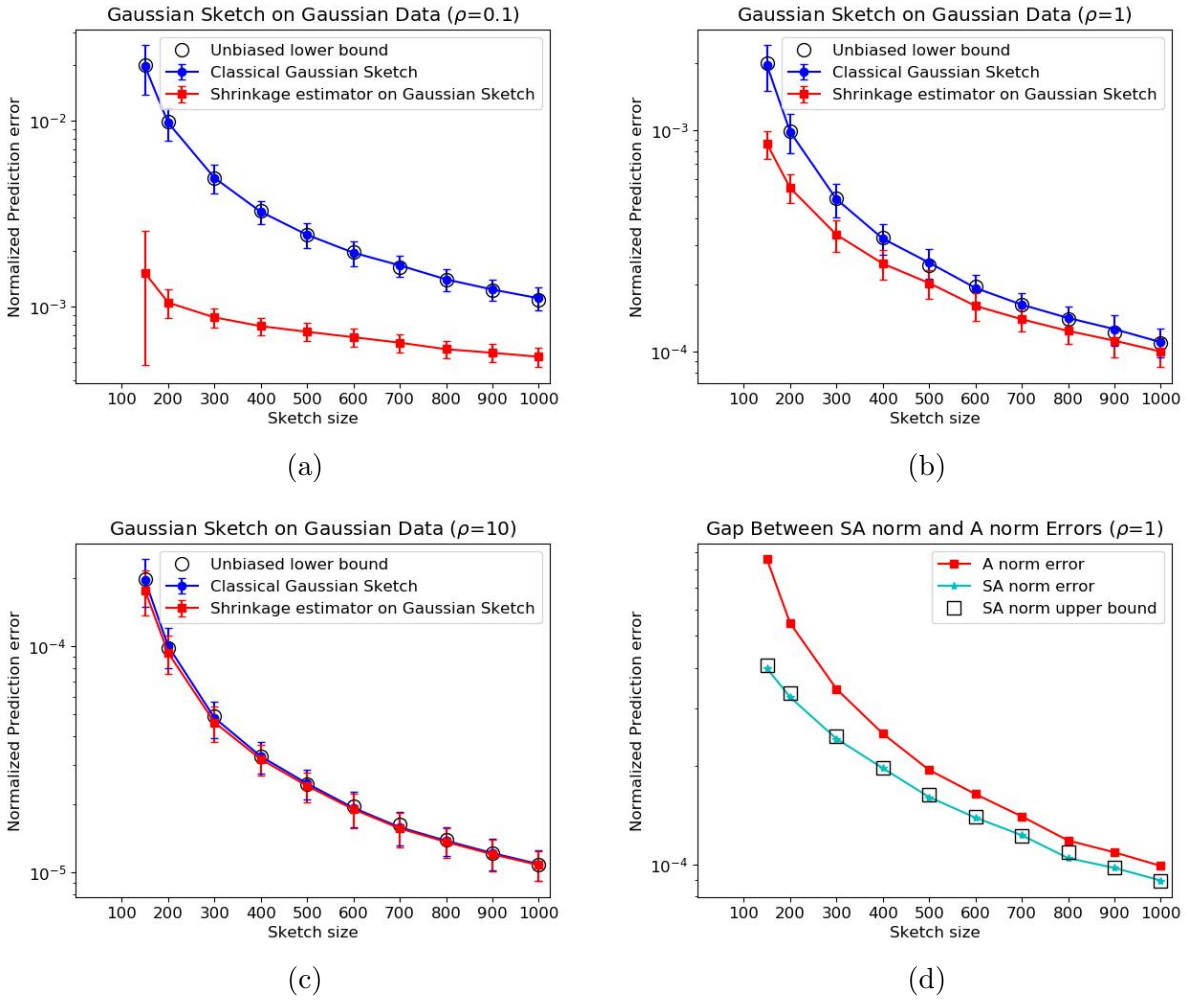
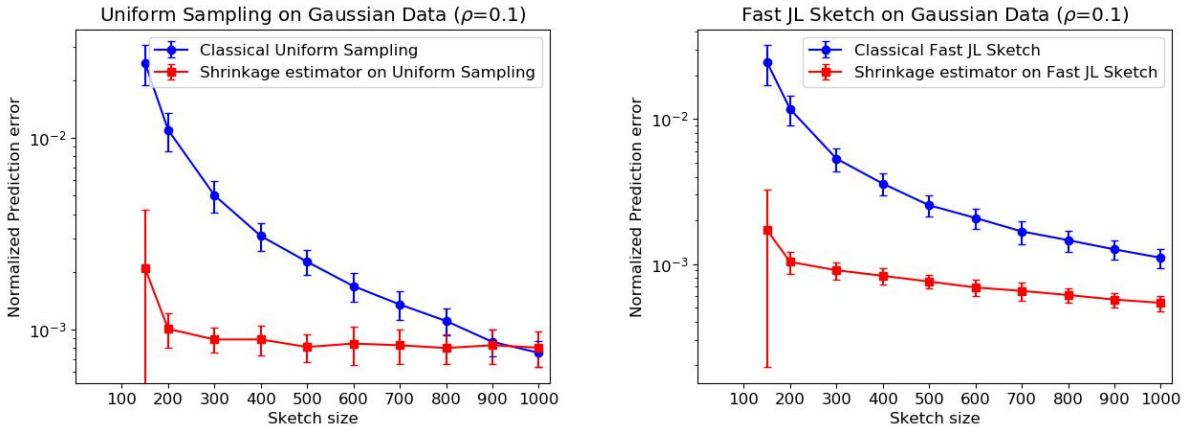


Figure 1: Shrinkage estimator applied to Gaussian data. (a-c): Error $\frac{1}{n}\|A(\widehat{x} - x_{LS})\|_2^2$ for the classical Gaussian sketch as well as the shrinkage estimator, versus the sketch size (m), for different values of $\rho(A, y)$. (d): Comparison of the errors $\frac{1}{n}\|A(\widehat{x}_{shr} - x_{LS})\|_2^2$ and $\frac{1}{n}\|SA(\widehat{x}_{shr} - x_{LS})\|_2^2$, and the upper bound (18). All the plots are averages over 100 independent sketches and the error bars show the standard deviation on each side.

can be the case when the measurement system itself gives a compressed measurement, or in scenarios with memory constraints. This gives us the alternate estimator

$$\widehat{x}_{alt} = \left(1 - \frac{(d-2)\|SA\widehat{x} - Sy\|_2^2}{(m-d)\|SA\widehat{x}\|_2^2}\right)\widehat{x} \tag{22}$$

Figure 3a shows that the true James-Stein estimator (14), the original approximation (16) and the alternate approximation (22) achieve similar expected errors.



(a) (b)

Figure 2: Comparing the classical uniform sampling and Fast JL Transform based sketches with the shrinkage estimator applied on them. All the plots are averages over 100 independent sketches and the error bars show the standard deviation on each side.

