A New Theory for Sketching in Linear Regression

Edgar Dobriban* and Sifan Liu†

December 6, 2018

Abstract

Large datasets create opportunities as well as analytic challenges. A recent development is to use *random projection* or *sketching* methods for dimension reduction in statistics and machine learning. In this work, we study the statistical performance of sketching algorithms for linear regression. Suppose we randomly project the data matrix and the outcome using a random sketching matrix reducing the sample size, and do linear regression on the resulting data. How much do we lose compared to the original linear regression? The existing theory does not give a precise enough answer, and this has been a bottleneck for using random projections in practice.

In this paper, we introduce a new mathematical approach to the problem, relying on very recent results from *asymptotic random matrix theory* and *free probability theory*. This is a perfect fit, as the sketching matrices are random in practice. We allow the dimension and sample sizes to have an arbitrary ratio. We study the most popular sketching methods in a unified framework, including random projection methods (Gaussian and iid projections, uniform orthogonal projections, subsampled randomized Hadamard transforms), as well as sampling methods (including uniform, leverage-based, and greedy sampling). We find precise and simple expressions for the accuracy loss of these methods. These go beyond classical Johnson-Lindenstrauss type results, because they are exact, instead of being bounds up to constants. Our theoretical formulas are surprisingly accurate in extensive simulations and on two empirical datasets.

1 Introduction

As big data continuously creates value in all areas of the world, the sheer volume of data also brings computational and statistical challenges. This motivates the development of new methodologies for large-scale data analysis. A recent development is to exploit *randomization* as a computational resource. More specifically, *sketching* or *random projection* methods are a fundamental tool in numerical linear algebra, of wide applicability to statistics, machine learning, data mining, optimization (e.g., Mahoney, 2011; Woodruff, 2014; Drineas and Mahoney, 2016).

In this work, we study the statistical performance of sketching algorithms in linear regression. Linear regression is a fundamental problem in statistics, the basis of many more sophisticated methods. Here we are interested in the accuracy loss introduced by different sketching algorithms.

*Wharton Statistics Department, University of Pennsylvania. E-mail: dobriban@wharton.upenn.edu.

†Department of Mathematical Sciences, Tsinghua University. E-mail: liusf15@mails.tsinghua.edu.cn. The bulk of this work was performed while SL was visiting the Wharton Statistics Department during Summer 2018.

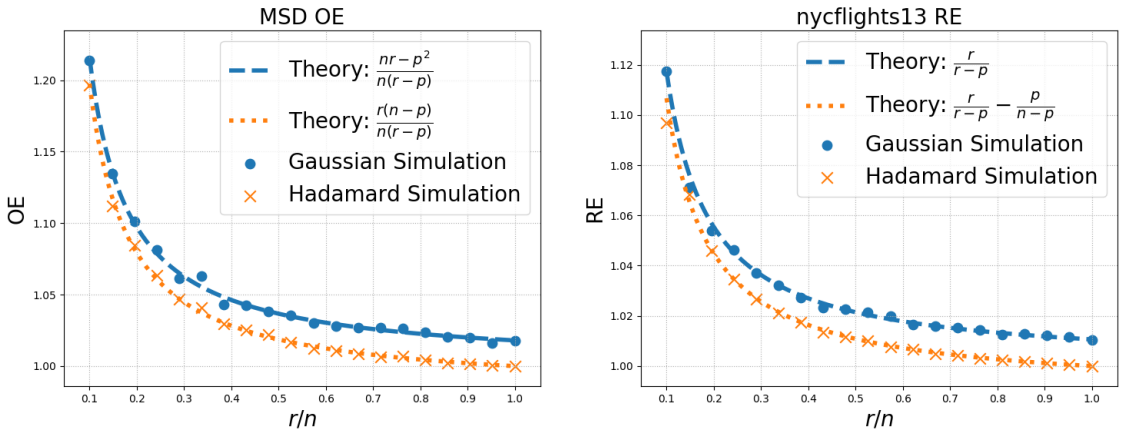

Figure 1: Experiments on two datasets. On the left, we use a subset of the Million Song Dataset (MSD) with $n = 5000$ samples and $p = 90$ features. The data is arranged into the $n \times p$ matrix X and the $n \times 1$ outcome vector Y. Instead of doing regression of Y on X, in sketching we randomly project X, Y into $(\tilde{X}, \tilde{Y}) = (SX, SY)$ using a random $r \times n$ sketching matrix S. The sketched data has a smaller effective sample size $r < n$. Then we do regression of $\tilde{Y}$ on $\tilde{X}$. How much does the statistical error increase? We evaluate the increase in test error (out-of-sample prediction error, OE) for Gaussian random projections and subsampled randomized Hadamard transform (SRHT), for a range of r. On the right plot, we use a subset of nycflights13 dataset, where $n = 2000$, and $p = 21$. Here we evaluate the increase in the residuals (RE). We plot both our theoretical formulas (lines) and the empirical results (dots). Our theory agrees well with the empirical results.

This fundamental problem has been studied in many works, for instance Drineas et al. (2006, 2011); Dhillon et al. (2013); Raskutti and Mahoney (2016); Ma et al. (2015); Thanei et al. (2017). However, the existing theory does not give a clear answer for how much accuracy is lost, and this has been a bottleneck for using random projections in practice. This has been recognized as a challenge in the community, as argued for instance by Ravi Kannan in his recent keynote talk on the "Foundations of Data Science" (Kannan, 2018),

In this paper, we introduce a new mathematical approach to the problem, relying on very recent results from *asymptotic random matrix theory* and *free probability theory*. This enables us to get highly accurate results for the performance of sketching. Using random matrix theory for sketching is a perfect fit, as the sketching matrices are already randomly generated.

In addition, we study the problem in a regime where we allow the relevant dimensions and sample sizes to have arbitrary aspect ratios. We are able to study many of the most popular and important sketching methods in a unified framework, including random projection methods (Gaussian and iid projections, uniform orthogonal—Haar—projections, subsampled randomized Hadamard transforms) as well as random sampling methods (including uniform, randomized leverage-based, and greedy leverage sampling). We find simple and elegant formulas for the accuracy loss of these methods compared to unprojected linear regression. We think that this is the "correct" way to study sketching, because our theoretical formulas are also accurate in extensive simulations and on two empirical datasets (see Figure 1 for a quick example, and Table 1 for details).

Table 1: How much do we lose by using sketching in linear regression? Suppose we have a linear model $Y = X\beta + \varepsilon$ of size $n \times p$, with n training datapoints X and p dimensions ($n > p$). The sketched linear model $SY = SX\beta + S\varepsilon$ has size $r \times p$, where $n \geqslant r > p$. Sketching leads to a performance loss. We show the degradation of three loss functions: VE (variance efficiency—increase in parameter estimation error), PE (prediction efficiency—increase in regression function estimation error), and OE (out-of-sample prediction efficiency–increase in test error), as defined in Section 1.2. For instance, when S is a matrix with iid entries, the variance efficiency is $1 + (n-p)/(r-p)$, so estimation error increases by that factor due to sketching. For the results on leverage-based sampling, see Section 2.5.2.

<table>
<tr><th>S</th><th>X</th><th>VE: $\frac{\mathbb{E}\|\hat{\beta}_s-\beta\|^2}{\mathbb{E}\|\hat{\beta}-\beta\|^2}$</th><th>$PE$: $\frac{\mathbb{E}\|X\hat{\beta}_s-X\beta\|^2}{\mathbb{E}\|X\hat{\beta}-X\beta\|^2}$</th><th>$OE$: $\frac{\mathbb{E}(x_t^\top\hat{\beta}_s-y_t)^2}{\mathbb{E}(x_t^\top\hat{\beta}-y_t)^2}$</th></tr>
<tr><td>iid entries</td><td rowspan="2">Fixed</td><td colspan="2" rowspan="2">$1 + \frac{n-p}{r-p}$</td><td>$\frac{nr-p^2}{n(r-p)}$</td></tr>
<tr><td>Haar/Hadamard</td><td rowspan="2">$\frac{r(n-p)}{n(r-p)}$</td></tr>
<tr><td>Uniform sampling</td><td>Ortho-invariant</td><td colspan="2">$\frac{n-p}{r-p}$</td></tr>
<tr><td>Leverage-based sampling</td><td>Elliptical: $WZ\Sigma^{1/2}$</td><td>$\frac{\eta_{sw^2}^{-1}(1-p/n)}{\eta_{w^2}^{-1}(1-p/n)}$</td><td>$1+\mathbb{E}[w^2(1-s)]\frac{\eta_{sw^2}^{-1}(1-p/n)}{p/n}$</td><td>$\frac{1+\mathbb{E}[w^2]\eta_{sw^2}^{-1}(1-\gamma)}{1+\mathbb{E}[w^2]\eta_{w^2}^{-1}(1-\gamma)}$</td></tr>
</table>

1.1 Sketched linear regression

Suppose we observe n datapoints (x_i, y_i), $i = 1, \ldots, n$, where x_i are the p-dimensional features (or predictors, covariates) of the i-th datapoint, and y_i are the outcomes (or responses). We assume the usual linear model $y_i = x_i^\top \beta + \varepsilon_i$, where β is an unknown p-dimensional parameter and ε_i is the noise of the i-th sample. In matrix form, we have

$$Y = X\beta + \varepsilon,$$

where X is the $n \times p$ data matrix, the i-th row $x_i^\top$, and Y is the $n \times 1$ outcome vector, the i-th entry y_i. We further assume that ε_i are uncorrelated with zero mean and have equal variance σ^2. Then the usual ordinary least squares (OLS) estimator is

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

Sketching reduces the size of the data before doing linear regression. We multiply (X, Y) by the $r \times n$ matrix S to obtain the *sketched data* $(\tilde{X}, \tilde{Y}) = (SX, SY)$. Instead of doing regression of Y on X, we do regression of $\tilde{Y}$ on $\tilde{X}$. The solution is

$$\hat{\beta}_s = (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{Y},$$

if $rank(SX) = p$. In the remainder, we assume that both X and SX have full column rank, so that the OLS estimators of β are well defined.

1.2 Error criteria

To compare the statistical efficiency of the two estimators $\hat{\beta}$ and $\hat{\beta}_s$, we evaluate the relative value of their mean squared error. If we use the full OLS estimator, we incur a mean squared error of

$\mathbb{E}\|\hat{\beta} - \beta\|^2$. If we use the sketched OLS estimator, we incur a mean squared error of $\mathbb{E}\|\hat{\beta}_s - \beta\|^2$ instead. To see how much efficiency we lose, it is natural and customary in statistics to consider the relative efficiency, which is their ratio. We call this the *variance efficiency* (VE), because the MSE for estimation can be viewed as the sum of variances of the OLS estimator. Hence, we define

$$VE(\hat{\beta}_s, \hat{\beta}) = \frac{\mathbb{E}\left[\|\hat{\beta}_s - \beta\|^2\right]}{\mathbb{E}\left[\|\hat{\beta} - \beta\|^2\right]}.$$

This quantity is greater than or equal to unity, so $VE \geqslant 1$, and *smaller is better*. An accurate sketching method would achieve an efficiency close to unity.

The expectations are taken with respect to all sources of randomness in the problem. This always includes the sampling noise and the sketching matrix. In general the data matrix X is fixed and arbitrary, but in some cases we also assume a bit of randomness on X to facilitate the proofs (see the theorem statements for details).

It is also of interest to estimate the regression function $\mathbb{E}(Y|X) = X\beta$, using a plug-in estimator $X\hat{\beta}$. We consider a similar relative efficiency, called the *prediction efficiency* (PE), as we are attempting to predict the regression function. Specifically,

$$PE(\hat{\beta}_s, \hat{\beta}) = \frac{\mathbb{E}\left[\|X\hat{\beta}_s - X\beta\|^2\right]}{\mathbb{E}\left[\|X\hat{\beta} - X\beta\|^2\right]}.$$

The variance efficiency and prediction efficiency can both be regarded as a usual ratio of mean squared errors of two estimators. For VE, the parameter is the vector of regression coefficients β, while for PE, it is the regression function $\mathbb{E}(Y|X) = X\beta$.

Finally, we also study the important problem of out-of-sample prediction or test error. We suppose we have a test datapoint (x_t, y_t), generated from the same model $y_t = x_t^\top \beta + \varepsilon_t$, where x_t, ε_t are independent of X, ε, and only x_t is observable. We want to use $x_t^\top \hat{\beta}$ to predict y_t. The *out-of-sample efficiency* (OE) tells us how much larger the test error becomes after sketching. Specifically,

$$OE(\hat{\beta}_s, \hat{\beta}) = \frac{\mathbb{E}\left[(x_t^\top \hat{\beta}_s - y_t)^2\right]}{\mathbb{E}\left[(x_t^\top \hat{\beta} - y_t)^2\right]}.$$

While these criteria are very natural, we are particularly inspired by Raskutti and Mahoney (2016) to study them. In particular, prediction efficiency was also studied by Raskutti and Mahoney (2016), however in a different framework (see the related work section for details). That work also considers the residual efficiency $\mathbb{E}\|Y - X\hat{\beta}_s\|^2 / \mathbb{E}\|Y - X\hat{\beta}\|^2$, which measures how much the residuals increase after sketching. It is easy to check that

$$RE(\hat{\beta}_s, \hat{\beta}) - 1 = \frac{p}{n-p}\left[PE(\hat{\beta}_s, \hat{\beta}) - 1\right], \tag{1}$$

so RE is a linear transform of PE. The formulas for RE can be derived from those for PE.

1.3 Summary of our results

We consider a "large data" asymptotic limit, where both the dimension p and the sample size n tend to infinity, and their aspect ratio converges to a constant. The size r of the sketched data is also proportional to the sample size. Specifically n, p, and r tend to infinity such that the original aspect ratio converges, $p/n \to \gamma \in (0, 1)$, while the data reduction factor also converges, $r/n \to \xi \in (\gamma, 1)$. Under these asymptotics, we find the limits of the relative efficiencies under various conditions on X and S. This allows the relevant dimensions and sample sizes to have arbitrary aspect ratios. This asymptotic setting is different from the usual one under which sketching is studied, where $n \gg r$; however our results are accurate even in that regime.

Our goal is to study popular random sketching methods, such as Gaussian projection, randomized Hadamard projection, etc. For some of these, we can find the efficiencies without any assumptions on X, while for others, we need to make stronger assumptions. Our results are summarized in Table 1. For instance, when S is a matrix with iid entries, the variance efficiency is $1 + (n - p)/(r - p)$, so estimation error increases by that factor due to sketching. The results are stated formally in theorems in the remainder of the paper.

Here are some observations:

1. Our formulas are *simple*. This is valuable, because practitioners can conveniently use them as guidance for comparing different sketching methods.

2. The formulas are *accurate*. This holds both in simulations and in two empirical data analysis examples. We show some examples in Figure 1. The experimental results agree well with our theory. In particular, the formulas are accurate in relatively small samples. Moreover, they go beyond Johnson-Lindentsrauss type results because they are accurate not just up to the rate, but also down to the precise constants.

3. From a theoretical point of view, it is surprising that they *do not depend on the data in a complicated way*. In particular, they do not depend on any un-estimable parameters of the data. The only complex formulas arise for leverage score sampling (see the relevant section for discussion).

4. The formulas enable a detailed *comparison* of the different sketching methods. For instance, in estimation error (VE), we have $VE_{\text{iid}} = VE_{\text{Haar}} + 1 = VE_{\text{Hadamard}} + 1$. This shows that estimation error for uniform orthogonal (Haar) random projections and the subsampled randomized Hadamard transform (SRHT) (Ailon and Chazelle, 2006) increases less than for iid random projections.

   A random projection with iid entries will *degrade the performance of OLS even if we do not reduce the sample size*. This is because iid projections distort the geometry of Euclidean space due to their non-orthogonality.