5.4 Positive Part Shrinkage Estimator

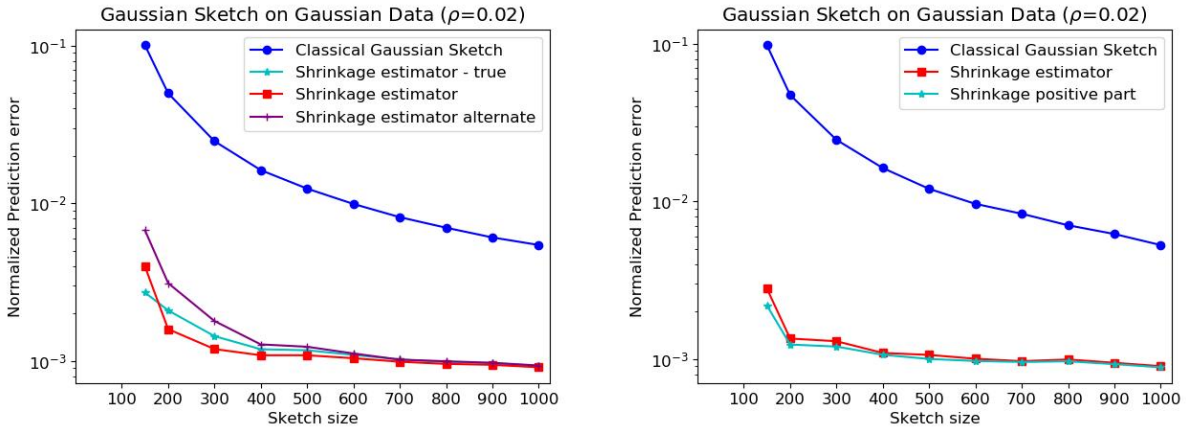
The James-Stein estimator shrinks the estimate of the mean towards 0. However it is possible that the shrinkage factor can be too large causing the estimator to switch signs. To avoid this, we can use a positive-part James-Stein estimator which is known to have strictly lower expected squared error than the James-Stein estimator too.

$$\widehat{x}_{pp} = \max\left\{\left(1 - \frac{(d-2)(m-d-1)\|A\widehat{x} - y\|_2^2}{m(m-1)\|SA\widehat{x}\|_2^2}\right), 0\right\}\widehat{x} \tag{23}$$

In these experiments, it was seen that there is an improvement in the error only at very low values of $\rho(A, y)$ (Figure 3b).

5.5 Results on Real Data

We present empirical results on 4 real datasets. The CPU dataset ($n = 8192, d = 12$) and the Years dataset ($n = 463715, d = 90$) [33] are regression datasets. We compare the shrinkage estimator and classical Gaussian sketch for both these datasets with varying levels of added noise, i.e. we add Gaussian noise of variance $\kappa\left\|y^{\perp}\right\|_2^2$ to y to vary the SNR ρ. The results



(a) (b)

Figure 3: (a): Comparison of the true James-Stein estimator, the proposed shrinkage estimator and the alternate approximation. (b): Comparison of the shrinkage estimator and the positive part estimator. All the plots are averages over 100 independent sketches

are seen in Figure 4a,b for CPU data and in Figure 4c,d for the Years data. The shrinkage estimator performs much better than the classical sketch when the data is very noisy and when $n \gg d$, and performs equally well as the classical sketch otherwise.

We also show results for least-squares based classification on the MNIST and SVHN datasets. For SVHN ($n = 20000, d = 3072$), we classify between two digits using vectorized images in the rows of A and 0-or-1 labels in y. For MNIST ($n = 48000, d = 1000$), we perform a multiclass classification with vectorized images passed through a randomly initialized ReLU layer in the rows of $A \in \mathbb{R}^{n\times d}$, and $Y \in \mathbb{R}^{n\times 10}$ is a matrix with one-hot vector labels. Thus we have a Frobenius norm regression problem: minimize $\|AX - Y\|_F^2$. With $\widehat{X}$ as the solution of the sketched version (minimize $\|SAX - SY\|_F^2$), the shrinkage estimator in this case is simply

$$\widehat{X}_{shr} = \left(1 - \frac{(d-2)(m-d-1)\left\|A\widehat{X} - Y\right\|_F^2}{m(m-1)\left\|SA\widehat{X}\right\|_F^2}\right)\widehat{X} \tag{24}$$

This estimator satisfies the same error upper bound (19) with the vector norms replaced by the Frobenius norms.

Figure 5 shows the prediction error on these two datasets. Since $\|A\widehat{x} - y\|_2^2 =$

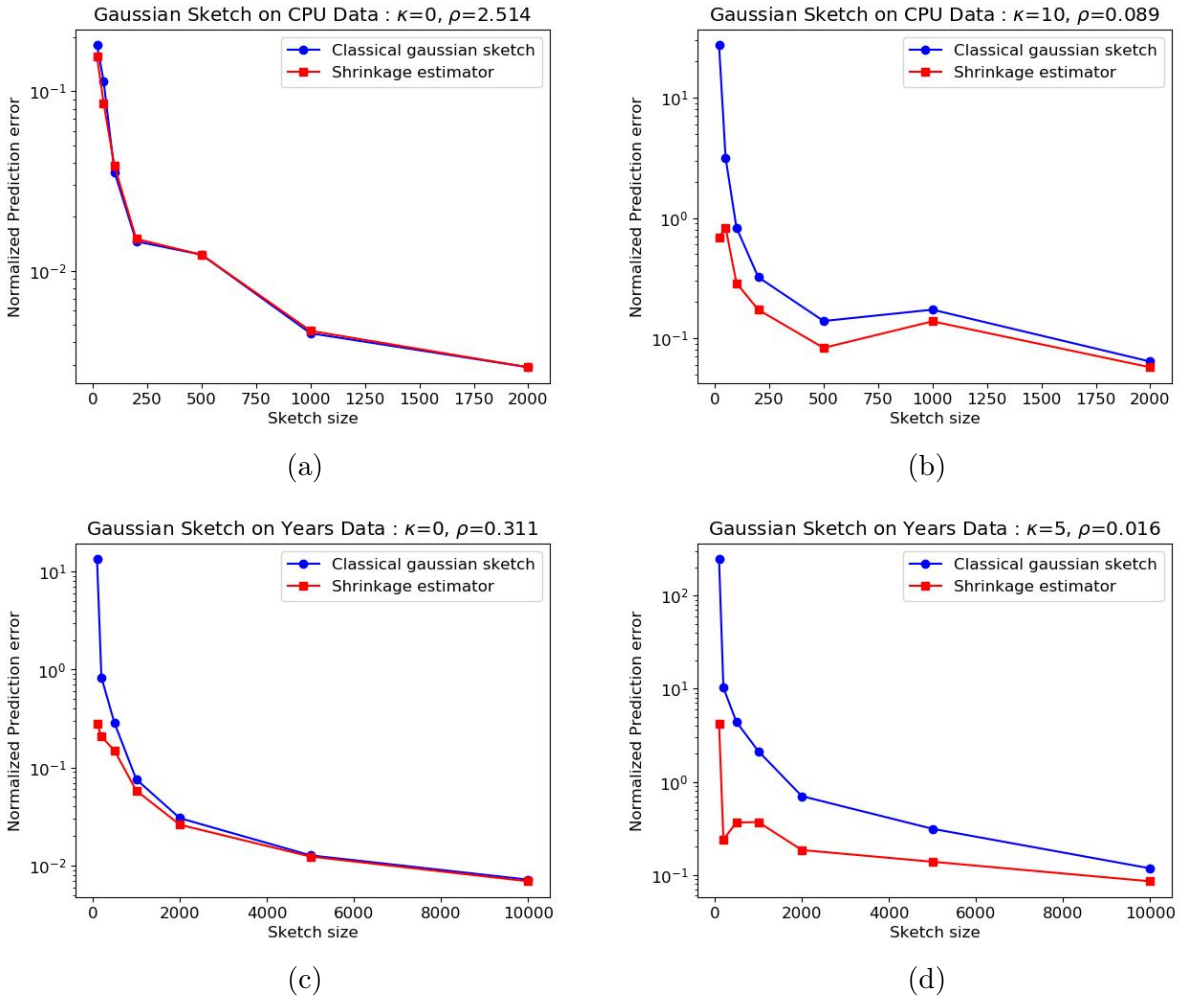
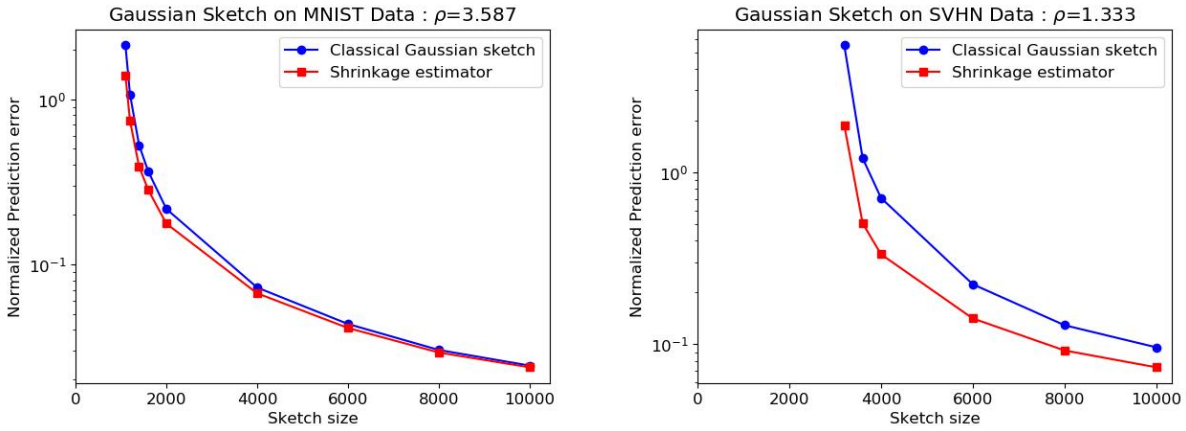


Figure 4: Results of shrinkage estimator and classical Gaussian sketch on the CPU data ($n = 8192, d = 12$) and Years data ($n = 463715, d = 90$). κ is the variance of the added noise scaled by $\left\|y^{\perp}\right\|_2^2$ and ρ is the SNR.

$\|A(\widehat{x} - x_{LS})\|_2^2 + \|Ax_{LS} - y\|_2^2$, the mean squared error loss is only a constant added to the prediction error. While the SNR of MNIST is high and the shrinkage estimator performs only slightly better than the classical Gaussian sketch, the shrinkage estimator performs much better than the classical Gaussian sketch in SVHN.



(a) (b)

Figure 5: Results of shrinkage estimator and classical Gaussian sketch on the MNIST ($n = 48000, d = 1000$) and SVHN data ($n = 20000, d = 3072$). ρ is the SNR.

6 Conclusion and Further Directions

Applying the James-Stein estimator to sketched least squares optimization results in smaller prediction error, especially when the data is more noisy. The estimator is easy to compute based on the classical sketching methods, and can allow us to reduce the sketch sizes greatly. The analysis is interesting as it looks at an optimization problem as a statistical estimation problem. This idea works irrespective of the distribution of the data. Empirically it works for many common sketching methods as well. From prior results that show that common sketching solutions asymptotically converge to the Gaussian sketch solution in distribution [18, 19], we can argue that our results apply to those sketches as well.

The lower bounds derived in this project are new, and easy to compute, unlike existing ones such as in [5]. We proved that the classical sketching method is the optimum unbiased estimator based on the Gaussian sketch. Additionally our lower bound can incorporate prior knowledge about the solution through constraints. The error of the shrinkage estimator can be made arbitrarily close to the lower bound for large m and d, when the data is well conditioned (upto the factor of π^2), so the bounds are tight.

Since the shrinkage estimator shrinks the sketched solution towards 0, it can be regarded

as a regularizer. To the best of our knowledge, there are no results showing whether an ℓ^2 norm regularized sketching solution strictly dominates the classical sketching solution from a statistical perspective. But our shrinkage estimator is shown to strictly dominate the classical Gaussian sketch solution. It would be interesting to study the generalization properties of the shrinkage estimator proposed here, and compare it with standard regularization methods.

Another interesting direction is to apply such shrinkage estimators for right sketching methods for the minimum norm solution of underdetermined equations. Since the minimum norm solution also has the same form and has a similar distribution under Gaussian sketch, such an estimator would give lower error in that case as well.

References

[1] J. Blocki, A. Blum, and O. Sheffet, "The johnson-lindenstrauss transform itself preserves differentialprivacy," 04 2012.

[2] M. Showkatbakhsh, C. Karakus, and S. Diggavi, "Privacy-utility trade-off of linear regression under random projections and additive noise," 06 2018, pp. 186–190.

[3] S. Zhou, K. Ligett, and L. Wasserman, "Differential privacy with compression," in *2009 IEEE International Symposium on Information Theory*, 2009, pp. 2718–2722.

[4] E. Dobriban and S. Liu, "Asymptotics for sketching in least squares regression," in *Advances in Neural Information Processing Systems*, 2019, pp. 3670–3680.

[5] M. Pilanci and M. J. Wainwright, "Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares," 2014.

[6] G. Raskutti and M. W. Mahoney, "A statistical perspective on randomized sketching for ordinary least-squares," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 7508–7538, 2016.

[7] M. W. Mahoney *et al.*, "Randomized algorithms for matrices and data," *Foundations and Trends® in Machine Learning*, vol. 3, no. 2, pp. 123–224, 2011.

[8] T. Sarlos, "Improved approximation algorithms for large matrices via random projections," in *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, 2006, pp. 143–152.

[9] S. S. Vempala, *The random projection method*. American Mathematical Soc., 2005, vol. 65.

[10] D. P. Woodruff, "Sketching as a tool for numerical linear algebra," *CoRR*, vol. abs/1411.4357, 2014. [Online]. Available: http://arxiv.org/abs/1411.4357

[11] P. Drineas, M. W. Mahoney, and S. Muthukrishnan, "Sampling algorithms for l2 regression and applications," in *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, ser. SODA 06. USA: Society for Industrial and Applied Mathematics, 2006, p. 11271136.

[12] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós, "Faster least squares approximation," *Numerische mathematik*, vol. 117, no. 2, pp. 219–249, 2011.

[13] M. Pilanci and M. J. Wainwright, "Randomized sketches of convex programs with sharp guarantees," *IEEE Transactions on Information Theory*, vol. 61, no. 9, pp. 5096–5115, 2015.

[14] R. D. Gill and B. Y. Levit, "Applications of the van trees inequality: a bayesian cramr-rao bound," *Bernoulli*, vol. 1, no. 1-2, pp. 59–79, 03 1995. [Online]. Available: https://projecteuclid.org:443/euclid.bj/1186078362

[15] C. Boutsidis and P. Drineas, "Random projections for the nonnegative least-squares problem," *Linear algebra and its applications*, vol. 431, no. 5-7, pp. 760–771, 2009.

[16] N. Ailon and B. Chazelle, "Approximate nearest neighbors and the fast johnson-lindenstrauss transform," in *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*, ser. STOC 06. New York, NY, USA: Association for Computing Machinery, 2006, p. 557563. [Online]. Available: https://doi.org/10.1145/1132516.1132597

[17] K. L. Clarkson and D. P. Woodruff, "Low-rank approximation and regression in input sparsity time," *Journal of the ACM (JACM)*, vol. 63, no. 6, pp. 1–45, 2017.