5. Therefore, the SRHT appears to be the most favorable one-step sketching method. It takes time $O(np \log(n))$ to apply and is as accurate as uniform Haar random projections.

6. When the data is near rotationally invariant, uniform sampling can work as well as Haar projections and Hadamard projections. But when the data is nonuniform, uniform sampling may fail to compete with Haar/Hadamard projections. In this case, leverage-based sampling has much better performance. This is consistent with prior findings by Drineas, Mahoney, Woodruff and others (e.g., Mahoney, 2011; Woodruff, 2014; Drineas and Mahoney, 2016).

7. The asymptotic framework where the dimension and sample sizes are proportional leads to new insights and observations that are not easily observable using other approaches.

The structure of the paper is as follows: We review some related work in Section 1.4. We state our theoretical results in Section 2. We study various types sketching matrices: Gaussian, iid, uniform (Haar), subsampled randomized Hadamard transform (SRHT), uniform and leverage sampling. We provide simulation experiments to support our theory in Section 3. We verify our theoretical results on two empirical datasets in Section 4. We provide the proofs of our results, some mathematical background, and additional simulations, in Section 5. The code for the paper can be found at `https://github.com/liusf15/Sketching-lr`.

1.4 Related work

In this section we review some related work. Due to space limitations, we can only consider the most closely related work. For general overviews of sketching and random projection methods from a numerical linear algebra perspective, see Halko et al. (2011); Mahoney (2011); Woodruff (2014); Drineas and Mahoney (2017). For a theoretical computer science perspective, see Vempala (2005).

A cornerstone result in this area is the Johnson-Lindenstrauss lemma. This states that norms, and thus also relative distances between points, are approximately preserved after sketching i.e., $(1-\delta)\|x_i\|^2 \leq \|Sx_i\|^2 \leq (1+\delta)\|x_i\|^2$ for $x_1, \ldots, x_n \in \mathbb{R}^p$. This is also known as the subspace embedding property. Each projection studied in this paper has the embedding property, and this can be used to derive bounds for regression. However, our results are much more refined, because they quantify the precise value of the error, while the bounds above are up to the constant δ and hold with some probability.

More recent results study the performance of random projections, sampling or sketching in linear regression. Drineas et al. (2006) show that leverage score sampling leads to better results than uniform sampling in regression. Drineas et al. (2012), show furthermore that leverage scores can be approximated fast (using the Hadamard transform). Drineas et al. (2011) propose the fast Hadamard transform for sketching in regression. They prove strong relative error bounds on the realized in-sample prediction error for arbitrary input data. Our results concern the a different setting that assumes a statistical model.

Ma et al. (2015) study leverage score sampling in a statistical linear model. They derive a Taylor series expansion of the sampled least squares estimator around the full estimator. They use this to study the bias and MSE of the sketched estimator. Our analysis is different, because we directly study the error of the sketched estimator around the true parameter.

One of the most related works, and a major inspiration for us, is Raskutti and Mahoney (2016). They study sketching algorithms from both statistical and algorithmic perspectives. However, they focus on a different setting, where $n \gg r$, and prove bounds on RE and PE. For instance, they discover that RE can be bounded even when r is not too large, proving bounds such as $RE \leqslant 1 + 44p/r$. In contrast, we show more precise results such as $RE \approx r/(r-p)$, (without the constant 44), under slightly stronger conditions. We show that these conditions are reasonable, because our results are accurate both in simulations and in empirical data analysis examples.

Another related work is Ahfock et al. (2017), who study the statistical properties of sketching algorithms. They develop confidence intervals for sketching estimators and bounds on the relative efficiency of different sketching algorithms. However, their perspective is different and complementary to ours, because they do not assume a linear model for the data.

Randomized dimension reduction can also be studied using stochastic geometry. Oymak and Tropp (2017) develop such results, and as an application show that for a sketching matrix with independent entries, the RE is bounded as $\|Y - X\hat{\beta}_s\|^2/\|Y - X\hat{\beta}\|^2 \leq (1+t)r/(r-p)$, for $t > 0$ with high probability depending on t. This is consistent with our results, which imply that the RE has a limit of $r/(r-p)$.

Pilanci and Wainwright (2015) study random projection in ℓ^2 regression with convex constraints. Our results are more accurate, albeit only for the unconstrained case. Lopes et al. (2018) develop methods for error estimation for randomized least-squares algorithms via the bootstrap. Our goal is different, namely to develop precise theoretical results.

Apart from reducing the sample size, one can also apply sketching to reduce the number of features. For instance, Maillard and Munos (2009); Kabán (2014); Thanei et al. (2017) also studied the column-wise compression, i.e. $\min_{\beta \in \mathbb{R}^k} \|Y - XS\beta\|$, where S is a $p \times k$ matrix. Moreover, Zhou et al. (2008); Fard et al. (2012) studied compressed regression with a sparsity assumption.

Another method, partial sketching (Dhillon et al., 2013; Ahfock et al., 2017) is to use a scalar multiple of $\hat{\beta}_{part} = (\tilde{X}^\top \tilde{X})^{-1} X^\top Y$. Ahfock et al. (2017) showed that when $Y^\top HY/(Y^\top Y)$ is close to unity (where H is the hat matrix), complete sketching is much more efficient. When it is close to 0, (unbiased) partial sketching is much more efficient. That work also studied the linear combination of full and partial sketching estimators which minimizes the mean squared error.

Moreover, it is important to see that random projection and sketching ideas are increasingly being used in various areas, for instance

1. In statistics, machine learning and data mining, they have recently been applied to nonparametric regression (Yang et al., 2017), ridge regression (Lu et al., 2013; Wang et al., 2018), two sample testing (Lopes et al., 2011; Srivastava et al., 2016), text and image mining (Bingham and Mannila, 2001), classification (Cannings and Samworth, 2017) and clustering (Fern and Brodley, 2003).
2. In convex optimization they are recently proving to be a promising approach (Pilanci and Wainwright, 2015, 2016, 2017).
3. In applied areas, they are starting to be used for instance in economics (Ng, 2017), forecasting (Schneider and Gupta, 2016) and genomics (Galinsky et al., 2016).
4. Random projection can be used for privacy preserving data analysis (e.g., Liu et al., 2006; Jassim et al., 2009).

2 Theoretical results

2.1 Gaussian projection

We begin with the simplest-to-analyze model, that of Gaussian random projections, and work towards more sophisticated methods. An advantage of Gaussian projections is that generating and multiplying Gaussian matrices is *embarrassingly parallel*, making it appropriate for certain distributed and cloud-computing architectures.

Theorem 2.1 (Gaussian projection)**.** *Suppose S is an $r \times n$ Gaussian random matrix with iid standard normal entries. Let X be an arbitrary $n \times p$ matrix with full column rank p, and suppose*

that $r - p > 1$. Then the efficiencies have the following form

$$VE(\hat{\beta}_s, \hat{\beta}) = PE(\hat{\beta}_s, \hat{\beta}) = 1 + \frac{n-p}{r-p-1},$$
$$OE(\hat{\beta}_s, \hat{\beta}) = \frac{1 + \left[1 + \frac{n-p}{r-p-1}\right] x_t^\top (X^\top X)^{-1} x_t}{1 + x_t^\top (X^\top X)^{-1} x_t}.$$

Suppose in addition that X is also random, having the form $X = Z\Sigma^{1/2}$, where $Z \in \mathbb{R}^{n\times p}$ has iid entries of zero mean, unit variance and finite fourth moment, and $\Sigma \in \mathbb{R}^{p\times p}$ is a deterministic positive definite matrix. If the test datapoint is drawn independently from the same population as X, i.e. $x_t = \Sigma^{1/2} z_t$, then as n, p, r grow to infinity proportionally, with $p/n \to \gamma \in (0, 1)$ and $r/n \to \xi \in (\gamma, 1)$, we have the simple formula for OE

$$\lim_{n\to\infty} OE(\hat{\beta}_s, \hat{\beta}) = \frac{\xi - \gamma^2}{\xi - \gamma} \approx \frac{nr - p^2}{n(r-p)}.$$

See Section 5.3 for the proof. In finite samples the formula for the OE can be approximated by $(nr - p^2)/[n(r-p)]$. These formulas have all the properties we claimed before: they are simple, accurate, and easy to interpret. We observe the following:

1. The relative efficiencies *decrease* with r/n, the ratio of preserved samples after sketching. This is because a larger number of samples leads to a higher accuracy.

2. When $\xi = \lim r/n = 1$, VE and PE reach a minimum of 2. Thus, taking a random Gaussian projection will *degrade the performance of OLS even if we do not reduce the sample size.* This is because iid projections distort the geometry of Euclidean space due to their non-orthogonality. We will see how to overcome this using orthogonal random projections.

2.2 iid projection

Our results so far only include Gaussian random projections. What happens for more general projections with iid entries? This is an important question, because *sparse projections* such as those with iid $0, 1, -1$ entries can speed up computation (see e.g., Achlioptas, 2001). In this section we show that we can use sparse matrices to do random projections as long as the sparsity level is fixed, and we obtain the same performance as with Gaussian projections. This is an instance of *universality*.

As a preliminary remark, in this case the probability of SX being singular is nonzero. However, this probability vanishes rapidly as $n \to \infty$, and so this will cause no problems. For clarity, when SX is singular, we define $(X^\top S^\top SX)^{-1}$ as the zero matrix.

Theorem 2.2 (Universality for iid projection)**.** *Suppose that S has iid entries of zero mean and finite fourth moment. Suppose also that X is a deterministic matrix, whose singular values are uniformly bounded away from zero and infinity. Then as n goes to infinity, while $p/n \to \gamma \in (0, 1)$, $r/n \to \xi \in (\gamma, 1)$, the efficiencies have the limits*

$$\lim_{n\to\infty} VE(\hat{\beta}_s, \hat{\beta}) = \lim_{n\to\infty} PE(\hat{\beta}_s, \hat{\beta}) = 1 + \frac{1-\gamma}{\xi - \gamma}.$$

Suppose in addition that X is also random, under the same model as in Theorem 2.1. Then the formula for OE given there still holds in this more general case.

See Section 5.4 for the proof. To summarize, asymptotically we get the same behavior for iid matrices as for Gaussian matrices.

2.3 Uniform (Haar) random projection

We saw that a random projection with iid entries will degrade the performance of OLS *even if we do not reduce the sample size*. That is, even if we use an $n \times n$ random projection matrix S with iid entries, our performance will degrade. Hence matrices with iid entries are not ideal for sketching. This is because iid projections distort the geometry of Euclidean space due to their non-orthogonality. Is it possible to overcome this using orthogonal random projections?

To study this, we next take our sketching matrix S to be a Haar random matrix uniformly distributed over the space of all $r \times n$ partial orthogonal matrices. This can also be seen as the first r rows of an $n \times n$ Haar matrix distributed uniformly over orthogonal matrices.

Recall that for an $n \times p$ matrix M with $n \geqslant p$, such that the eigenvalues of $n^{-1}M^\top M$ are λ_j, the empirical spectral distribution (e.s.d.) of M is the cdf of the eigenvalues. Formally, it is the mixture

$$\frac{1}{p}\sum_{j=1}^{p}\delta_{\lambda_j},$$

where δ_λ denotes a point mass distribution at λ.

Theorem 2.3 (Haar projection)**.** *Suppose that S is an $r \times n$ Haar-distributed random matrix. Suppose also that X is a deterministic matrix whose e.s.d. converges weakly to some fixed probability distribution with compact support bounded away from the origin. Then as n tends to infinity, while $p/n \to \gamma \in (0,1)$, $r/n \to \xi \in (\gamma, 1)$, the efficiencies have the limits*

$$\lim_{n\to\infty} VE(\hat{\beta}_s, \hat{\beta}) = \lim_{n\to\infty} PE(\hat{\beta}_s, \hat{\beta}) = \frac{1-\gamma}{\xi-\gamma}.$$

Suppose in addition that the training and test data X and x_t are also random, under the same model as in Theorem 2.1. Then

$$\lim_{n\to\infty} OE(\hat{\beta}_s, \hat{\beta}) = \frac{1-\gamma}{1-\gamma/\xi}.$$

See Section 5.5 for the proof, which shows (see Section 5.5.1) that the limiting eigenvalue spectrum of SX is the *free multiplicative convolution* of the limiting eigenvalue spectra of S and X (combined with a point mass at zero).

We observe that orthogonal projections are *uniformly better* than iid projections. For instance, in estimation error (VE),

$$VE_{\text{iid}} = VE_{\text{Haar}} + 1.$$

This shows that estimation error for Haar random projections increases much less than for iid random projections. However, there is still a tradeoff between statistical accuracy and computational

cost, since the time complexity of generating a Haar matrix using the Gram-Schmidt procedure is $O(np^2)$.

When we do not reduce the size of the original matrix, so that $r = n$ and $\xi = 1$, VE, PE and OE are all equal to unity. This means that we do not lose any information in the linear model. Geometrically, left-multiplying X and Y by an $n \times n$ orthogonal matrix S rotates the columns of X and Y. The geometry between Y and the column space of X does not change. In contrast, Gaussian projections introduce more distortions than rotation.

2.4 Subsampled randomized Hadamard transform

We saw that Haar random projections have a better performance than iid random projections. However, they are still slow to generate and apply, having a cubic complexity. Can we get the same good statistical performance as Haar projections with faster methods? Here we consider the subsampled randomized Hadamard transform (SRHT) (Ailon and Chazelle, 2006), also known as the Fast Johnson-Lindentsrauss transform (FJLT). This is faster as it relies on the Fast Fourier Transform, and is often viewed as a standard reference point for comparing sketching algorithms.

An $n \times n$ possibly complex-valued matrix H is called a *Hadamard matrix* if $H/\sqrt{n}$ is orthogonal and the absolute values of its entries are unity, $|H_{ij}| = 1$ for $i, j = 1, \ldots, n$. A prominent example, the *Walsh-Hadamard matrix* is defined recursively by

$$H_n = \begin{pmatrix} H_{n/2} & H_{n/2} \\ H_{n/2} & -H_{n/2} \end{pmatrix},$$

with $H_1 = (1)$. This requires n to be a power of 2. Another construction is the discrete Fourier transform (DFT) matrix with the (u, v)-th entry equal to $H_{uv} = n^{-1/2} e^{-2\pi i (u-1)(v-1)/n}$. Multiplying this matrix from the right by X is equivalent to applying the discrete Fourier transform to each column of X, up to scaling. The time complexity for the matrix-matrix multiplication for both the transforms is $O(np \log n)$ due to the Fast Fourier Transform. This is near-quadratic, an order of magnitude faster than other random projections.

Now we consider the subsampled randomized Hadamard transform. Define the $n \times n$ subsampled randomized Hadamard matrix as

$$S = BHDP, \tag{2}$$

where $B \in \mathbb{R}^{n \times n}$ is a diagonal *sampling matrix* of iid Bernoulli random variables with success probability r/n. Then the expected number of sampled rows is r. Also, $H \in \mathbb{R}^{n \times n}$ is the Hadamard matrix; and $D \in \mathbb{R}^{n \times n}$ is a diagonal matrix of iid sign random variables, equal to ± 1 with probability one half. Finally, $P \in \mathbb{R}^{n \times n}$ is a uniformly distributed permutation matrix, i.e. $P_{ij} = \delta_{i,\pi(j)}$ for a permutation π chosen uniformly from the set of n-element permutations. In the definition of S, the Hadamard matrix H is deterministic, while the other matrices B, D and P are random. At the last step, we discard the zero rows of S, so that it becomes an $\tilde{r} \times n$ orthogonal matrix where $\tilde{r} \approx r$. Thus, we can expect the SRHT to be similar to the uniform orthogonal projection. The following theorem verifies our intuition.