[18] D. Ahfock, W. J. Astle, and S. Richardson, "Statistical properties of sketching algorithms," 2017.

[19] P. Li, T. J. Hastie, and K. W. Church, "Very sparse random projections," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 287–296.

[20] P. Ma, M. W. Mahoney, and B. Yu, "A statistical perspective on algorithmic leveraging," *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 861–911, 2015.

[21] L. P. Barnes, Y. Han, and A. Özgür, "Learning distributions from their samples under communication constraints," *CoRR*, vol. abs/1902.02890, 2019. [Online]. Available: http://arxiv.org/abs/1902.02890

[22] A. K. Gupta and D. K. Nagar, *Matrix variate distributions*. Chapman and Hall/CRC, 2018, ch. 3.

[23] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, 1956, pp. 197–206. [Online]. Available: https://projecteuclid.org/euclid.bsmsp/1200501656

[24] W. James and C. Stein, "Estimation with quadratic loss," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Berkeley, Calif.: University of California Press, 1961, pp. 361–379. [Online]. Available: https://projecteuclid.org/euclid.bsmsp/1200512173

[25] B. Efron and C. Morris, "Stein's paradox in statistics," *Scientific American - SCI AMER*, vol. 236, pp. 119–127, 05 1977.

[26] K. H. Yung, "Explaining the stein paradox," 1999.

[27] R. J. Samworth and S. Cambridge, "Steins paradox."

[28] B. Efron and C. Morris, "Multivariate empirical bayes and estimation of covariance matrices," *Ann. Statist.*, vol. 4, no. 1, pp. 22–32, 01 1976.

[29] B. E. Hansen, "Generalized shrinkage estimators," *manuscript, University of Wisconsin*, 2008.

[30] P. D. Hoff, "Shrinkage estimators," *Course notes. University of Washington*, 2012.

[31] P. Y.-S. Shao and W. E. Strawderman, "Improving on the james-stein positive-part estimator," *The Annals of Statistics*, pp. 1517–1538, 1994.

[32] J. Nelson, *Oblivious Subspace Embeddings*. New York, NY: Springer New York, 2016, pp. 1430–1434.

[33] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, May 2011. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

[34] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 9, no. 6, pp. 1135–1151, 11 1981.

[35] R. C. Rogers, *The calculus of several variables*, 2011, p. 289.

[36] A. A. Borovkov, *Mathematical statistics*. Australia: Gordon and Breach Science Publishers, 1998.

7 Proofs

7.1 Proof of Lemma 1

The solution x_{LS} of (1) satisfies $\nabla\|Ax - y\|_2^2 = 0$, i.e. $A^T(Ax_{LS} - y) = 0$. By solving this equation we get $x_{LS} = (A^TA)^{-1}A^Ty = A^\dagger y$ (assuming that A is full column rank, so that A^TA is invertible). By setting $y^\perp = y - Ax_{LS}$, we get $A^Ty^\perp = 0$. Similarly, the sketched solution $\widehat{x}$ of (2) satisfies $(SA)^T(SA\widehat{x} - Sy) = 0$. Solving this, we get

$$\begin{aligned}
\widehat{x} &= (A^TS^TSA)^{-1}A^TS^TSy \\
&= (A^TS^TSA)^{-1}A^TS^TS(Ax_{LS} + y^\perp) \\
&= x_{LS} + (A^TS^TSA)^{-1}A^TS^TSy^\perp \\
&= x_{LS} + (SA)^\dagger Sy^\perp
\end{aligned}$$

Since each element of SA and $Sy^\perp$ is a sum of independent Gaussians, each column of SA, and $Sy^\perp$ follow a multivariate Gaussian distribution. In particular, $Sy^\perp \sim \mathcal{N}(0, \frac{1}{m}\|y^\perp\|_2^2 I_m)$. Further,

$$\mathbb{E}[(SA)^TSy^\perp] = \mathbb{E}[A^TS^TSy^\perp] = A^T\mathbb{E}[S^TS]y^\perp = A^Ty^\perp = 0 \quad \text{since } \mathbb{E}[S^TS] = I$$

Thus, SA and $Sy^\perp$ are uncorrelated Gaussian random variables, so they are independent too. Finally, we note that conditioned on SA, $\widehat{x}$ must have a Gaussian distribution (since

$Sy^\perp$ is Gaussian) with $\mathbb{E}[\widehat{x} \mid SA] = x_{LS}$ since $\mathbb{E}[Sy^\perp \mid SA] = \mathbb{E}[Sy^\perp] = 0$, and

$$\begin{aligned}
\mathrm{Cov}[\widehat{x} \mid SA] &= \mathbb{E}[(\widehat{x} - x_{LS})(\widehat{x} - x_{LS})^T \mid SA] \\
&= (SA)^\dagger \mathbb{E}[(Sy^\perp)(Sy^\perp)^T \mid SA](SA)^{\dagger T} \\
&= (SA)^\dagger \mathrm{Cov}[Sy^\perp \mid SA](SA)^{\dagger T} \\
&= (SA)^\dagger \mathrm{Cov}[Sy^\perp](SA)^{\dagger T} \\
&= (A^T S^T SA)^{-1} A^T S^T \left(\frac{1}{m} \left\| y^\perp \right\|_2^2 I_m \right) SA (A^T S^T SA)^{-1} \\
&= \frac{1}{m} \left\| y^\perp \right\|_2^2 (A^T S^T SA)^{-1}
\end{aligned}$$

■

7.2 Proof of Lemma 2

From Lemma 1,

$$\widehat{x} \sim \mathcal{N}\left(x_{LS}, \frac{1}{m} \left\| y^\perp \right\|_2^2 (A^T S^T SA)^{-1} \right) \mid SA$$

Hence we have

$$\begin{aligned}
\mathbb{E}\left[\left\| A(\widehat{x} - x_{LS}) \right\|_2^2 \mid SA \right] &= \mathbb{E}\,\mathrm{tr}\left[A(\widehat{x} - x_{LS})(\widehat{x} - x_{LS})^T A^T \mid SA \right] \\
&= \mathrm{tr}\left[A \mathbb{E}\left[(\widehat{x} - x_{LS})(\widehat{x} - x_{LS})^T \mid SA \right] A^T \right] \\
&= \mathrm{tr}\left[A \left(\frac{1}{m} \left\| y^\perp \right\|_2^2 (A^T S^T SA)^{-1} \right) A^T \right]
\end{aligned}$$

Taking expectation with respect to SA on both sides,

$$\begin{aligned}
\mathbb{E}\left[\left\| A(\widehat{x} - x_{LS}) \right\|_2^2 \right] &= \mathbb{E}_{SA}\,\mathrm{tr}\left[A \left(\frac{1}{m} \left\| y^\perp \right\|_2^2 (A^T S^T SA)^{-1} \right) A^T \right] \\
&= \frac{1}{m} \left\| y^\perp \right\|_2^2 \mathrm{tr}\left[A \mathbb{E}_{SA}\left[(A^T S^T SA)^{-1} \right] A^T \right]
\end{aligned}$$

Since each row of SA has an independent multivariate Gaussian distribution $\mathcal{N}\left(0, \frac{1}{m} A^T A\right)$,

$A^T S^T S A$ follows a Wishart distribution $\mathcal{W}_d\left(m, \frac{1}{m}A^T A\right)$, and so its inverse follows an inverted Wishart distribution that satisfies $\mathbb{E}\left[(A^T S^T S A)^{-1}\right] = \frac{m}{m-d-1}(A^T A)^{-1}$ [22].

$$\begin{aligned}
\mathbb{E}\left[\|A(\widehat{x} - x_{LS})\|_2^2\right] &= \frac{1}{m}\left\|y^{\perp}\right\|_2^2 \mathrm{tr}\left[A\frac{m}{m-d-1}(A^T A)^{-1}A^T\right] \\
&= \frac{1}{m-d-1}\left\|y^{\perp}\right\|_2^2 \mathrm{tr}\left[A(A^T A)^{-1}A^T\right] \\
&= \frac{1}{m-d-1}\left\|y^{\perp}\right\|_2^2 \mathrm{tr}\left[(A^T A)^{-1}A^T A\right] \\
&= \frac{d}{m-d-1}\left\|y^{\perp}\right\|_2^2
\end{aligned}$$

■

7.3 Proof of Lemma 3

The proof is a consequence of Stein's lemma from [34] which can be proved using the integration by parts formula for multivariable calculus. With $\widehat{\theta}$ as in (7),

$$\begin{aligned}
\mathbb{E}[(\widehat{\theta} - \theta)^T \Sigma^{-1}(\widehat{\theta} - \theta)] &= \mathbb{E}\left[\left(X - \theta - \frac{(d-2)X}{X^T\Sigma^{-1}X}\right)^T \Sigma^{-1}\left(X - \theta - \frac{(d-2)X}{X^T\Sigma^{-1}X}\right)\right] \\
&= \mathbb{E}[(X-\theta)^T\Sigma^{-1}(X-\theta)] - 2(d-2)\mathbb{E}\left[\frac{(X-\theta)^T\Sigma^{-1}X}{X^T\Sigma^{-1}X}\right] \\
&\quad + (d-2)^2\mathbb{E}\left[\frac{1}{X^T\Sigma^{-1}X}\right]
\end{aligned}$$

The first term can be simplified as

$$\begin{aligned}
\mathbb{E}[(X-\theta)^T\Sigma^{-1}(X-\theta)] &= \mathbb{E}\,\mathrm{tr}[(X-\theta)^T\Sigma^{-1}(X-\theta)] \\
&= \mathbb{E}\,\mathrm{tr}[\Sigma^{-1}(X-\theta)(X-\theta)^T] \\
&= \mathrm{tr}(\Sigma^{-1}\mathbb{E}[(X-\theta)(X-\theta)^T]) \\
&= \mathrm{tr}(\Sigma^{-1}\Sigma) \\
&= d
\end{aligned}$$

For the second term, we use the following identity from multivariable calculus [35]:

$$\int_\Omega u\,\nabla\cdot\mathbf{V}\,d\Omega \;=\; \int_\Gamma u\mathbf{V}\cdot\hat{\mathbf{n}}\,d\Gamma - \int_\Omega \nabla(u)\cdot\mathbf{V}\,d\Omega$$

where $\Omega \subseteq \mathbb{R}^d$, $u:\Omega\to\mathbb{R}$, $\mathbf{V}:\Omega\to\mathbb{R}^d$, Γ is the boundary of Ω and $\hat{\mathbf{n}}$ is the outward normal to Γ. $\nabla(u)$ is the gradient of u and $\nabla\cdot\mathbf{V}$ is the divergence of $\mathbf{V}$.

We choose $\Omega=\mathbb{R}^d$, $u=f(x)$ as the density function of $\mathcal{N}(\theta,\Sigma)$ such that $\nabla(u) = -f(x)\Sigma^{-1}(x-\theta)$, and $V=\frac{x}{x^T\Sigma^{-1}x}$. Since the density function goes to zero at the boundary of Ω, the first integral on the right side is zero.

$$\begin{aligned}
\mathbb{E}\left[\frac{(X-\theta)^T\Sigma^{-1}X}{X^T\Sigma^{-1}X}\right] &= \int_\Omega f(x)\frac{(x-\theta)^T\Sigma^{-1}x}{x^T\Sigma^{-1}x}\,d\Omega \\
&= \int_\Omega -\nabla f(x)\cdot\frac{x}{x^T\Sigma^{-1}x}\,d\Omega \\
&= \int_\Omega f(x)\,\nabla\cdot\left(\frac{x}{x^T\Sigma^{-1}x}\right)\,d\Omega \\
&= \mathbb{E}\left[\nabla\cdot\left(\frac{X}{X^T\Sigma^{-1}X}\right)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{d}\frac{X^T\Sigma^{-1}X - 2X_i(\Sigma^{-1}X)_i}{(X^T\Sigma^{-1}X)^2}\right] \\
&= (d-2)\mathbb{E}\left[\frac{1}{X^T\Sigma^{-1}X}\right]
\end{aligned}$$

Putting these parts together, we get

$$\mathbb{E}[(\widehat{\theta}-\theta)^T\Sigma^{-1}(\widehat{\theta}-\theta)] = d-(d-2)^2\mathbb{E}\left[\frac{1}{X^T\Sigma^{-1}X}\right]$$

■

7.4 Proof of Theorem 4

We first note by the definition of $y^\perp$ that

$$\widetilde{y} = Sy = SAx_{LS} + Sy^\perp$$

where SAx_{LS} and $Sy^\perp$ are both Gaussian random vectors, and are independent to each other (from Lemma 1). So, conditioned on SA, $\widetilde{y}$ is also a Gaussian random vector with mean SAx_{LS} and the vector $Sy^\perp$ acting as independent, zero-mean Gaussian noise, i.e.

$$\widetilde{y} \sim \mathcal{N}\left(SAx_{LS}, \frac{1}{m}\left\|y^\perp\right\|_2^2 I_m\right) \mid SA \tag{25}$$

Our problem is to estimate the parameter x_{LS} in the mean of this Gaussian distribution. Let $f(\widetilde{y}; x_{LS})$ be the density function of the Gaussian shown above, parameterized by x_{LS}. Then we define the Fisher information matrix for estimation of x_{LS} from $\widetilde{y}$ as

$$I(\widetilde{y}; x_{LS}) = \mathbb{E}\left[\nabla_{x_{LS}} \log f(\widetilde{y}; x_{LS}) \nabla_{x_{LS}} \log f(\widetilde{y}; x_{LS})^T\right]$$

Since $\log f(\widetilde{y}; x_{LS}) = -(1/2)(\widetilde{y} - SAx_{LS})^T \left(\frac{1}{m}\left\|y^\perp\right\|_2^2 I_m\right)^{-1} (\widetilde{y} - SAx_{LS}) +$ constant,

$$\begin{aligned} I(\widetilde{y}; x_{LS}) &= \mathbb{E}\left[(SA)^T \left(\frac{1}{m}\left\|y^\perp\right\|_2^2 I_m\right)^{-1} (\widetilde{y} - SAx_{LS})(\widetilde{y} - SAx_{LS})^T \left(\frac{1}{m}\left\|y^\perp\right\|_2^2 I_m\right)^{-1} SA\right] \\ &= \frac{m}{\left\|y^\perp\right\|_2^2} A^T S^T SA \quad \text{since } \mathbb{E}\left[(\widetilde{y} - SAx_{LS})(\widetilde{y} - SAx_{LS})^T\right] = \frac{1}{m}\left\|y^\perp\right\|_2^2 I_m \end{aligned}$$