Theorem 2.4 (Subsampled randomized Hadamard projection)**.** *Let S be an $n \times n$ subsampled randomized Hadamard matrix. Suppose also that X is an $n \times p$ deterministic matrix whose e.s.d. converges weakly to some fixed probability distribution with compact support bounded away from the origin. Then as n tends to infinity, while $p/n \to \gamma \in (0, 1)$, $r/n \to \xi \in (\gamma, 1)$, the efficiencies have the same limits as for Haar projection in Theorem 2.3.*

See Section 5.6 for the proof, which uses some basic notions from free probability theory (see Section 5.1 for a brief review). Although the subsampled randomized Hadamard matrix has a similar performance to a Haar random matrix, it has much less randomness. To prove the above result, we use an entirely different approach, based on recent results on *asymptotically liberating sequences* from free probability (Anderson and Farrell, 2014). This is a key technical innovation of our work.

The randomized Hadamard projection that we study is slightly different from the standard randomized Hadamard projection, in e.g., Ailon and Chazelle (2006). Our transform first *randomly permutes* the rows of X and Y, before applying the classical transform. This has negligible cost, and *breaks the non-uniformity* in the data. If we assume that the rows of X are permutation-invariant in distribution, our result also holds for the usual randomized Hadamard projection.

2.5 Random sampling

2.5.1 Uniform random sampling

Fast orthogonal transforms such as the Hadamard transforms are considered as a baseline for sketching methods, because they are efficient and work well quite generally. However, if the data are very uniform, for instance if the data matrix can be assumed to be nearly orthogonally invariant, then *sampling methods* can work just as well, as will be shown below.

The simplest sampling method is uniform subsampling, where we take r of the n rows of X with equal probability, with or without replacement. Here we analyze a nearly equivalent method, where we sample each row of X independently with probability r/n, so that the expected number of sampled rows is r. For large r and n, the number of sampled rows concentrates around r.

Moreover, we also assume that X is random, and the distribution of X is *rotationally invariant*, i.e. for any $n \times n$ orthogonal matrix U and any $p \times p$ orthogonal matrix V, the distribution of $UXV^\top$ is the same as the distribution of X. This holds for instance if X has iid Gaussian entries. Then the following theorem states the surprising fact that uniform sampling performs just like Haar projection. See Section 5.7 for the proof.

Theorem 2.5 (Uniform sampling)**.** *Let S be an $n \times n$ diagonal uniform sampling matrix with iid Bernoulli(r/n) entries. Let X be an $n \times p$ rotationally invariant random matrix. Suppose that n tends to infinity, while $p/n \to \gamma \in (0, 1)$, and $r/n \to \xi \in (\gamma, 1)$, and the e.s.d. of X converges almost surely in distribution to a compactly supported probability measure bounded away from the origin. Then the efficiencies have the same limits as for Haar matrices in Theorem 2.3.*

Therefore, when the data is near-orthogonally invariant, we can simply use subsampling instead of complicated random projection methods.

2.5.2 Leverage-based sampling

Uniform sampling can work poorly when the data are highly non-uniform, because some datapoints are more influential than others for the regression fit. There are more advanced sampling methods that sample each row of X with some non-uniform probability π_i which relates to the importance of the ith sample. The simplest non-uniform sampling method is ℓ_2 sampling (Drineas et al., 2006), where π_i is proportional to the squared norm of the ith row, i.e., $\pi_i \propto \|X_{i,\cdot}\|^2$. However, this method does not have the best possible performance (Drineas and Mahoney, 2016). For instance,

when X has a column with only one nonzero entry in the jth row, then any subsample is rank deficient unless it contains that row. However, the ℓ_2 norm of the row may be small.

In linear regression, we can instead measure the importance of the ith row of X by $(X_{i,.}^\top \beta)^2/\|X\beta\|^2$, the contribution of the ith row to the regression function $X\beta$. Since β is unknown, we consider the maximum over all $\beta \in \mathbb{R}^p$. Suppose X is of full column rank and has the SVD decomposition $X = U\Lambda V^\top$. Then

$$\sup_{\beta\in\mathbb{R}^p} \frac{(X_{i,.}^\top \beta)^2}{\|X\beta\|^2} = \sup_{\beta\in\mathbb{R}^p} \frac{(U_{i,.}^\top \Lambda V^\top \beta)^2}{\|U\Lambda V^\top \beta\|^2} = \sup_{\beta\in\mathbb{R}^p} \frac{(U_{i,.}^\top \beta)^2}{\|U\beta\|^2} = \|U_{i,.}\|^2 = x_i^\top (X^\top X)^{-1} x_i = h_{ii},$$

which means that the largest possible contribution of the ith row of X to the regression function is h_{ii}. The quantity h_{ii} is also known as the leverage score, and can be thought of as the "leverage of response value Y_i on the corresponding value $\hat{Y}_i$".

Leverage score sampling is based on this intuition, and samples each row of X with probability $\pi_i \propto h_{ii}$. However, this method suffers from a large variance when the leverage scores are highly non-uniform. Several approaches were proposed to improve this methods. Ma et al. (2015) proposed "shrinkage" leverage score sampling, where $\pi_i = \theta h_{ii} + (1-\theta)\frac{1}{n}$. This method decreases the influence of the large-leverage samples and increases the influence of low-leverage samples, thus reducing the variance. Another approach is to let π_i be proportional to some power of h_{ii} (Chen et al., 2016). One can also take the r rows with largest leverage scores in a deterministic way (see e.g., Papailiopoulos et al., 2014). We call this method greedy leverage sampling.

In this section, we give a unified framework to study these sampling methods. Since leverage-based sampling does not introduce enough randomness for the results to be as simple and universal as before, we need to assume some more randomness via a model for X. Here we consider the *elliptical model*

$$x_i = w_i \Sigma^{1/2} z_i, i = 1, \ldots, n, \tag{3}$$

where the *scale variables* w_i are deterministic scalars bounded away from zero, and $\Sigma^{1/2}$ is a $p \times p$ positive definite matrix. Also, z_i are iid $p \times 1$ random vectors whose entries are all iid random variables of zero mean and unit variance. This model has a long history in multivariate statistics (e.g., Mardia et al., 1979). If a scale variable w_i is much larger than the rest, that point will have a large leverage score, and will "stick out" from the datacloud. Therefore, this model allows us to study the effect of unequal leverage scores. Similarly to uniform sampling, we analyze the model where each row is sampled independently with some probability.

Recall that η-transform of a distribution F is defined by

$$\eta_F(z) = \int \frac{1}{1+zx} dF(x),$$

for $z \in \mathbb{C}^+$. In the next result, we assume that the scalars w_i^2, $i = 1, \ldots, n$, have a limiting distribution F_{w^2} as the dimension increases.

Theorem 2.6 (Sampling for elliptical model)**.** *Suppose X is sampled from the elliptical model defined in* (3)*. Suppose the e.s.d. of Σ converges in distribution to some probability measure with compact support bounded away from the origin. Let n tend to infinity, while $p/n \to \gamma \in (0,1)$ and $r/n \to \xi \in (\gamma, 1)$. Suppose also that the $4+\eta$-th moment of z_i is uniformly bounded, for some $\eta > 0$.*

Consider the sketching method where we sample the i-th row of X with probability π_i independently, where π_i may only depend on w_i, and $\pi_i, i = 1, \dots, n$ have a limiting distribution F_π. Let $s|\pi$ be a Bernoulli random variable with success probability π, then

$$\lim_{n\to\infty} VE(\hat{\beta}_s, \hat{\beta}) = \frac{\eta_{sw^2}^{-1}(1-\gamma)}{\eta_{w^2}^{-1}(1-\gamma)},$$
$$\lim_{n\to\infty} PE(\hat{\beta}_s, \hat{\beta}) = 1 + \frac{1}{\gamma}\mathbb{E}\left[w^2(1-s)\right]\eta_{sw^2}^{-1}(1-\gamma),$$
$$\lim_{n\to\infty} OE(\hat{\beta}_s, \hat{\beta}) = \frac{1 + \mathbb{E}\left[w^2\right]\eta_{sw^2}^{-1}(1-\gamma)}{1 + \mathbb{E}\left[w^2\right]\eta_{w^2}^{-1}(1-\gamma)},$$

where η_{w^2} and η_{sw^2} are the η-transforms of w^2 and sw^2, respectively. Moreover, the expectation is taken with respect to the joint distribution of s, w^2 as defined above.

See Section 5.8 for the proof. Next we use this theorem to study leverage-based sampling (including leverage score sampling, greedy leverage sampling) and uniform sampling for the elliptical model.

Since the leverage scores sum to p, and we want to sample an average of r rows, it is a natural idea to take each row with probability $r/p \cdot h_{ii}$. However, since this success probability can be larger than unity, we instead sample with probability $\min\{r/p \cdot h_{ii}, 1\}$. This decreases slightly the expected number of rows. Therefore, when we compare leverage sampling and uniform sampling with the same parameter r, even an equal performance would favor leverage sampling, as it uses a smaller number of rows.

Corollary 2.7 (Leverage score sampling)**.** *Under the conditions of Theorem 2.6, suppose that for $p < r < n$, we sample the i-th row of X independently with probability* $\min[r/p \cdot h_{ii}, 1]$ *and do linear regression on the resulting subsample of X and Y. Let $s|w$ be a Bernoulli random variable with success probability* $\min\left[\frac{r}{p}\left(1 - \frac{1}{1+w^2\eta_{w^2}^{-1}(1-\gamma)}\right), 1\right]$*. Then the limiting efficiencies have the form given in Theorem 2.6 with the above choice of s.*

See Section 5.8.1 for the proof.

If w_i-s are all equal to unity, one can check that the results are the same as for orthogonal projection or uniform sampling on rotationally invariant X. This is because all leverage scores are nearly equal. For more general w, VE and PE may not be equal.

In addition to randomized leverage-score sampling, it is also of interest to understand the greedy method where we sample the datapoints with the largest leverage scores.

Corollary 2.8 (Greedy leverage sampling)**.** *Under the conditions of Theorem 2.6, suppose that for $p < r < n$, we take the r rows of X with the highest leverage scores and do linear regression on the resulting subsample of X, Y. Let $\tilde{w}^2 = w^2 1_{[w^2 > F_{w^2}^{-1}(1-\xi)]}$ denote the distribution of F_{w^2} truncated*

at $1-\xi$. *Then*

$$\lim_{n\to\infty} VE(\hat{\beta}_s, \hat{\beta}) = \frac{\eta_{\tilde{w}^2}^{-1}(1-\gamma)}{\eta_{w^2}^{-1}(1-\gamma)},$$
$$\lim_{n\to\infty} VE(\hat{\beta}_s, \hat{\beta}) = 1 + \frac{1}{\gamma}\mathbb{E}\left[w^2 1_{[w^2 < F_{w^2}^{-1}(1-\xi)]}\right] \eta_{\tilde{w}^2}^{-1}(1-\gamma/\xi),$$
$$\lim_{n\to\infty} OE(\hat{\beta}_s, \hat{\beta}) = \frac{1+\mathbb{E}\left[w^2\right]\eta_{\tilde{w}^2}^{-1}(1-\gamma)}{1+\mathbb{E}\left[w^2\right]\eta_{w^2}^{-1}(1-\gamma)},$$

where η_{w^2} *and* $\eta_{\tilde{w}^2}$ *are the* η*-transforms of* F_{w^2} *and* $F_{\tilde{w}^2}$*, respectively, and the expectations are taken with respect to those limiting distributions.*

Proof. The proof is exactly the same as for Corollary 2.7. □

The following corollary states the result for uniform sampling under the elliptical model. Note that this result is different from that obtained in Theorem 2.5, where X is assumed to be orthogonally invariant.

Corollary 2.9 (Uniform sampling for elliptical model)**.** *Under the conditions of Theorem 2.6, suppose we sample each row of* X *with probability* r/n *and do linear regression on the resulting subsample. Then*

$$\lim_{n\to\infty} VE(\hat{\beta}_s, \hat{\beta}) = \frac{\eta_{w^2}^{-1}(1-\gamma/\xi)}{\eta_{w^2}^{-1}(1-\gamma)},$$
$$\lim_{n\to\infty} PE(\hat{\beta}_s, \hat{\beta}) = 1 + \frac{1-\xi}{\gamma}\mathbb{E}\left[w^2\right]\eta_{w^2}^{-1}(1-\gamma/\xi),$$
$$\lim_{n\to\infty} OE(\hat{\beta}_s, \hat{\beta}) = \frac{1+\mathbb{E}\left[w^2\right]\eta_{w^2}^{-1}(1-\gamma/\xi)}{1+\mathbb{E}\left[w^2\right]\eta_{w^2}^{-1}(1-\gamma)},$$

Proof. The proof follows directly from Theorem 2.6 and because $\pi_i = r/n$ is independent of z_i. □

The results for leverage-based sampling are not as explicit as our previous results. This makes them harder to interpret and to use. However, as a first step, we can evaluate the formulas numerically for certain models of the leverage scores. This allows us to get a better sense of the advantages of leverage sampling compared to uniform sampling. We will see an example below.

We consider a simple example where w follows a discrete distribution, with $\mathbb{P}[w_i = \pm d_1] = \mathbb{P}[w_i = \pm d_2] = 1/4$. Z is a standard Gaussian random matrix and Σ is the identity matrix. We calculate the efficiencies using our theory in Theorem 2.7, 2.8 and 2.9 (see Section 5.9 for the details). In Figure 2, we plot simulation results as well as our theory for leverage score sampling, greedy leverage scores and uniform sampling. Our theory agrees very well with the simulations.

We also observe that the greedy leverage sampling outperforms random leverage sampling, especially for relatively small r. Moreover, leverage sampling and greedy leverage scores have much better performances than uniform sampling. This is because the leverage scores are highly nonuniform in this example. One can also show that in this example, Haar/Hadamard projections outperform uniform sampling, but perform worse than leverage sampling and greedy leverage. Surprisingly, Gaussian/iid projections work even worse than uniform sampling. This phenomenon also appear in the empirical data analysis, as shown in Figure 9 and 10.

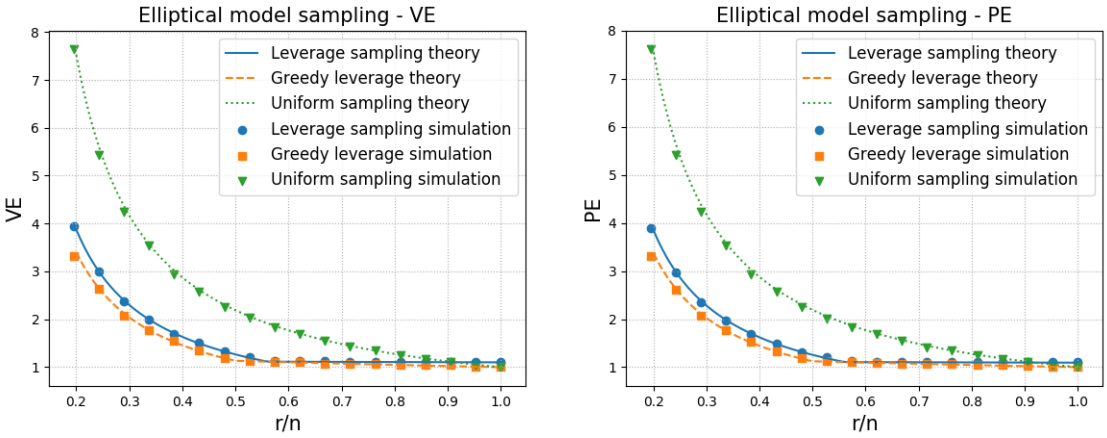

Figure 2: Leverage sampling, greedy leverage sampling and uniform sampling for elliptical model. We generate the data matrix X from the elliptical model defined in (3), and we take $d_1 = 1, d_2 = 3, n = 20000, p = 1000$ while Z is generated with iid $\mathcal{N}(0,1)$ entries and Σ is the identity. We let r range from 4000 to 20000. At each dimension r we repeat the experiments 50 times and take the average. For leverage sampling, we sample each row of X independently with probability $\min(r/p \cdot h_{ii}, 1)$. For greedy leverage scores, we take the r rows of X with the largest leverage scores. For uniform sampling, we uniformly sample r rows of X. The theoretical lines for leverage sampling, greedy leverage scores and uniform sampling are drawn according to Theorems 2.7, 2.8 and 2.9, respectively. We see a good match between theory and simulations.