For unbiased estimators, we have a lower bound which is easily derived using the Cramér-Rao lower bound. According to the following version of the Cramér-Rao lower bound which only holds for unbiased estimators [36, page 147],

$$\mathbb{E}\left[(\widetilde{x} - x_{LS})(\widetilde{x} - x_{LS})^T \mid SA\right] \succeq I(\widetilde{y}; x_{LS})^{-1}$$

where the inequality $A \succeq B$ means that $A - B$ is positive semi-definite. Since the Gaussian distribution and the Fisher information above are conditioned on SA, this lower bound holds under the same conditioning. We can lower bound the mean squared error as follows:

$$\begin{aligned}\mathbb{E}\left[\|A(\widetilde{x} - x_{LS})\|_2^2 \mid SA\right] &= \operatorname{tr}\left[A\mathbb{E}\left[(\widetilde{x} - x_{LS})(\widetilde{x} - x_{LS})^T \mid SA\right] A^T\right] \\ &\geq \frac{\left\|y^{\perp}\right\|_2^2}{m} \operatorname{tr}\left[A(A^T S^T SA)^{-1} A^T\right]\end{aligned}$$

After taking expectation with respect to SA on both sides, we follow the same steps as in the proof of Lemma 2 to get

$$\mathbb{E}\left[\|A(\widetilde{x} - x_{LS})\|_2^2\right] \geq \frac{d}{m - d - 1}\left\|y^{\perp}\right\|_2^2$$

■

7.5 Proof of Theorem 5

Define the Fisher information matrix $I(\widetilde{y}; x_{LS})$ as in the proof for Theorem 4. In this proof, we will have to use slightly different and more sophisticated tools because the Cramér-Rao bound used in the proof for Theorem 4 holds only for unbiased estimators.

Let $\lambda(x_{LS})$ be the density function of some prior distribution for x_{LS}, which has support on $[-B, B]^d$ and let $I(\lambda)$ be the matrix defined by

$$I(\lambda) = \mathbb{E}\left[\nabla \log \lambda(x_{LS}) \nabla \log \lambda(x_{LS})^T\right]$$

Let $\widetilde{x}$ be an estimator of x_{LS} derived from $\widetilde{y}$. Let $\Lambda = (A^T A)^{-1}$. Our problem is to estimate the parameter $x_{LS} \in \mathcal{C}$ such that $[-B, B]^d \subseteq \mathcal{C}$. We use the following special case of Theorem 1 from [14], which holds under certain regularity conditions which are explained therein. These regularity conditions are easily satisfied by the Gaussian distribution. [14] states a lower bound on the expected error averaged with respect to the density $\lambda(x_{LS})$.

Since we do not consider any probabilistic model on the data, we can lower bound the worst case error by the average error.

$$\mathbb{E}\left[(\widetilde{x}-x_{LS})^T\Lambda^{-1}(\widetilde{x}-x_{LS})\mid SA\right] \geq \frac{d^2}{\mathbb{E}_{x_{LS}}\mathrm{tr}\left[\Lambda^T I(\widetilde{y};x_{LS})\right]+\mathbb{E}_{x_{LS}}\mathrm{tr}\left[\Lambda^T I(\lambda)\right]}$$

In this case, the two quantities in the denominator do not depend on x_{LS} and hence the expectations with respect to x_{LS} are trivial. We then take the expectation of both sides with respect to SA.

$$\begin{aligned}
\mathbb{E}\left[\|A(\widetilde{x}-x_{LS})\|_2^2\right] &= \mathbb{E}_{SA}\left[\mathbb{E}\left[\|A(\widetilde{x}-x_{LS})\|_2^2\mid SA\right]\right] \\
&\geq \mathbb{E}_{SA}\left[\frac{d^2}{\mathrm{tr}\left[\Lambda^T I(\widetilde{y};x_{LS})\right]+\mathrm{tr}\left[\Lambda^T I(\lambda)\right]}\right] \\
&\geq \frac{d^2}{\mathbb{E}_{SA}\,\mathrm{tr}[\Lambda^T I(\widetilde{y};x_{LS})]+\mathrm{tr}\left[\Lambda^T I(\lambda)\right]} \qquad \text{from } \mathbb{E}\frac{1}{X}\geq\frac{1}{\mathbb{E}X}
\end{aligned}$$

We upper bound the two quantities in the denominator as follows:

$$\begin{aligned}
\mathbb{E}_{SA}\,\mathrm{tr}[\Lambda^T I(\widetilde{y};x_{LS})] &= \frac{m}{\|y^\perp\|_2^2}\mathbb{E}_{SA}\,\mathrm{tr}[(A^TA)^{-1}A^TS^TSA] \\
&= \frac{m}{\|y^\perp\|_2^2}\mathrm{tr}[(A^TA)^{-1}\mathbb{E}_{SA}[A^TS^TSA]] \\
&= \frac{m}{\|y^\perp\|_2^2}\mathrm{tr}[(A^TA)^{-1}A^TA] \qquad \text{since } \mathbb{E}[S^TS]=I \\
&= \frac{md}{\|y^\perp\|_2^2}
\end{aligned}$$

$$\begin{aligned}
\mathrm{tr}[\Lambda^T I(\lambda)] &= \mathrm{tr}[(A^TA)^{-1}I(\lambda)] \\
&\leq \left\|(A^TA)^{-1}\right\|_2\mathrm{tr}[I(\lambda)] \qquad \text{because } (A^TA)^{-1} \text{ is symmetric and } I(\lambda)\succeq 0 \\
&= \frac{1}{\sigma_{min}(A^TA)}\mathrm{tr}[I(\lambda)] \qquad \left\|(A^TA)^{-1}\right\|_2=\sigma_{max}(A^TA)^{-1}=\left(\sigma_{min}(A^TA)\right)^{-1}
\end{aligned}$$

The minimum value of $\mathrm{tr}[I(\lambda)]$ over all distributions $\lambda(.)$ which satisfy the regularity condi-

tions and have support in $[-B, B]^d$, is $d\pi^2/B^2$. This result can be found in [36, page 172] and has been used in [21] as well. Since the worst case bound holds for any prior distribution, we use this minimum value.

$$\begin{aligned}\mathbb{E}\left[\|A(\widetilde{x} - x_{LS})\|_2^2\right] &\geq \frac{d^2}{\frac{md}{\|y^\perp\|_2^2} + \frac{d\pi^2}{B^2\sigma_{min}(A^TA)}} \\ &\geq \frac{d}{m}\|y^\perp\|_2^2\left(1 - \frac{\pi^2\|y^\perp\|_2^2}{mB^2\sigma_{min}(A^TA)}\right) \qquad \text{from } (1+x)^{-1} \geq 1 - x\end{aligned}$$

This bound holds for any $B > 0$ such that the hypercube $[-B, B]^d$ lies within the set of feasible x_{LS} values. For unconstrained optimization, we can take $B \to \infty$ and we get

$$\mathbb{E}\left[\|A(\widetilde{x} - x_{LS})\|_2^2\right] \geq \frac{d}{m}\|y^\perp\|_2^2$$

■

7.6 Proof of Theorem 6

From Lemma 1, we have

$$\widehat{x} \sim \mathcal{N}\left(x_{LS}, \frac{1}{m}\|y^\perp\|_2^2(A^TS^TSA)^{-1}\right) \mid SA$$