3 Simulation studies

In this section, we test our theoretical results in some simulation experiments. The code for reproducing these simulations can be found at `https://github.com/liusf15/Sketching-lr`.

In the simulations below, we take $n = 2000$, $p = 100$, with the target dimension r ranging from 300 to 2000. We generate the data matrix X and the coefficients β at the beginning and fix them in the simulation. Here X is generated from a standard Gaussian distribution and β is generated from a uniform distribution on $[0, 1]$. At each dimension r, we randomly generate 50 sketching matrices S, and compute the sketched and full OLS estimators. Since in our theory we take expectations w.r.t. the sketching matrices S as well as the noise ε, we randomly generate the Gaussian noise ε and compute $Y = X\beta + \varepsilon$ in each simulation. Then we calculate the VE, PE, RE and OE defined in Section 1.2. For OE, we generate the test data matrix X_{test}, which also has size $n \times p$, from the standard Gaussian distribution. We generate $\varepsilon_{test} \in \mathbb{R}^n$ from the standard Gaussian distribution and compute $Y_{test} = X_{test}\beta + \varepsilon_{test}$. Although we only show simulation results with Gaussian data due to space limitations, we have also obtained good results for more general X, such as multivariate t distributions.

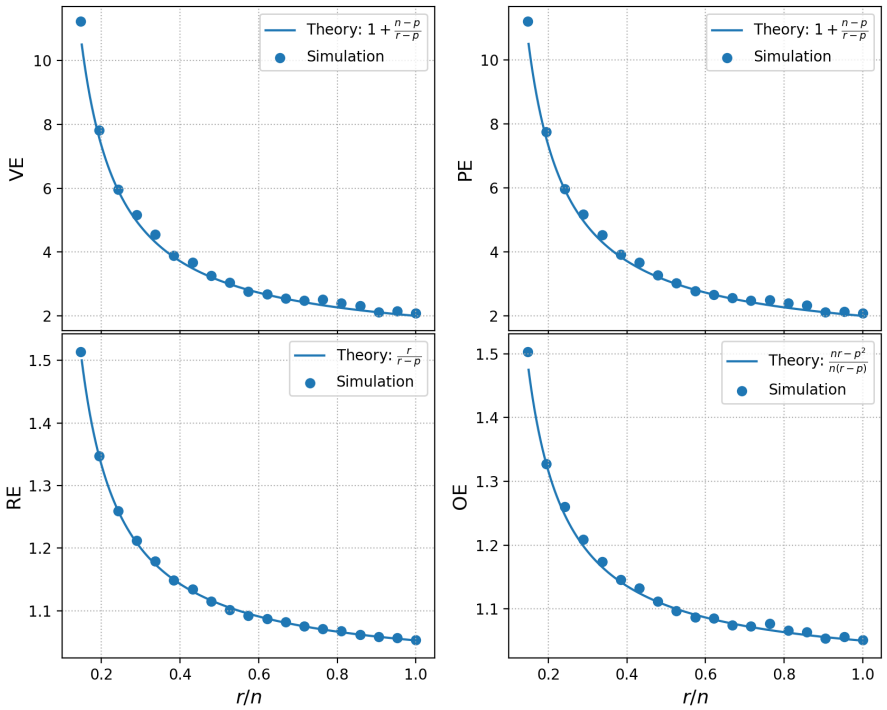

Figure 3: Gaussian projection. In this simulation, we let $n = 2000$, the aspect ratio $\gamma = 0.05$, with r/n ranging from 0.15 to 1. The four subplots plot VE, PE, RE and OE, respectively. The data matrix X is generated from a Gaussian distribution, while the coefficient β is generated from a uniform distribution; both are fixed at the beginning. The relative efficiencies are averaged over 50 simulations. In each simulation, we generate the noise ε as well as the Gaussian sketching matrix S. For OE, we generate n test data points from a Gaussian distribution.

3.1 Gaussian and iid projections

Here we check our theoretical results for Gaussian and iid projections discussed in Sections 2.1, 2.2. In Figure 3, we show results for standard Gaussian projections. We generate the sketching matrix S with iid $\mathcal{N}(0, 1)$ entries. In Figure 4, we show results for S with iid random entries, approximately 10% equal to ± 1, the rest being 0. The theoretical lines for VE, PE, and OE are drawn according to Theorem 2.1 and Theorem 2.2. The theory for RE is given by equation (1). It is shown that our theory agrees well with the simulation results.

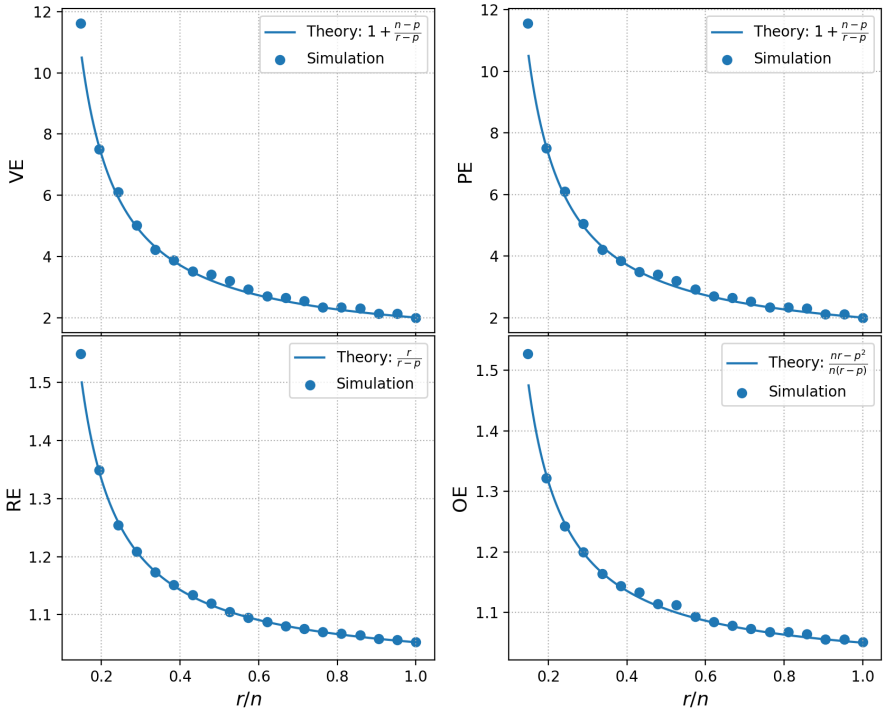

Figure 4: Sparse projection. The protocol is as in Figure 3. The sketching matrix S is sparse, with entries equal ± 1 with equal probability 0.05, or 0 with probability 0.9.

3.2 Orthogonal and Hadamard projections

Here we check our theoretical results for orthogonal and randomized Hadamard projections discussed in Sections 2.3, 2.4. In Figure 5, we show results for orthogonal projection. We generate the sketching matrix S uniformly from the space of $r \times n$ orthogonal matrices. In Figure 6, we show results for subsampled randomized Hadamard projection. We first permute the rows of X and flip the sign of each row with probability 0.5. Then we apply discrete cosine transform to each column of X and next divide the first row by $\sqrt{2}$ to ensure it is an orthogonal transform. The theoretical lines are drawn according to Theorems 2.3 and 2.4.

3.3 Sampling

Here we check our theoretical results for uniform and leverage based sampling from Sections 2.5.1, 2.5.2. In Figure 7, we show results for uniform sampling for orthogonally invariant X. We sample each row of X independently with probability $\frac{r}{n}$. In Figure 8, we show results for leverage based sampling. We sample the i-th row of X independently with probability $\min(\frac{r}{p}h_{ii}, 1)$, where h_{ii} is

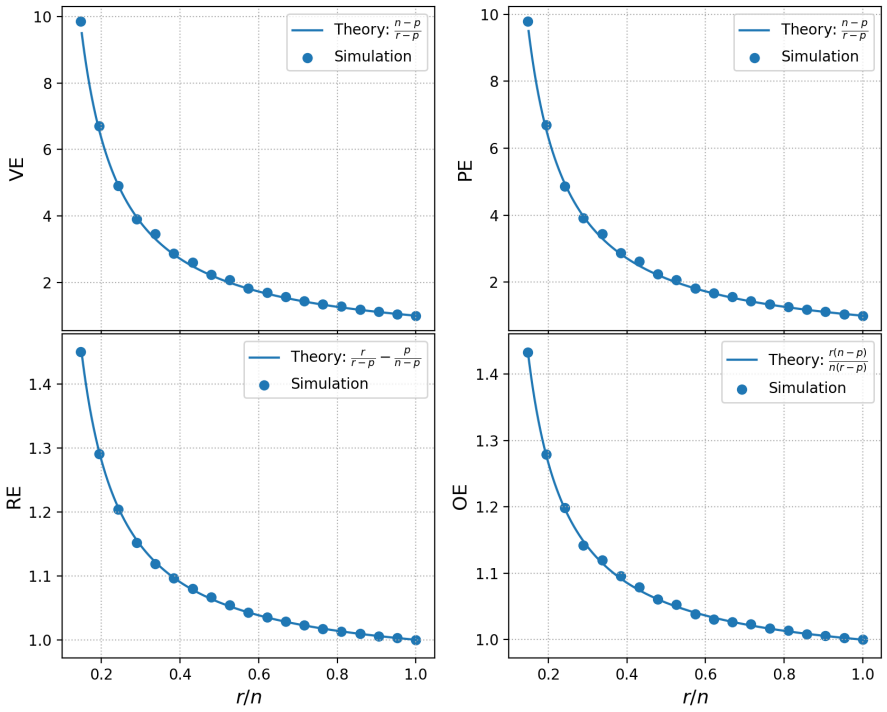

Figure 5: Orthogonal projection. The protocol is as in Figure 3. The sketching matrix S is an uniformly distributed orthogonal projection matrix.

the i-th leverage score of X. We estimate the leverage scores using the algorithms proposed in Drineas et al. (2012). We first apply discrete cosine transform to each column of X, sample 150 rows uniformly to obtain $X_1 \in \mathbb{R}^{150\times 100}$, compute the QR decomposition of X_1 as $X_1 = QR$ and let $\Omega = XR^{-1}$. Let $\ell_i = \|\Omega_{i,\cdot}\|^2$ and normalize them such that $\sum_{i=1}^n \ell_i = p$. Then ℓ_i-s are taken as the approximation of the leverage scores h_{ii}, $i = 1, \ldots, n$. The theoretical lines are drawn according to Theorems 2.5 and Corollary 2.7.

4 Experiments on empirical data

In this section, we verify our theoretical results on two empirical datasets.

4.1 Million Song dataset

We use a subset of the Million Song Year Prediction Dataset (MSD) (Bertin-Mahieux et al., 2011) with $n = 5000$ samples and $p = 90$ features. We compare four different sketching methods: Gaussian

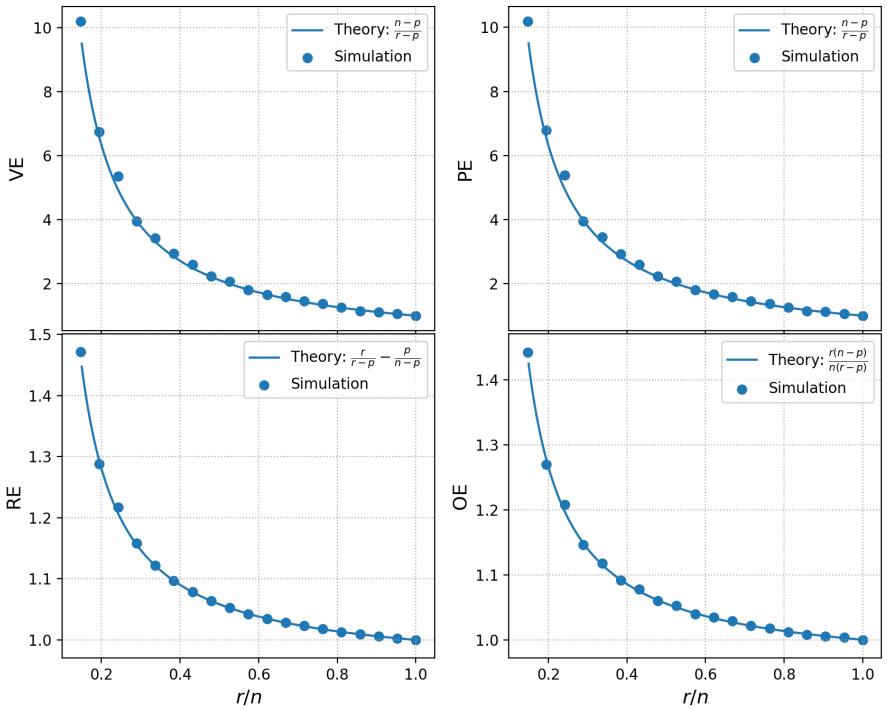

Figure 6: Randomized Hadamard projection. The protocol is as in Figure 3. The sketching matrix S is a subsampled randomized Hadamard projection.

projection, randomized Hadamard projection, uniform sampling and leverage based sampling. The results are shown in Figure 9. For each target dimension r, we average RE and OE over 50 independent experiments. For OE, we randomly choose 5000 test datapoints in each simulation. The theoretical lines are drawn according to Theorems 2.1 and 2.4.

For Gaussian and Hadamard projections, our theory agrees very well with the experimental results, while for uniform sampling and leverage based sampling, our theory is somewhat less accurate. This is because our theory for uniform sampling requires the data matrix X to be rotationally invariant, and for leverage based sampling requires X to follow an elliptical model. These assumptions may not hold for this dataset.

We also have the following observations. In terms of both RE and OE, Hadamard projection is uniformly better than the other sketching methods. Based on this experiment, Hadamard projection seems to be the best method for sketching in linear regression. For small r, Gaussian projection is better than uniform and leverage sampling. But as r/n exceeds around 0.4, Gaussian projection performs worse than the two sampling methods. In terms of RE, leverage sampling is better than uniform sampling when r/n is less than 0.4, but worse when r/n is larger. In terms of OE, leverage

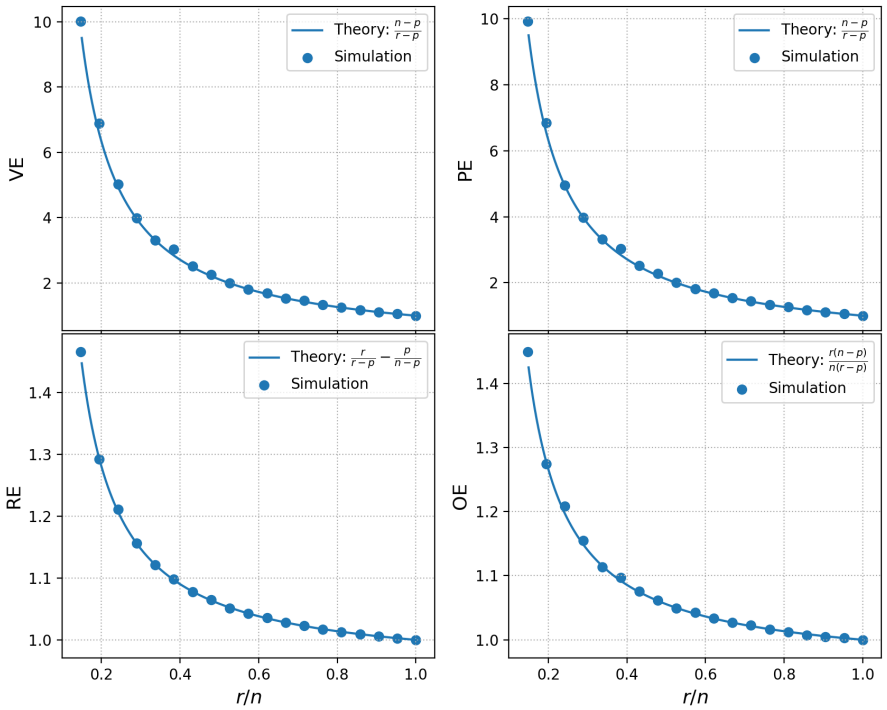

Figure 7: Uniform sampling. The protocol is as in Figure 3. The sketching matrix S samples each row of X independently with probability $\frac{r}{n}$.

sampling is better than uniform sampling when r/n is less than 0.8. When r/n is close to 1, the two sampling methods perform similarly.

4.2 Flights dataset

Here we test our results on the New York flights dataset (Wickham, 2018). We use a subset containing $n = 5000$ samples and $p = 21$ features. The experiments are the same as described in Section 4.1. The results are shown in Figure 10. We see that for RE, with Gaussian and Hadamard projections our theory agrees very well with the experimental results. However, for OE our theory is less accurate. This may be because this dataset is not rotationally invariant. Moreover, the different features are severely multicollinear, which may also cause inaccuracy.

From this experiment, we also conclude that Hadamard projection outperforms the other three sketching methods.

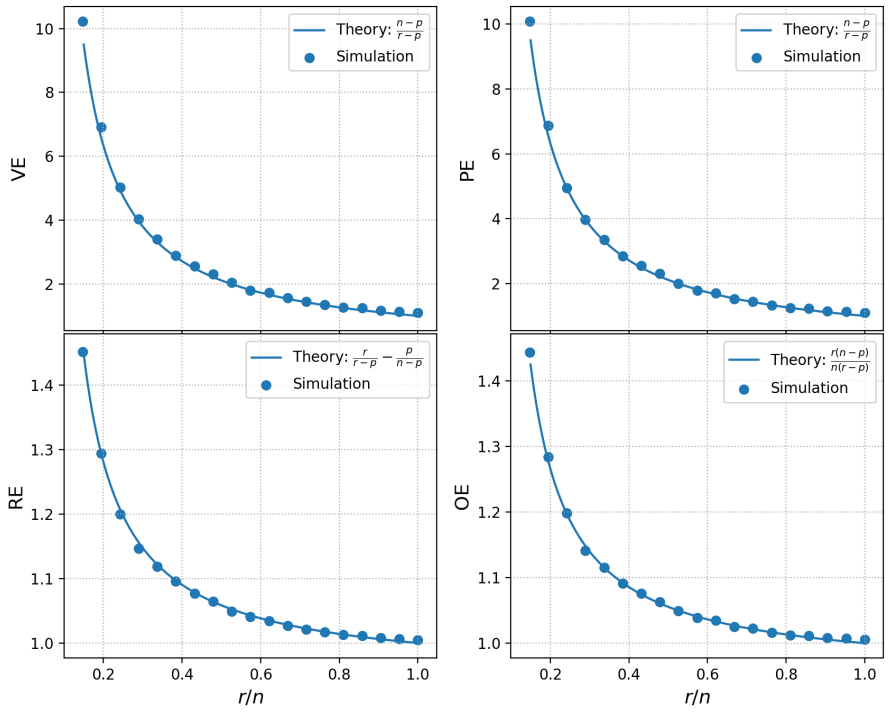

Figure 8: Leverage sampling. The protocol is as in Figure 3. The sketching matrix S samples the rows of X independently with probability $\frac{r}{p}h_{ii}$, where h_{ii}-s are the estimated leverage scores of X.

Acknowledgments

The authors thank Miles Lopes helpful discussions. ED was partially supported by NSF BIGDATA grant IIS 1837992. SL was partially supported by a Tsinghua University Summer Research award.

5 Appendix

5.1 Mathematical background

In this section we introduce a few needed definitions from random matrix theory and free probability. See Bai and Silverstein (2010); Paul and Aue (2014); Yao et al. (2015) for references on random matrix theory and Voiculescu et al. (1992); Hiai and Petz (2006); Nica and Speicher (2006); Anderson et al. (2010) for references on free probability. See also our earlier works (Dobriban, 2015; Liu et al., 2016; Dobriban, 2017b; Dobriban et al., 2017; Dobriban, 2017a; Dobriban and Owen, 2017; Dobriban and Wager, 2018; Dobriban and Sheng, 2018) for applications to statistics.

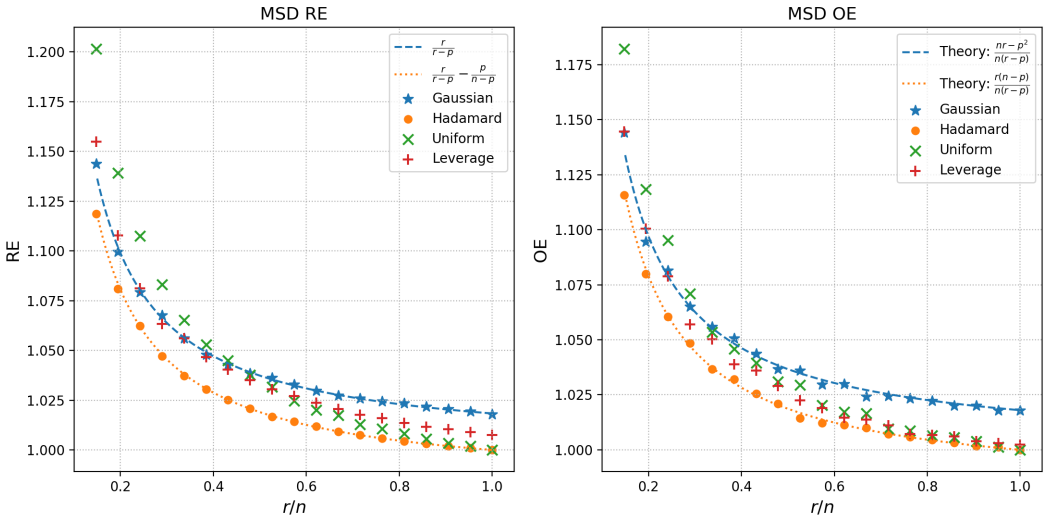

Figure 9: Million Song Dataset. We take $n = 5000$, $p = 90$, with r/n ranging from 0.15 to 1. For each target dimension r, we take the average of RE and OE over 50 simulations. For OE, we randomly choose 5000 test datapoints in each simulation.

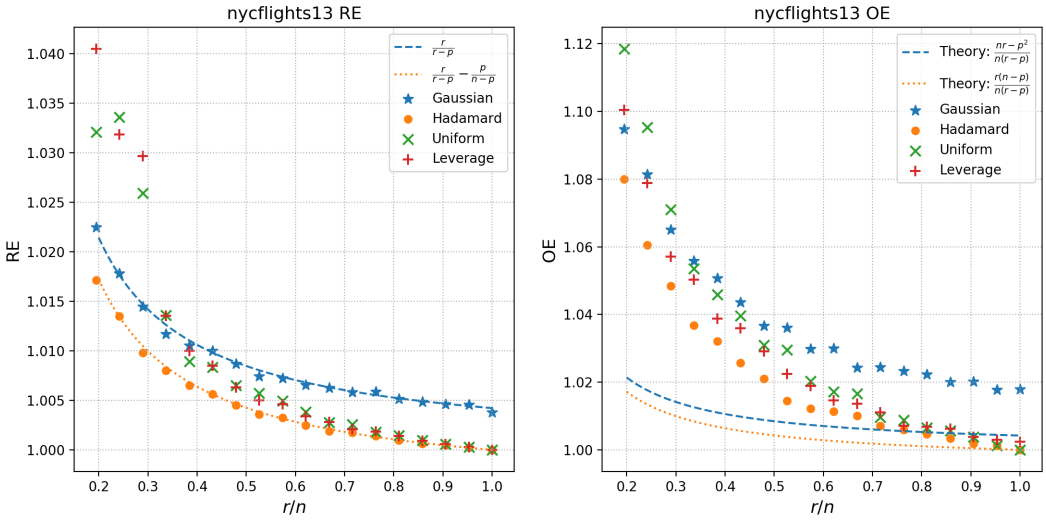

Figure 10: nycflights13 dataset. We take $n = 5000$, $p = 21$, with r/n ranging from 0.15 to 1. At each target dimension r, we repeat the simulation 50 times and take the averaged RE and OE of the four sketching methods. For OE, we randomly choose 5000 testing points (different from the training set) in each simulation.

The data are the $n \times p$ matrix X and contain p features of n samples. Recall that for an $n \times p$ matrix M with $n \geqslant p$, such that the eigenvalues of $n^{-1}M^\top M$ are λ_j, the empirical spectral distribution (e.s.d.) of M is the cdf of the eigenvalues. Formally, it is the mixture $\frac{1}{p}\sum_{j=1}^{p}\delta_{\lambda_j}$ where δ_λ denotes a point mass distribution at λ.

The aspect ratio of X is $\gamma_p = p/n$. We consider limits with $p \to \infty$ and $\gamma_p \to \gamma \in (0, \infty)$. If the e.s.d. converges weakly, as $n, p, \to \infty$, to some distribution F, this is called the limiting spectral distribution (l.s.d.) of X.

The Stieltjes transform of a distribution F is defined for $z \in \mathbb{C}^+ = \{z \in \mathbb{C} : \text{Imag}(z) > 0\}$ as

$$m(z) = \int \frac{dF(x)}{x - z}.$$

This can be used to define the S-transform of a distribution F, which is a key tool for free probability. This is defined as the solution to the equation, which is unique under certain conditions (see Voiculescu et al. (1992)),

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

In addition to Stieltjes transform, there are other useful transforms of a distribution. The η-transform of F is defined by

$$\eta_F(z) = \int \frac{1}{1+zx} dF(x) = \frac{1}{z} m_F(-\frac{1}{z}). \tag{4}$$

Now let us give a typical and key example of a result from asymptotic random matrix theory. Suppose the rows of X are iid p-dimensional observations x_i, for $i = 1, \ldots, n$. Let Σ be the covariance matrix of x_i. We consider a model of the form $X = Z\Sigma^{1/2}$, where the entries of Z are iid with zero mean and unit variance, and the e.s.d. of Σ converges weakly to a probability distribution H. Then the Marchenko-Pastur theorem (see Marchenko and Pastur (1967); Bai and Silverstein (2010)) states that the e.s.d. of the sample covariance matrix $n^{-1}X^\top X$ converges almost surely in distribution to a distribution F_γ, whose Stieltjes transform is the unique solution of a certain fixed point equation. A lot of information can be extracted from this equation, and we will see examples in the proofs.

Random matrix theory is related to free probability. Here we briefly introduce a few concepts in free probability that will be used in the proofs. A non-commutative probability space is a pair $(\mathcal{A}, \tau)$, where $\mathcal{A}$ is a non-commutative algebra with the unit 1 and $\tau : \mathcal{A} \to \mathbb{R}$ is a linear functional such that $\tau(1) = 1$. If $\tau(ab) = \tau(ba)$ for all $a, b \in \mathcal{A}$, then τ is called a trace. If $\tau(a^*a) \geq 0$, for all $a \in \mathcal{A}$ and the equality holds iff $a = 0$, then the trace τ is called faithful. There is also an inner product, and thus a norm, induced by τ:

$$\langle a, b\rangle = \tau(a^*b),\ \|a\|^2 = \langle a, a\rangle.$$

For $a \in \mathcal{A}$ with $a = a^*$, the spectral radius $\rho(a)$ is defined by $\rho(a) = \lim_{k\to\infty} |\tau(a^{2k})|^{\frac{1}{2k}}$, whenever this limit exists. The elements in $\mathcal{A}$ are called (non-commutative) random variables, and the law (or distribution) of a random variable $a \in \mathcal{A}$ is a linear functional on the polynomial algebra $[X]$ that maps any $P(x) \in [X]$ to $\tau(P(a))$. The connection between the non-commutative probability space and classical probability theory is the spectral theorem, stating that for all $a \in \mathcal{A}$ with bounded

spectral radius, there exists a unique Borel probability measure μ_a such that for any polynomial $P(x) \in [X]$,

$$\tau(P(x)) = \int P(t) d\mu_a(t).$$

We can also define the Stieltjes transform of $a \in \mathcal{A}$ by

$$m_a(z) = \tau((a-z)^{-1}) = -\sum_{k=0}^{\infty} \frac{\tau(a^k)}{z^{k+1}},$$

which is the same as the Stieltjes transform of the probability measure μ_a associated with a.

In the space of random matrices, one can easily verify that

$$(\mathcal{A} = (L^{\infty-} \otimes M_n(\mathbb{R})), \tau = \frac{1}{n}\mathbb{E}\,\mathrm{tr})$$

is a non-commutative probability space and $\tau = \frac{1}{n}\mathbb{E}\,\mathrm{tr}$ is a faithful trace, where $L^{\infty-}$ denotes the collection of random variables with all moments finite. For $X \in L^{\infty-} \otimes M_n(\mathbb{R})$, the spectral radius is $\|\|X\|_{op}\|_{L^\infty}$, the essential supremum of the operator norm. The probability measure corresponding to the law of X is the expected empirical spectral distribution

$$\mu_X(t) = \frac{1}{n}\mathbb{E}\sum_{i=1}^{n} \delta_{\lambda_i},$$

where λ_i-s are the eigenvalues of X.

A collection of random variables $\{a_1, \ldots, a_k\} \subset \mathcal{A}$ are said to be freely independent (free) if

$$\tau[\Pi_{j=1}^{m} P_j(a_{i_j} - \tau(P_j(a_{i_j})))] = 0,$$

for any positive integer m, any polynomials $P_1, \ldots, P_m$ and any indices $i_1, \ldots, i_m \in [k]$ with no two adjacent i_j equal. A sequence of random variables $\{a_{1,n}, \ldots, a_{k,n}\}_{n\geq 1} \subset \mathcal{A}$ is said to be asymptotically free if

$$\tau[\Pi_{j=1}^{m} P_j(a_{i_j,n} - \tau(P_j(a_{i_j,n})))] \to 0,$$

for any positive integer m, any polynomials $P_1, \ldots, P_m$ and any indices $i_1, \ldots, i_m \in [k]$ with no two adjacent i_j equal. If $a, b \in \mathcal{A}$ are free, then the law of their product is called their freely multiplicative convolution, and is denoted $a \boxtimes b$.

A fundamental result is that the S-transform of $a \boxtimes b$ equals the products of $S_a(z)$ and $S_b(z)$. In addition, random matrices with sufficiently independent entries and "near-uniformly" distributed eigenvectors tend to be asymptotically free in the high-dimensional limit. This is a powerful tool to find the l.s.d. of a product of random matrices.

5.2 Finite-sample results for fixed matrices

We start with finite-sample results that are true for any fixed sketching matrix S. These results will be fundamental in all remaining work. Later, to simplify these results, we will make probabilistic assumptions. First we find a more explicit form of the relative efficiencies.