We use the estimator $\widehat{\theta}(X)$ from Lemma 3 with $X = \widehat{x}$, $\theta = x_{LS}$, and $\Sigma = \frac{1}{m}\|y^\perp\|_2^2(A^TS^TSA)^{-1}$ to get the estimator

$$\begin{aligned}\widehat{x}_S &= \left(1 - \frac{(d-2)}{X^T\Sigma^{-1}X}\right)X \\ &= \left(1 - \frac{(d-2)\|y^\perp\|_2^2}{m\|SA\widehat{x}\|_2^2}\right)\widehat{x}\end{aligned}$$

We look at the approximation of $\left\|y^{\perp}\right\|_2^2$. For proving part (a),

$$\begin{aligned}
\mathbb{E}\left[\|A\widehat{x}-y\|_2^2\right] &= \|Ax_{LS}-y\|_2^2 + \mathbb{E}\left[\|A(\widehat{x}-x_{LS})\|_2^2\right] && \text{since } A^T y^{\perp}=0 \text{ from Lemma 1} \\
&= \left\|y^{\perp}\right\|_2^2 + \frac{d}{m-d-1}\left\|y^{\perp}\right\|_2^2 && \text{from Lemma 2} \\
&= \frac{m-1}{m-d-1}\left\|y^{\perp}\right\|_2^2
\end{aligned}$$

Alternatively, we can also approximate $\left\|y^{\perp}\right\|_2^2$ as $\frac{m}{m-d}\|SA\widehat{x}-Sy\|_2^2$ (see Section 5). This is also an unbiased estimator, as seen below. Firstly, $(SA)^T(Sy-SA\widehat{x})=0$ from the optimality condition for the sketched problem. Then,

$$\begin{aligned}
\mathbb{E}\left[\|SA\widehat{x}-Sy\|_2^2\right] &= \mathbb{E}\left[\|SAx_{LS}-Sy\|_2^2\right] - \mathbb{E}\left[\|SA(\widehat{x}-x_{LS})\|_2^2\right] \\
&= \|Ax_{LS}-y\|_2^2 - \mathbb{E}_{SA}\left[\mathbb{E}\left[\|SA(\widehat{x}-x_{LS})\|_2^2 \mid SA\right]\right] && \text{since } \mathbb{E}[S^TS]=I \\
&= \left\|y^{\perp}\right\|_2^2 - \frac{d}{m}\left\|y^{\perp}\right\|_2^2 && \text{since } \widehat{x}\sim\mathcal{N}\left(x_{LS}, \frac{1}{m}\left\|y^{\perp}\right\|_2^2(A^TS^TSA)^{-1}\right) \mid SA \\
&= \frac{m-d}{m}\left\|y^{\perp}\right\|_2^2
\end{aligned}$$

For the purpose of deriving the upper bound in part (b), we will assume that we make two independent draws of the random matrix S : S_1 and S_2. We will first solve the sketched problem using S_1 to obtain a solution $\widehat{x}_1$, and use that to compute $\|A\widehat{x}_1-y\|_2^2$, for our approximation of $\left\|y^{\perp}\right\|_2^2$. Then we solve the sketched problem using S_2 to obtain a solution $\widehat{x}_2$, and then compute the final estimator as

$$\widehat{x}_{shr} = \left(1-\frac{(d-2)(m-d-1)\|A\widehat{x}_1-y\|_2^2}{m(m-1)\|S_2A\widehat{x}_2\|_2^2}\right)\widehat{x}_2$$

Note that this is only a trick used to simplify the derivation of the error bound. In practice, it is enough to use a single instance of S to compute the estimator.

Let us start by analyzing the approximation term. Since $y^\perp = y - Ax_{LS}$ and $A^T y^\perp = 0$,

$$\begin{aligned}
\|A\widehat{x}_1 - y\|_2^2 &= \left\|y^\perp\right\|_2^2 + \|A(\widehat{x}_1 - x_{LS})\|_2^2 \\
\|A\widehat{x}_1 - y\|_2^4 &= \left\|y^\perp\right\|_2^4 + 2\left\|y^\perp\right\|_2^2\|A(\widehat{x}_1 - x_{LS})\|_2^2 + \|A(\widehat{x}_1 - x_{LS})\|_2^4 \\
\mathbb{E}\left[\|A\widehat{x}_1 - y\|_2^4\right] &= \left\|y^\perp\right\|_2^4 + \frac{2d}{m-d-1}\left\|y^\perp\right\|_2^4 + \mathbb{E}\left[\|A(\widehat{x}_1 - x_{LS})\|_2^4\right] \qquad \text{from Lemma 2}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{m-d-1}{m-1}\|A\widehat{x}_1 - y\|_2^2\right)^2\right] &= \frac{(m-d-1)(m+d-1)}{(m-1)^2}\left\|y^\perp\right\|_2^4 \\
&\quad + \frac{(m-d-1)^2}{(m-1)^2}\mathbb{E}\left[\|A(\widehat{x}_1 - x_{LS})\|_2^4\right] \\
&= \left\|y^\perp\right\|_2^4 - \frac{d^2}{(m-1)^2}\left\|y^\perp\right\|_2^4 + \frac{(m-d-1)^2}{(m-1)^2}\mathbb{E}\left[\|A(\widehat{x}_1 - x_{LS})\|_2^4\right]
\end{aligned} \tag{26}$$

To compute the last term on the right hand side, let $z = A(\widehat{x}_1 - x_{LS})$, then

$$z \sim \mathcal{N}\left(0, K = \frac{1}{m}\left\|y^\perp\right\|_2^2 A(A^T S_1^T S_1 A)^{-1} A^T\right) \mid S_1 A$$

We require the 4^{th} order moments of this Gaussian, which is as follows:

$$\begin{aligned}
\mathbb{E}\left[\|A(\widehat{x}_1 - x_{LS})\|_2^4 \mid S_1 A\right] &= \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n z_i^2 z_j^2\right] \\
&= \sum_{i=1}^n \sum_{j=1}^n \left(2K_{ij}^2 + K_{ii}K_{jj}\right)
\end{aligned}$$

Since each row of $S_1 A$ has an independent multivariate Gaussian distribution $\mathcal{N}\left(0, \frac{1}{m}A^T A\right)$, $A^T S_1^T S_1 A$ follows a Wishart distribution $\mathcal{W}_d\left(m, \frac{1}{m}A^T A\right)$, and so its inverse follows an inverted Wishart distribution $\mathcal{W}_d^{-1}\left(m, m(A^T A)^{-1}\right)$, and therefore

$$K \sim \mathcal{W}_d^{-1}\left(m, \left\|y^\perp\right\|_2^2 A(A^T A)^{-1} A^T\right)$$

Let $P = A(A^TA)^{-1}A^T$ be the "hat matrix". Using the covariance of elements of an inverted Wishart matrix from [22],

$$\begin{aligned}
\mathbb{E}\left[\|A(\widehat{x}_1 - x_{LS})\|_2^4\right] &= \mathbb{E}_{S_1A}\left[\sum_{i=1}^{n}\sum_{j=1}^{n}\left(2K_{ij}^2 + K_{ii}K_{jj}\right)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\left(2\mathbb{E}\left[K_{ij}\right]^2 + 2Var\left(K_{ij}\right) + \mathbb{E}\left[K_{ii}\right]\mathbb{E}\left[K_{jj}\right] + Cov\left(K_{ii}, K_{jj}\right)\right) \\
&= \left\|y^{\perp}\right\|_2^4\left(\frac{2\|P\|_F^2}{(m-d-1)^2} + \frac{2(m-d+1)\|P\|_F^2 + 2(m-d-1)\operatorname{tr}(P)^2}{(m-d)(m-d-1)^2(m-d-3)}\right. \\
&\quad \left. + \frac{\operatorname{tr}(P)^2}{(m-d-1)^2} + \frac{2\operatorname{tr}(P)^2 + 2(m-d-1)\|P\|_F^2}{(m-d)(m-d-1)^2(m-d-3)}\right) \\
&= \frac{\left\|y^{\perp}\right\|_2^4\left(2\|P\|_F^2 + \operatorname{tr}(P)^2\right)}{(m-d-1)(m-d-3)}
\end{aligned}$$