Proposition 5.1 (Finite n results)**.** *Taking expectations only over the noise ε and ε_t, fixing X and S, the efficiencies have the following forms:*

$$\begin{aligned}
VE(\hat{\beta}_s, \hat{\beta})|X, S &= \frac{\operatorname{tr}[Q_1]}{\operatorname{tr}[(X^\top X)^{-1}]}, \\
PE(\hat{\beta}_s, \hat{\beta})|X, S &= \frac{\operatorname{tr}[Q_2]}{p}, \\
OE(\hat{\beta}_s, \hat{\beta})|X, S &= \frac{1 + x_t^\top Q_1 x_t}{1 + x_t^\top (X^\top X)^{-1} x_t},
\end{aligned}$$

where $Q_0 = (X^\top S^\top S X)^{-1} X^\top S^\top S$, while $Q_1 = Q_0 Q_0^\top$, and $Q_2 = X Q_1 X^\top$.

Proof. The OLS before and after sketching give the estimators $\hat{\beta}$ and $\hat{\beta}_s$

$$\begin{aligned}
\hat{\beta}_{full} &= (X^\top X)^{-1} X^\top Y = \beta + (X^\top X)^{-1} X^\top \varepsilon, \\
\hat{\beta}_{sub} &= (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \tilde{Y} = \beta + (\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top \varepsilon = \beta + Q_0 \varepsilon.
\end{aligned}$$

We define the "hat" matrices

$$\begin{aligned}
H &= X(X^\top X)^{-1} X^\top, \\
\tilde{H} &= X(\tilde{X}^\top \tilde{X})^{-1} \tilde{X}^\top S = X(X^\top S^\top S X)^{-1} X^\top S^\top S = X Q_0.
\end{aligned}$$

These are both projection matrices, i.e., they satisfy the relations $H^2 = H, \tilde{H}^2 = \tilde{H}$. By our assumptions, we have the following facts:

$$\mathbb{E}_\varepsilon[\varepsilon] = 0_n,\ \mathbb{E}_\varepsilon\left[\varepsilon\varepsilon^\top\right] = \sigma^2 I_n,\ \operatorname{tr}[H] = \operatorname{tr}[\tilde{H}] = p.$$

Therefore, we can calculate as follows.

1. Variance efficiency:

$$\begin{aligned}
\mathbb{E}_\varepsilon\left[\|\hat{\beta} - \beta\|^2\right] &= \mathbb{E}_\varepsilon\left[\|(X^\top X)^{-1} X^\top \varepsilon\|^2\right] = \sigma^2 \operatorname{tr}[(X^\top X)^{-1}] \\
\mathbb{E}_\varepsilon\left[\|\hat{\beta}_s - \beta\|^2\right] &= \mathbb{E}_\varepsilon\left[\|Q_0 \varepsilon\|^2\right] = \sigma^2 \operatorname{tr}(Q_0 Q_0^\top).
\end{aligned}$$

   This proves the formula for VE.

2. Prediction efficiency:

$$\begin{aligned}
\mathbb{E}_\varepsilon\left[\|X\beta - X\hat{\beta}\|^2\right] &= \mathbb{E}_\varepsilon\left[\|H\varepsilon\|^2\right] = \sigma^2 \operatorname{tr}[H] = p\sigma^2, \\
\mathbb{E}_\varepsilon\left[\|X\beta - X\hat{\beta}_s\|^2\right] &= \mathbb{E}_\varepsilon\left[\|\tilde{H}\varepsilon\|^2\right] = \sigma^2 \operatorname{tr}[\tilde{H}^\top \tilde{H}] = \sigma^2 \operatorname{tr}(Q_2).
\end{aligned}$$

   This finishes the calculation for PE.

3. Out-of-sample efficiency:

$$\begin{aligned}
\mathbb{E}_{\varepsilon,\varepsilon_t}\left[(y_t - x_t^\top \hat{\beta})^2\right] &= \mathbb{E}_{\varepsilon,\varepsilon_t}\left[(\varepsilon_t - x_t^\top (X^\top X)^{-1} X^\top \varepsilon)^2\right] \\
&= \mathbb{E}_{\varepsilon,\varepsilon_t}\left[\varepsilon_t^2 + \varepsilon^\top X (X^\top X)^{-1} x_t x_t^\top (X^\top X)^{-1} X^\top \varepsilon\right] \\
&= \sigma^2(1 + x_t^\top (X^\top X)^{-1} x_t), \\
\mathbb{E}_{\varepsilon,\varepsilon_t}\left[(y_t - x_t^\top \hat{\beta}_s)^2\right] &= \mathbb{E}_{\varepsilon,\varepsilon_t}\left[(\varepsilon_t - x_t^\top Q_0 \varepsilon)^2\right] = \sigma^2(1 + x_t^\top Q_0 Q_0^\top x_t).
\end{aligned}$$

This finishes the proof. □

The expressions simplify considerably for orthogonal matrices S. Suppose that S is an $r \times n$ matrix such that $SS^\top = I_r$, then we have the following result:

Proposition 5.2 (Finite n results for orthogonal S)**.** *When S is an orthogonal matrix, the above formulas simplify to*

$$VE = \frac{\mathrm{tr}[(X^\top S^\top SX)^{-1}]}{\mathrm{tr}[(X^\top X)^{-1}]}, \quad PE = \frac{\mathrm{tr}[(X^\top S^\top SX)^{-1}X^\top X]}{p},$$

$$OE = \frac{1 + x_t^\top (X^\top S^\top SX)^{-1} x_t}{1 + x_t^\top (X^\top X)^{-1} x_t}.$$

Proof. Since S satisfies $SS^\top = I_r$, we have $(S^\top S)^2 = S^\top S$. Thus, $Q_1 = Q_0 Q_0^\top = (X^\top S^\top SX)^{-1}$. With this, the results follow directly from Proposition 5.1. □

Actually these formulas hold for any S s.t. $X^\top S^\top SX$ is nonsingular and $S^\top S$ is idempotent.

5.3 Proof of Theorem 2.1

The proof below utilizes the orthogonal invariance of Gaussian matrices and properties of Wishart matrices. For any $X \in \mathbb{R}^{n\times p}$ with $n \geqslant p$ and with full column rank, we have the singular value decomposition (SVD) $X = U\Lambda V^\top$, where $U \in \mathbb{R}^{n\times p}, V \in \mathbb{R}^{p\times p}$ are both orthogonal matrices, while $\Lambda \in \mathbb{R}^{p\times p}$ is a diagonal matrix, whose diagonal entries are the singular values of X. Therefore

$$\begin{aligned}
VE(\hat\beta_s, \hat\beta) &= \frac{\mathbb{E}\left[\mathrm{tr}((X^\top S^\top SX)^{-2}X^\top (S^\top S)^2 X)\right]}{\mathbb{E}\left[\mathrm{tr}[(X^\top X)^{-1}]\right]} \\
&= \frac{\mathbb{E}\left[\mathrm{tr}(\Lambda^{-2}(U^\top S^\top SU)^{-1}U^\top (S^\top S)^2 U(U^\top S^\top SU)^{-1})\right]}{\mathbb{E}\left[\mathrm{tr}(\Lambda^{-2})\right]}, \\
PE(\hat\beta_s, \hat\beta) &= \frac{\mathbb{E}\left[\mathrm{tr}((X^\top S^\top SX)^{-1}X^\top X(X^\top S^\top SX)^{-1}X^\top (S^\top S)^2 X)\right]}{p} \\
&= \frac{\mathbb{E}\left[\mathrm{tr}((U^\top S^\top SU)^{-2}U^\top (S^\top S)^2 U)\right]}{p}, \\
OE(\hat\beta_s, \hat\beta) &= \frac{1 + \mathbb{E}\left[x_t^\top (X^\top S^\top SX)^{-1}X^\top (S^\top S)^2 X(X^\top S^\top SX)^{-1}x_t\right]}{1 + \mathbb{E}\left[x_t^\top (X^\top X)^{-1}x_t\right]} \\
&= \frac{1 + \mathbb{E}\left[x_t^\top V\Lambda^{-1}(U^\top S^\top SU)^{-1}U^\top (S^\top S)^2 U(U^\top S^\top SU)^{-1}\Lambda^{-1}V^\top x_t\right]}{1 + \mathbb{E}\left[x_t^\top V\Lambda^{-2}V^\top x_t\right]}.
\end{aligned}$$

We can see that the first two relative efficiencies do not depend on the right singular vectors of X.

We denote by $U^\perp \in \mathbb{R}^{n\times(n-p)}$ a complementary orthogonal matrix of U, such that $UU^\top + U^\perp U^{\perp\top} = I_n$. Let $S_1 = SU$, $S_2 = SU^\perp$, of sizes $r \times p$, and $r \times (n-p)$, respectively. Then S_1 and S_2 both have iid $\mathcal{N}(0,1)$ entries and they are independent from each other, because of the orthogonal invariance of a Gaussian random matrix. Also note that

$$SS^\top = S(UU^\top + U^\perp U^{\perp\top})S^\top = S_1 S_1^\top + S_2 S_2^\top,$$

and

$$S_1^\top S_1 \sim \mathcal{W}_p(I_p, r), \quad S_2 S_2^\top \sim \mathcal{W}_r(I_r, n-p),$$

where $\mathcal{W}_p(\Sigma, r)$ is the Wishart distribution with r degrees of freedom and scale matrix Σ. Then by the properties of Wishart distribution (e.g., Anderson, 2003), when $r - p > 1$, we have

$$\mathbb{E}\left[(S_1^\top S_1)^{-1}\right] = \frac{I_p}{r-p-1}, \mathbb{E}\left[S_2 S_2^\top\right] = (n-p) I_r.$$

Hence the numerator of VE equals

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{tr}\left(\Lambda^{-2}(U^\top S^\top S U)^{-1} U^\top (S^\top S)^2 U (U^\top S^\top S U)^{-1}\right)\right] \\
&= \mathbb{E}\left[\operatorname{tr}\left(\Lambda^{-2}(S_1^\top S_1)^{-1} S_1 (S_1 S_1^\top + S_2 S_2^\top) S_1^\top (S_1^\top S_1)^{-1}\right)\right] \\
&= \operatorname{tr}\left(\Lambda^{-2}(I_p + \mathbb{E}\left[(S_1^\top S_1)^{-1} S_1^\top S_2 S_2^\top S_1 (S_1^\top S_1)^{-1}\right])\right) \\
&= \operatorname{tr}\left(\Lambda^{-2}(I_p + \mathbb{E}\left[(S_1^\top S_1)^{-1} S_1^\top (n-p) I_p S_1 (S_1^\top S_1)^{-1}\right])\right) \\
&= \operatorname{tr}\left(\Lambda^{-2}(I_p + (n-p)\mathbb{E}\left[(S_1^\top S_1)^{-1}\right])\right) \\
&= \operatorname{tr}\left(\Lambda^{-2}(1 + \frac{n-p}{r-p-1})\right),
\end{aligned}$$

and the denominator $\operatorname{tr}[(X^\top X)^{-1}] = \operatorname{tr}[(V\Lambda^2 V^\top)^{-1}] = \operatorname{tr}(\Lambda^{-2})$, so we have

$$VE(\hat{\beta}_s, \hat{\beta}) = 1 + \frac{n-p}{r-p-1}.$$

Similarly for the numerator of PE, we have

$$\begin{aligned}
\mathbb{E}\left[\operatorname{tr}((U^\top S^\top S U)^{-2} U^\top (S^\top S)^2 U)\right] &= \mathbb{E}\left[\operatorname{tr}((S_1^\top S_1)^{-2} S_1^\top (S_1 S_1^\top + S_2 S_2^\top) S_1)\right] \\
&= \mathbb{E}\left[\operatorname{tr}(I_p + (S_1^\top S_1)^{-2} S_1^\top S_2 S_2^\top S_1)\right] \\
&= p + \operatorname{tr}(\mathbb{E}\left[(S_1^\top S_1)^{-2} S_1^\top (n-p) I_r S_1\right]) \\
&= p + (n-p)\operatorname{tr}(\mathbb{E}\left[(S_1^\top S_1)^{-1}\right]) \\
&= p + \frac{(n-p)p}{r-p-1},
\end{aligned}$$

therefore

$$PE(\hat{\beta}_s, \hat{\beta}) = \frac{p + \frac{(n-p)p}{r-p-1}}{p} = 1 + \frac{n-p}{r-p-1}.$$

For the numerator of OE, note that

$$\begin{aligned}
&\mathbb{E}\left[x_t^\top V \Lambda^{-1} (U^\top S^\top S U)^{-1} U^\top (S^\top S)^2 U (U^\top S^\top S U)^{-1} \Lambda^{-1} V^\top x_t\right] \\
&= \operatorname{tr}[\mathbb{E}\left[(S_1^\top S_1)^{-1} S_1^\top (S_1 S_1^\top + S_2 S_2^\top) S_1 (S_1^\top S_1)^{-1}\right] \Lambda^{-1} V^\top x_t x_t^\top V \Lambda^{-1}] \\
&= \operatorname{tr}[(I_p + \mathbb{E}\left[(S_1^\top S_1)^{-1} S_1^\top (n-p) I_r S_1 (S_1^\top S_1)^{-1}\right]) \Lambda^{-1} V^\top x_t x_t^\top V \Lambda^{-1}] \\
&= \operatorname{tr}[(I_p + \frac{n-p}{r-p-1} I_p) \Lambda^{-1} V^\top x_t x_t^\top V \Lambda^{-1}] \\
&= (1 + \frac{n-p}{r-p-1}) x_t^\top V \Lambda^{-2} V^\top x_t.
\end{aligned}$$

Therefore

$$OE(\hat{\beta}_s, \hat{\beta}) = \frac{1 + (1 + \frac{n-p}{r-p-1}) x_t^\top (X^\top X)^{-1} x_t}{1 + x_t^\top (X^\top X)^{-1} x_t}.$$

Additionally, if $x_t = \Sigma^{1/2} z_t$ and $X = Z\Sigma^{1/2}$, we have $x_t^\top (X^\top X)^{-1} x_t = z_t^\top (Z^\top Z)^{-1} z_t$. Since z_t has iid entries of zero mean and unit variance, we have

$$\mathbb{E}\left[z_t^\top (Z^\top Z)^{-1} z_t\right] = \mathrm{tr}[\mathbb{E}\left[(Z^\top Z)^{-1}\right] \mathbb{E}\left[z_t z_t^\top\right]] = \mathrm{tr}[\mathbb{E}\left[(Z^\top Z)^{-1}\right]]$$

Note that the e.s.d. of $\frac{1}{n} Z^\top Z$ converges almost surely to the standard $Marčenko - Pastur$ law (Marchenko and Pastur, 1967; Bai and Silverstein, 2010) whose Stieltjes transform $m(z)$ satisfies the equation

$$m(z) = \frac{1}{1 - \gamma - z - z\gamma m(z)}$$

for $z \notin [(1 - \sqrt{\gamma})^2, (1 + \sqrt{\gamma})^2]$. Letting $z = 0$, we have $m(0) = 1/(1 - \gamma)$, thus

$$\mathrm{tr}[(\frac{1}{n} ZZ^\top)^{-1}] \xrightarrow{a.s.} \frac{1}{1-\gamma}, \quad \mathrm{tr}[(\frac{1}{p} ZZ^\top)^{-1}] \xrightarrow{a.s.} \frac{\gamma}{1-\gamma}.$$

Therefore $\mathbb{E}\left[x_t^\top (X^\top X)^{-1} x_t\right] \xrightarrow{a.s.} \frac{\gamma}{1-\gamma}$ and almost surely

$$OE(\hat{\beta}_s, \hat{\beta}) \to \frac{1 + (1 + \frac{1-\gamma}{\xi-\gamma}) \frac{\gamma}{1-\gamma}}{1 + \frac{\gamma}{1-\gamma}} = \frac{\xi - \gamma^2}{\xi - \gamma}, \text{ as } n \to \infty.$$

This finishes the proof.

5.4 Proof of Theorem 2.2

The proof idea is to use a Lindeberg swapping argument to show that the results from Gaussian matrices extend to iid matrices provided that the first two moments match.

Since the error criteria are invariant under the scaling of S, we can assume without loss of generality that the entries of S are $n^{-1/2} s_{ij}$, where s_{ij} are iid random variables of zero mean, unit variance, and finite fourth moment. We also let $T = n^{-1/2} t_{ij}$, t_{ij} being iid standard Gaussian random variables, for all $i \in [r]$, $j \in [n]$.

Let s (respectively, t) be the rn-dimensional vector whose entries are s_{ij} (respectively, t_{ij}) aligned by columns. Then there is a bijection from s to S, and from t to T. We already know that the desired results for VE and PE hold if $S = T$, and they only depend on $\mathbb{E}\left[\mathrm{tr}(Q_1)\right]$ and $\mathbb{E}\left[\mathrm{tr}(Q_2)\right]$.

For OE, under the extra assumptions that $X = Z\Sigma^{1/2}$, we already proved in Theorem 2.1 that

$$\mathbb{E}\left[x_t^\top (\frac{1}{p} X^\top X)^{-1} x_t\right] - \mathrm{tr}[(\frac{1}{p} Z^\top Z)^{-1}] \xrightarrow{a.s.} 0,$$
$$\mathbb{E}\left[x_t^\top Q_1 x_t\right] - \mathrm{tr}(Q_1) \xrightarrow{a.s.} 0,$$

so the results for OE will only depend on $\mathbb{E}\left[\mathrm{tr}[Q_1]\right]$ as well. Thus we only need to show that $\mathbb{E}\left[\mathrm{tr}[Q_1(S,X)]\right.$ has the same limit as $\mathbb{E}\left[\mathrm{tr}[Q_1(T,X)]\right]$, and $\mathbb{E}\left[\mathrm{tr}[Q_2(S,X)]\right.$ has the same limit as $\mathbb{E}\left[\mathrm{tr}[Q_2(T,X)]\right]$, as n goes to infinity.

Since SX has a nonzero chance of being singular, it is necessary first to show the universality for a regularized trace. See Section 5.4.1 for the proof of Lemma 5.3 below. In the rest of the proof, we let $N = rn$.

Lemma 5.3 (Universality for regularized trace functionals)**.** *Let $z_n = \frac{i}{n} \in \mathbb{C}$, where i is the imaginary unit. Define the functions $f_N, g_N : \mathbb{R}^N \to \mathbb{R}$ as*

$$f_N(s) = \frac{1}{p}\,\mathrm{tr}[(X^\top S^\top SX - z_n I_p)^{-2} X^\top (S^\top S)^2 X], \tag{5}$$

$$g_N(s) = \frac{1}{p}\,\mathrm{tr}[(X^\top S^\top SX - z_n I_p)^{-1} X^\top X (X^\top S^\top SX - z_n I_p)^{-1} X^\top (S^\top S)^2 X], \tag{6}$$

Then

$$\lim_{n\to\infty} \left|\mathbb{E}\left[f_N(s)\right] - \mathbb{E}\left[f_N(t)\right]\right| = 0,$$
$$\lim_{n\to\infty} \left|\mathbb{E}\left[g_N(s)\right] - \mathbb{E}\left[g_N(t)\right]\right| = 0.$$

Next we show that the regularized trace functionals have the same limit as the ones we want. See Section 5.4.2 for the proof.

Lemma 5.4 (Convergence of trace functionals)**.** *Define the functions $f_\infty, g_\infty : \mathbb{R}^N \to \mathbb{R}$*

$$f_\infty(s) = \frac{1}{p}\,\mathrm{tr}[(X^\top S^\top SX)^{-2} X^\top (S^\top S)^2 X] = \frac{1}{p}\,\mathrm{tr}[Q_1(S,X)], \tag{7}$$

$$g_\infty(s) = \frac{1}{p}\,\mathrm{tr}[(X^\top S^\top SX)^{-1} X^\top X (X^\top S^\top SX)^{-1} X^\top (S^\top S)^2 X] = \frac{1}{p}\,\mathrm{tr}[Q_2(S,X)]. \tag{8}$$

Then

$$\lim_{n\to\infty} \left|\mathbb{E}\left[f_N(s)\right] - \mathbb{E}\left[f_\infty(s)\right]\right| = \lim_{n\to\infty} \left|\mathbb{E}\left[f_N(t)\right] - \mathbb{E}\left[f_\infty(t)\right]\right| = 0,$$
$$\lim_{n\to\infty} \left|\mathbb{E}\left[g_N(s)\right] - \mathbb{E}\left[g_\infty(s)\right]\right| = \lim_{n\to\infty} \left|\mathbb{E}\left[g_N(t)\right] - \mathbb{E}\left[g_\infty(t)\right]\right| = 0.$$

According to lemma 5.3 and 5.4, we know that

$$\lim_{n\to\infty} \frac{1}{p}\mathbb{E}\left[\mathrm{tr}[Q_1(S,X)]\right] = \lim_{n\to\infty} \frac{1}{p}\mathbb{E}\left[\mathrm{tr}[Q_1(T,X)]\right],$$
$$\lim_{n\to\infty} \frac{1}{p}\mathbb{E}\left[\mathrm{tr}[Q_2(S,X)]\right] = \lim_{n\to\infty} \frac{1}{p}\mathbb{E}\left[\mathrm{tr}[Q_2(T,X)]\right],$$

which concludes the proof of Theorem 2.2.

5.4.1 Proof of Lemma 5.3

The proof of this lemma relies on the Lindeberg Principle, similar to the Generalized Lindeberg Principle, Theorem 1.1 of Chatterjee (2006). The first claim shows universality assuming bounded third derivatives.

Lemma 5.5 (Universality theorem)**.** *Suppose s and t are two independent random vectors in $\mathbb{R}^N$ with independent entries, satisfying $\mathbb{E}[s_i] = \mathbb{E}[t_i]$ and $\mathbb{E}\left[s_i^2\right] = \mathbb{E}\left[t_i^2\right]$ for all $1 \leq i \leq N$, and $\mathbb{E}\left[|s_i|^3 + |t_i|^3\right] \leq M < \infty$. Suppose $f_N \in C^3(\mathbb{R}^N, \mathbb{R})$ and $|\frac{\partial^3 f_N}{\partial s_i^3}|$ is bounded above by L_N for all $1 \leq i \leq N$ and almost surely as N goes to infinity,then*

$$|\mathbb{E}[f_N(s) - f_N(t)]| = O(L_N N), \text{ as } N \to \infty.$$

The lemma below shows that the third derivatives are actually bounded for our functions of interest, and that the L_N are of order $N^{-3/2}$.

Since we know the singular values of X are uniformly bounded away from zero and infinity, there exists a constant $c > 0$, such that

$$\frac{1}{c} \leq \sigma_{\min}(X) \leq \sigma_{\max}(X) \leq c.$$

Lemma 5.6 (Bounding the third derivatives)**.** *Let $f_N(s)$ and $g_N(s)$ be defined in* (5) *and* (6)*, where the entries of s are independent, of zero mean, unit variance and finite fourth moment. Then there exists some constant $\phi = \phi(c, \xi, \gamma) > 0$, such that for any partial derivative $\partial_\alpha = \frac{\partial}{\partial_{ij}}$, $\forall i \in [r], j \in [n]$,*

$$|\partial_\alpha^3 f_N| \leq \phi N^{-5/4}, \quad |\partial_\alpha^3 g_N| \leq \phi N^{-5/4}$$

hold almost surely as n goes to infinity.

The above two lemmas conclude the proof of Lemma 5.3. Next we prove them in turn.

Proof. (Proof of Lemma 5.5) The main idea of this proof is borrowed from the proof of Theorem 1.1 of Chatterjee (2006). For each fixed N, We write

$$s = (s_1, \ldots, s_N), \quad t = (t_1, \ldots, t_N).$$

For each $i = 0, 1, \ldots, N$, define

$$z_i = (s_1, \ldots, s_{i-1}, s_i, t_{i+1}, \ldots, t_N),$$
$$z_i^0 = (s_1, \ldots, s_{i-1}, 0, t_{i+1}, \ldots, t_N).$$

Note that $z_0 = t, z_N = s$. By a Taylor expansion, we have almost surely that

$$|f_N(z_i) - f_N(z_i^0) - \partial_i f_N(z_i^0) s_i - \frac{1}{2}\partial_i^2 f_N(z_i^0) s_i^2| \leq \frac{1}{6} L_N |s_i|^3,$$
$$|f_N(z_{i-1}) - f_N(z_i^0) - \partial_i f_N(z_i^0) t_i - \frac{1}{2}\partial_i^2 f_N(z_i^0) t_i^2| \leq \frac{1}{6} L_N |t_i|^3.$$

Thus

$$|f_N(z_i) - f_N(z_{i-1}) - \partial_i f_N(z_i^0)(s_i - t_i) - \frac{1}{2}\partial_i^2 f_N(z_i^0)(s_i^2 - t_i^2)| \le \frac{1}{6}(|s_i|^3 + |t_i|^3)L_N.$$

Since

$$f_N(s) - f_N(t) = \sum_{i=1}^{N} f_N(z_i) - f_N(z_{i-1}),$$

we have

$$|f_N(s) - f_N(t) - \sum_{i=1}^{N} \partial_i f_N(z_i^0)(s_i - t_i) - \sum_{i=1}^{N} \frac{1}{2}\partial_i^2 f_N(z_i^0)(s_i^2 - t_i^2)| \le \sum_{i=1}^{N} \frac{1}{6}(|s_i|^3 + |t_i|^3)L_N$$

almost surely as N goes to infinity. By the bounded convergence theorem, and because the first two moments of s, t match, we have

$$|\mathbb{E}\left[f_N(s) - f_N(t)\right]| \le \frac{1}{6}\mathbb{E}\left[(|s_i|^3 + |t_i|^3)\right] L_N N,$$

thus

$$|\mathbb{E}\left[f_N(s) - f_N(t)\right]| \le O(L_N N).$$

This proves Lemma 5.5. □

Proof. (Proof of Lemma 5.6) We will show that the third derivative of f_N and g_N are both bounded in magnitude by $N^{-5/4}$, or equivalently, $n^{-5/2}$. For any $\alpha = (i, j) \in [r] \otimes [n]$, denote $\partial_\alpha = \frac{\partial}{\partial_{ij}}$. Define

$$G_n(S) = (X^\top S^\top S X - z_n I_p)^{-2} X^\top (S^\top S)^2 X,$$

then we have $f_N(s) = \frac{1}{p}\operatorname{tr}(G_n(S))$ and

$$(X^\top S^\top S X - z_n I_p)^2 G_n(S) = X^\top (S^\top S)^2 X. \tag{9}$$

Take derivative w.r.t. α on both sides and we get

$$\partial_\alpha[(X^\top S^\top S X - z_n I_p)^2] \cdot G_n(S) + (X^\top S^\top S X - z_n I_p)^2 \cdot \partial_\alpha G_n(S) = \partial_\alpha[X^\top (S^\top S)^2 X]. \tag{10}$$

We have

$$\begin{aligned}\partial_\alpha[(X^\top S^\top S X - z_n I_p)^2] &= \partial_\alpha[(X^\top S^\top S X)^2] - 2z_n \partial_\alpha (X^\top S^\top S X) \\ &= \partial_\alpha (X^\top S^\top S X) \cdot (X^\top S^\top S X) + (X^\top S^\top S X) \cdot \partial_\alpha (X^\top S^\top S X) \\ &\quad - 2z_n \partial_\alpha (X^\top S^\top S X),\end{aligned}$$

and

$$\partial_\alpha (X^\top S^\top S X) = X^\top [\partial_\alpha (S^\top) \cdot S + S^\top \cdot \partial_\alpha S] X = X^\top (n^{-1/2} E_{ji} S + S^\top n^{-1/2} E_{ij}) X,$$

where $E_{ij} \in \mathbb{R}^{r\times n}$ whose (i,j)-th entry is 1 and the rest are all zeros, and $E_{ji} = E_{ij}^\top$. Therefore

$$\begin{aligned}\partial_\alpha[(X^\top S^\top SX - z_n I_p)^2] = [&X^\top(E_{ji}S + S^\top E_{ij})XX^\top S^\top SX + X^\top S^\top SXX^\top(E_{ji}S + S^\top E_{ij})X \\ &- 2z_n X^\top(E_{ji}S + S^\top E_{ij})X]n^{-1/2}. \end{aligned} \tag{11}$$

Similarly,

$$\begin{aligned}\partial_\alpha[X^\top(S^\top S)^2 X] &= X^\top[\partial_\alpha(S^\top S)\cdot(S^\top S) + (S^\top S)\cdot\partial_\alpha(S^\top S)]X \\ &= \left\{X^\top[(E_{ji}S + S^\top E_{ij})(S^\top S) + (S^\top S)(E_{ji}S + S^\top E_{ij})]X\right\} n^{-1/2}. \end{aligned} \tag{12}$$

Denoting $P(S) = X^\top S^\top SX$ and $Q(S) = E_{ji}S + S^\top E_{ij}$, substituting (11),(12) into (10), we get

$$\begin{aligned}\partial_\alpha G(S) = (P(S) - z_n I_p)^{-2}\{&X^\top[Q(S)S^\top S + S^\top SQ(S)]X \\ &- [X^\top Q(S)XP(S) + P(S)X^\top Q(S)X - 2z_n X^\top Q(S)X]\}G(S)n^{-1/2}. \end{aligned} \tag{13}$$

Next we will show that the trace of $\partial_\alpha G(S)$ is bounded by $n^{-1/2}$. By the inequality $\|AB\| \le \|A\|\|B\|$ and the lemma 5.7 below, we only need to show that the sum of the absolute values of the eigenvalues of $Q(S)$ and the spectral norms of

$$X^\top X, \quad S^\top S, \quad P(S), \quad (P(S) - z_n I_p)^{-2}, \quad G(S)$$

are all bounded above by some constants only dependent on c and ξ.

Lemma 5.7. *(Trace of products). Suppose A, B are two $n \times n$ diagonalizable complex matrices, then*

$$|\operatorname{tr}(AB)| \le |\lambda|_{\max}(A)\sum_{i=1}^n |\mu_i|,$$

where $|\lambda|_{\max}(A)$ is the largest absolute value of eigenvalues of A and μ_i are the eigenvalues of B.

Note that

$$Q(S) = E_{ji}S + S^\top E_{ij} = e_j S_{i\cdot} + S_{i\cdot}^\top e_j^\top,$$

where e_j is an $n \times 1$ vector with the jth entry equal to 1 and the rest equal to 0, $S_{i\cdot}$ is the ith row of S. The eigenvalues of $Q(S)$ are $S_{ij} \pm \|S_{i\cdot}\|$, according to Lemma 5.8 below.

Lemma 5.8. *(Rank two matrices.) Let $u, v \in \mathbb{R}^n$ and $u^\top v \neq 0$, then the nonzero eigenvalues of $uv^\top + vu^\top$ are $u^\top v \pm \|u\|\|v\|$, both with multiplicity 1.*

First note that $|S_{ij}| \le \sigma_{\max}(S)$ and $\|S_{i\cdot}\| \xrightarrow{a.s.} 1$ by the law of large number. It is also known that as $n \to \infty$ and $r/n \to \xi$, we have

$$\lambda_{\min}(S^\top S) \xrightarrow{a.s.} (1 - \sqrt{\xi})^2, \quad \lambda_{\max}(S^\top S) \xrightarrow{a.s.} (1 + \sqrt{\xi})^2,$$

see Bai and Silverstein (2010). So the sum of the absolute values of the eigenvalues of $Q(S)$ is bounded above by $2(2 + \sqrt{\xi})$, almost surely as n tends to infinity.

By our assumption, the eigenvalues of $X^\top X$ are bounded in the interval $[\frac{1}{c^2}, c^2]$.