For the hat matrix, $\operatorname{tr}(P) = \operatorname{tr}(A(A^TA)^{-1}A^T) = \operatorname{tr}((A^TA)^{-1}A^TA) = d$, and $\|P\|_F^2 = \operatorname{tr}(P^TP) = \operatorname{tr}(P) = d$.

$$\mathbb{E}\left[\|A(\widehat{x}_1 - x_{LS})\|_2^4\right] = \frac{\left\|y^{\perp}\right\|_2^4 d(d+2)}{(m-d-1)(m-d-3)}$$

Substituting this in (26),

$$\mathbb{E}\left[\left(\frac{m-d-1}{m-1}\|A\widehat{x}_1 - y\|_2^2\right)^2\right] = \left\|y^{\perp}\right\|_2^4 - \frac{d^2}{(m-1)^2}\left\|y^{\perp}\right\|_2^4 + \frac{d(d+2)(m-d-1)}{(m-1)^2(m-d-3)}\left\|y^{\perp}\right\|_2^4 \tag{27}$$

Now, we look at the following conditional expectation.

$$\mathbb{E}\left[(\widehat{x}_{shr} - x_{LS})^T\Sigma^{-1}(\widehat{x}_{shr} - x_{LS}) \mid S_1A, S_2A\right]$$

Since $\widehat{x}_1$ and $\widehat{x}_2$ are independent (even when conditioned on S_1A, S_2A), we can first take

expectation with respect to $\widehat{x}_1$, and using (27), this becomes

$$\mathbb{E}\left[(\widehat{x}_S - x_{LS})^T \Sigma^{-1} (\widehat{x}_S - x_{LS}) \mid S_2 A\right] - \frac{d^2(d-2)^2}{(m-1)^2} \mathbb{E}\left[\frac{1}{\widehat{x}_2^T \Sigma^{-1} \widehat{x}_2}\right] + \frac{d(d+2)(d-2)^2(m-d-1)}{(m-1)^2(m-d-3)} \mathbb{E}\left[\frac{1}{\widehat{x}_2^T \Sigma^{-1} \widehat{x}_2}\right]$$

where $\Sigma = \frac{1}{m}\left\|y^{\perp}\right\|_2^2 (A^T S_2^T S_2 A)^{-1}$ and $\widehat{x}_S$ (the un-approximated estimator) is calculated using S_2 and $\widehat{x}_2$. Then we use the normalized error from Lemma 3 to get

$$\mathbb{E}\left[(\widehat{x}_S - x_{LS})^T \Sigma^{-1} (\widehat{x}_S - x_{LS}) \mid S_2 A\right] = d - (d-2)^2 \mathbb{E}\left[\frac{1}{\widehat{x}_2^T \Sigma^{-1} \widehat{x}_2}\right]$$

Substituting $\Sigma = \frac{1}{m}\left\|y^{\perp}\right\|_2^2 (A^T S_2^T S_2 A)^{-1}$, and using $\mathbb{E}\frac{1}{X} \geq \frac{1}{\mathbb{E}X}$,

$$\begin{aligned}
&\frac{m}{\|y^{\perp}\|_2^2} \mathbb{E}\left[\|S_2 A(\widehat{x}_{shr} - x_{LS})\|_2^2 \mid S_2 A\right] \\
&= d - \mathbb{E}\left[\frac{1}{\widehat{x}_2^T \Sigma^{-1} \widehat{x}_2}\right] \left((d-2)^2 + \frac{d^2(d-2)^2}{(m-1)^2} - \frac{d(d+2)(d-2)^2(m-d-1)}{(m-1)^2(m-d-3)}\right) \\
&\leq d - \frac{(d-2)^2}{d + m\|S_2 A x_{LS}\|_2^2 / \|y^{\perp}\|_2^2} \left(1 + \frac{d^2}{(m-1)^2} - \frac{d(d+2)(m-d-1)}{(m-1)^2(m-d-3)}\right)
\end{aligned}$$

At this stage we replace the notation S_2 back by S which has the identical distribution.

Taking expectation with respect to SA on both sides,

$$
\begin{aligned}
&\mathbb{E}\left[\|SA(\widehat{x}_{shr}-x_{LS})\|_2^2\right] \\
&\leq \frac{\left\|y^{\perp}\right\|_2^2}{m}\mathbb{E}_{SA}\left[d-\frac{(d-2)^2\left(1+\frac{d^2}{(m-1)^2}-\frac{d(d+2)(m-d-1)}{(m-1)^2(m-d-3)}\right)}{d+m\|SAx_{LS}\|_2^2/\|y^{\perp}\|_2^2}\right] \\
&\leq \frac{\left\|y^{\perp}\right\|_2^2}{m}\left(d-\frac{(d-2)^2\left(1+\frac{d^2}{(m-1)^2}-\frac{d(d+2)(m-d-1)}{(m-1)^2(m-d-3)}\right)}{d+m\mathbb{E}_{SA}[\|SAx_{LS}\|_2^2]/\|y^{\perp}\|_2^2}\right) && \text{from } \mathbb{E}\frac{1}{X}\geq\frac{1}{\mathbb{E}X} \\
&= \frac{d}{m}\left\|y^{\perp}\right\|_2^2\left(1-\frac{(d-2)^2/d\left(1+\frac{d^2}{(m-1)^2}-\frac{d(d+2)(m-d-1)}{(m-1)^2(m-d-3)}\right)}{d+m\|Ax_{LS}\|_2^2/\|y^{\perp}\|_2^2}\right) && \text{since } \mathbb{E}[S^TS]=I \\
&= \frac{d}{m}\left\|y^{\perp}\right\|_2^2\left(1-\frac{1-\epsilon'(d,m)}{1+(m/d)\|Ax_{LS}\|_2^2/\|y^{\perp}\|_2^2}\right)
\end{aligned}
$$

where $\epsilon'(d,m)=\frac{4(d-1)}{d^2}+\frac{2(d-2)^2}{d(m-1)(m-d-3)}$.

Finally for proving part (c), we use the fact that S is an Oblivious Subspace Embedding [32] so that when $m>\mathcal{O}\left((d+\log(1/\delta))/\epsilon^2\right)$, with probability greater than $1-\delta$,

$$(1-\epsilon)\|Ax\|_2^2\leq\|SAx\|_2^2\leq(1+\epsilon)\|Ax\|_2^2$$

Then for sufficiently small ϵ such that $(1-\epsilon)^{-1}\approx(1+\epsilon)$,

$$
\begin{aligned}
\mathbb{E}\left[\|A(\widehat{x}_{shr}-x_{LS})\|_2^2\right] &\leq (1+\epsilon)\mathbb{E}\left[\|SA(\widehat{x}_{shr}-x_{LS})\|_2^2\right] \\
&\leq (1+\epsilon)\frac{d}{m}\left\|y^{\perp}\right\|_2^2\left(1-\frac{1-\epsilon'(d,m)}{1+(m/d)\|Ax_{LS}\|_2^2/\|y^{\perp}\|_2^2}\right)
\end{aligned}
$$