Suppose the eigenvalues of $X^\top S^\top SX$ are $\lambda_1 \geq \ldots \geq \lambda_p$. So almost surely,

$$\lambda_p \geq \lambda_{\min}(X^\top X)\lambda_{\min}(S^\top S) \geq \frac{1}{c^2}(1-\sqrt{\xi})^2,$$
$$\lambda_1 \leq \lambda_{\max}(X^\top X)\lambda_{\max}(S^\top S) \leq c^2(1+\sqrt{\xi})^2,$$

Since the complex matrix $X^\top S^\top SX - z_n I_p$ is diagonalizable, and its eigenvalues are $\lambda_1 - z_n, \ldots, \lambda_p - z_n$. Thus the eigenvalues of $(X^\top S^\top SX - z_n I_p)^{-2}$ are $\frac{1}{(\lambda_1 - z_n)^2}, \ldots, \frac{1}{(\lambda_p - z_n)^2}$. Because $\lambda_i \in \mathbb{R}$, $z_n = i/n$ and $|\lambda_i - z_n| > |\lambda_i|$, the largest absolute eigenvalue of $(X^\top S^\top SX - z_n I_p)^{-2}$ is bounded above by $\frac{1}{\lambda_p^2}$, that is, $\|(P(S) - z_n I_p)^{-2}\| \leq \frac{1}{\lambda_p^2} \leq \frac{c^4}{(1-\sqrt{\xi})^4}$.

We also have

$$\begin{aligned}\|G(S)\| &\leq \|(P(S) - z_n I_p)^{-2}\| \|X^\top (S^\top S)^2 X\| \\ &\leq \frac{c^4}{(1-\sqrt{\xi})^4} c^2 (1+\sqrt{\xi})^4 = c^6 \frac{(1+\sqrt{\xi})^4}{(1-\sqrt{\xi})^4}.\end{aligned}$$

Thus $\operatorname{tr}[\partial_\alpha G(S)]$ is bounded by $O(n^{-1/2})$. Since $p/n \to \gamma$, there exists a constant $\phi_1(c, \gamma, \xi)$, such that

$$|f_N| = \frac{1}{p} |\operatorname{tr}[\partial_\alpha G(S)]| \leq \phi_1(c, \gamma, \xi) n^{-3/2}.$$

Next we will bound the second derivative of f_N from above by n^{-2}. Take the second derivative w.r.t. to α on both sides of (9), we have

$$\begin{aligned}\partial_\alpha^2[(X^\top S^\top SX - z_n I_p)^2] \cdot G(S) + 2\partial_\alpha[(X^\top S^\top SX - z_n I_p)^2] \cdot \partial_\alpha G(S) + (X^\top S^\top SX - z_n I_p)^2 \partial_\alpha^2 G(S) \\ = \partial_\alpha^2[X^\top (S^\top S)^2 X], \quad (14)\end{aligned}$$

and thus

$$\begin{aligned}\partial_\alpha^2 G(S) = (X^\top S^\top SX - z_n I_p)^{-2} [\partial_\alpha^2[X^\top (S^\top S)^2 X] - \partial_\alpha^2[(X^\top S^\top SX - z_n I_p)^2] \cdot G(S) - \\ 2\partial_\alpha[(X^\top S^\top SX - z_n I_p)^2] \cdot \partial_\alpha G(S)]. \quad (15)\end{aligned}$$

Using (11), we have

$$\begin{aligned}\partial_\alpha^2[(X^\top S^\top SX - z_n I_p)^2] &= \partial_\alpha[X^\top (E_{ji}S + S^\top E_{ij}) X X^\top S^\top SX + X^\top S^\top SX X^\top (E_{ji}S + S^\top E_{ij}) X \\ &\quad - 2z_n X^\top (E_{ji}S + S^\top E_{ij}) X] n^{-1/2} \\ &= \{X^\top (E_{ji}E_{ij} + E_{ji}E_{ij}) X X^\top S^\top SX + \\ &\quad X^\top (E_{ji}S + S^\top E_{ij}) X X^\top (E_{ji}S + S^\top E_{ij}) X + \\ &\quad X^\top (E_{ji}S + S^\top E_{ij}) X X^\top (E_{ji}S + S^\top E_{ij}) X + \\ &\quad X^\top S^\top SX X^\top (E_{ji}E_{ij} + E_{ji}E_{ij}) X - \\ &\quad 2z_n X^\top (E_{ji}E_{ij} + E_{ji}E_{ij}) X\} \frac{1}{n} \\ &= \{2(X^\top (E_{ji}S + S^\top E_{ij}) X)^2 \\ &\quad + 2X^\top E_{jj} X X^\top S^\top SX + 2X^\top S^\top SX X^\top E_{jj} X - 4z_n X^\top E_{jj} X\} \frac{1}{n}.\end{aligned}$$

Using (12), we have

$$\begin{aligned}\partial_\alpha^2[X^\top(S^\top S)^2X] &= \partial_\alpha[\{X^\top[(E_{ji}S+S^\top E_{ij})(S^\top S)+(S^\top S)(E_{ji}S+S^\top E_{ij})]X\}]n^{-1/2}\\ &= X^\top[2E_{jj}S^\top S+2(E_{ji}S+S^\top E_{ij})^2+2S^\top SE_{jj}]X\frac{1}{n}.\end{aligned}$$

By the same arguments, we can show that the traces of the three terms on the right hand side of (15) are bounded above by n^{-1} in magnitude, therefore the second derivative of f_N is bounded by n^{-2}. Also by the same reasoning, we can show that there exists some constant $\phi_3(c,\xi,\gamma)$, such that $|\partial_\alpha^3 f_N(s)| \le \phi_3(c,\xi,\gamma)N^{-5/4}$, holds almost surely as n goes to infinity.

We then use similar methods to bound the third derivative of $g_N(s)$. Define

$$H_n(S) = (X^\top S^\top SX - z_nI_p)^{-1}X^\top X(X^\top S^\top SX - z_nI_p)^{-1}X^\top(S^\top S)^2X,$$

then

$$g_N(s) = \frac{1}{p}\operatorname{tr}[H_n(s)].$$

Note also that

$$(X^\top S^\top SX - z_nI_p)(X^\top X)^{-1}(X^\top S^\top SX - z_nI_p)H_n(S) = X^\top(S^\top S)^2X.$$

Taking derivative w.r.t. to α on both sides we have

$$\begin{aligned}&n^{-1/2}[X^\top(E_{ji}S+S^\top E_{ij})X(X^\top X)^{-1}(X^\top S^\top SX - z_nI_p)H_n(S)+\\ &(X^\top S^\top SX - z_nI_p)(X^\top X)^{-1}X^\top(E_{ji}S+S^\top E_{ij})XH_n(S)]+\\ &(X^\top S^\top SX - z_nI_p)(X^\top X)^{-1}(X^\top S^\top SX - z_nI_p)\partial_\alpha H_n(S)\\ &= n^{-1/2}[X^\top(E_{ji}S+S^\top E_{ij})S^\top SX + X^\top S^\top S(E_{ji}S+S^\top E_{ij})X].\end{aligned}$$

Using similar techniques, we can show that almost surely $\frac{1}{p}|\operatorname{tr}[\partial_\alpha H_n(S)]|$ is bounded in magnitude by $n^{-3/2}$, $\frac{1}{p}|\operatorname{tr}[\partial_\alpha^2 H_n(S)]|$ is bounded in magnitude by n^{-2}, and $\frac{1}{p}|\operatorname{tr}[\partial_\alpha^3 H_n(S)]|$ is bounded in magnitude by $n^{-5/2}$. Therefore almost surely $|\partial_\alpha^3 g_N(s)| \le \phi_3' N^{-5/4}$, for some $\phi_3' = \phi_3'(c,\xi,\gamma)$. Take $\phi = \max(\phi_3, \phi_3')$, and the proof of Lemma 5.6 is done.

□

Proof. (Proof of Lemma 5.7) Consider the eigendecompositions of A, B,

$$A = Q\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{pmatrix}Q^\top,\ B = P\begin{pmatrix}\mu_1 & & \\ & \ddots & \\ & & \mu_n\end{pmatrix}P^\top,$$

then

$$\operatorname{tr}(AB) = \operatorname{tr}(Q\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{pmatrix}Q^\top P\begin{pmatrix}\mu_1 & & \\ & \ddots & \\ & & \mu_n\end{pmatrix}P^\top).$$

Denote the n columns of $Q^\top P$ as $v_1, \ldots, v_n$, which are orthonormal. Then

$$\begin{aligned}|\operatorname{tr}(AB)| &= |\operatorname{tr}(\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} \sum_{i=1}^n \mu_i v_i v_i^\top)| \\ &= |\sum_{i=1}^n \mu_i v_i^\top \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} v_i| \leq \sum_{i=1}^n |\mu_i||\lambda|_{\max}(A).\end{aligned}$$

This finishes the proof. □

Proof. (Proof of Lemma 5.8) It is easy to see that $uv^\top + vu^\top$ has rank 2 and

$$\begin{aligned}(uv^\top + vu^\top)(\frac{u}{\|u\|} + \frac{v}{\|v\|}) &= (u^\top v + \|u\|\|v\|)(\frac{u}{\|u\|} + \frac{v}{\|v\|}), \\ (uv^\top + vu^\top)(\frac{u}{\|u\|} - \frac{v}{\|v\|}) &= (u^\top v - \|u\|\|v\|)(\frac{u}{\|u\|} - \frac{v}{\|v\|}).\end{aligned}$$

This finishes the proof. □

5.4.2 Proof of Lemma 5.4

Let $A = X^\top S^\top S X$ and $B = X^\top S^\top S X - z_n I_n$, and note that we have the relationship

$$A^{-2} - B^{-2} = B^{-1}(B - A)A^{-2} + B^{-2}(B - A)A^{-1} = -z_n(B^{-1}A^{-2} + B^{-2}A^{-1}).$$

Thus

$$\begin{aligned}f_N(s) - f_\infty(t) &= \frac{1}{p}\operatorname{tr}[(A^{-2} - B^{-2})X^\top (S^\top S)^2 X] \\ &= -z_n \frac{1}{p}\operatorname{tr}[(B^{-1}A^{-2} + B^{-2}A^{-1})X^\top (S^\top S)^2 X].\end{aligned}$$

If the eigenvalues of A are $\lambda_1 \geq \ldots \geq \lambda_p > 0$, then the eigenvalues of B are $\lambda_1 - z_n, \ldots, \lambda_p - z_n$. By Lemma 5.7, we have

$$\begin{aligned}\frac{1}{p}|\operatorname{tr}[B^{-1}A^{-2}X^\top (S^\top S)^2 X]| &\leq \|A^{-2}X^\top (S^\top S)^2 X\| \frac{1}{p}\sum_{i=1}^p \frac{1}{|\lambda_i - z_n|} \\ &\leq \frac{1}{\lambda_p^2}\|X^\top X\|\|S^\top S\|^2 \frac{1}{\lambda_p}.\end{aligned}$$

Recall that $\lambda_p \geq \frac{1}{c^2}(1 - \sqrt{\xi})^2$, then we have

$$\frac{1}{p}|\operatorname{tr}[B^{-1}A^{-2}X^\top (S^\top S)^2 X]| \leq c^8 \frac{(1 + \sqrt{\xi})^4}{(1 - \sqrt{\xi})^6}.$$

By the same argument, we have

$$\frac{1}{p}|\operatorname{tr}[B^{-2}A^{-1}X^\top (S^\top S)^2 X]| \leq c^8 \frac{(1 + \sqrt{\xi})^4}{(1 - \sqrt{\xi})^6}.$$

Hence

$$|f_N(s) - f_\infty(s)| \leq \frac{1}{p} 2c^8 \frac{(1+\sqrt{\xi})^4}{(1-\sqrt{\xi})^6}$$

holds almost surely. Hence, $f_N(s) - f_\infty(s) \xrightarrow{a.s.} 0$. By the bounded convergence theorem, we have $\lim_{n\to\infty} |\mathbb{E}[f_N(s)] - \mathbb{E}[f_\infty(s)]| = 0$. The other three limit statements can be proved similarly. This finishes the proof.

5.5 Proof of Theorem 2.3

Suppose that X has the SVD factorization $X = U\Lambda V^\top$ and let $S_1 = SU$. The majority of the proof will deal with the following quantities:

$$\begin{aligned}
\operatorname{tr}[(X^\top X)^{-1}] &= \operatorname{tr}(\Lambda^{-2}), \\
\operatorname{tr}[(X^\top S^\top S X)^{-1}] &= \operatorname{tr}[(\Lambda U^\top S^\top S U \Lambda)^{-1}] = \operatorname{tr}[(\Lambda S_1^\top S_1 \Lambda)^{-1}], \\
\operatorname{tr}[(X^\top S^\top S X)^{-1} X^\top X] &= \operatorname{tr}[(U^\top S^\top S U)^{-1}] = \operatorname{tr}[(S_1^\top S_1)^{-1}].
\end{aligned}$$

Since we are finding the limits of these quantities, we add the subscript n to matrices like S_n, U_n from now on. Since both S_n and U_n are rectangular orthogonal matrices, we embed them into full orthogonal matrices as

$$\mathbb{S}_n = \begin{pmatrix} S_n \\ S_n^\perp \end{pmatrix}, \mathbb{U}_n = \begin{pmatrix} U_n \\ U_n^\perp \end{pmatrix}.$$

Suppose $\frac{1}{p}\Lambda_n S_{1,n}^\top S_{1,n} \Lambda_n$ has an l.s.d. bounded away from zero. Then, the limit of $\frac{1}{p}\operatorname{tr}[(\frac{1}{p}\Lambda_n S_{1,n}^\top S_{1,n}\Lambda_n)^{-1}]$ must equal to the Stieltjes transform of its l.s.d. evaluated at zero. Therefore, we first find the Stieltjes transforms of the l.s.d. of the matrices $\frac{1}{p}\Lambda_n S_{1,n}^\top S_{1,n}\Lambda_n$. The same applies to $\operatorname{tr}[(S_{1,n}^\top S_{1,n})^{-1}]$, except that we replace Λ_n with the identity matrix.

Since $\Lambda_n S_{1,n}^\top S_{1,n}\Lambda_n$ and $S_{1,n}\Lambda_n^2 S_{1,n}^\top$ have the same non-zero eigenvalues, we first find the l.s.d. of $\frac{1}{n} S_{1,n}\Lambda_n^2 S_{1,n}^\top$. Note that

$$\begin{aligned}
S_{1,n} = S_n U_n &= \begin{pmatrix} I_r & 0 \end{pmatrix} \begin{pmatrix} S_n \\ S_n^\perp \end{pmatrix} \begin{pmatrix} U_n & U_n^\perp \end{pmatrix} \begin{pmatrix} I_p \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} I_r & 0 \end{pmatrix} \mathbb{S}_n \mathbb{U}_n \begin{pmatrix} I_p \\ 0 \end{pmatrix}.
\end{aligned}$$

Let $\mathbb{W}_n = \mathbb{S}_n \mathbb{U}_n$, which is again an $n \times n$ Haar-distributed matrix due to the orthogonal invariance of the Haar distribution. Then

$$S_{1,n}\Lambda_n^2 S_{1,n}^\top = \begin{pmatrix} I_r & 0 \end{pmatrix} \mathbb{W}_n \begin{pmatrix} I_p \\ 0 \end{pmatrix} \Lambda_n^2 \begin{pmatrix} I_p & 0 \end{pmatrix} \mathbb{W}_n^\top \begin{pmatrix} I_r \\ 0 \end{pmatrix}.$$

Define

$$C_n = \frac{1}{n} \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \mathbb{W}_n \begin{pmatrix} \Lambda_n^2 & 0 \\ 0 & 0 \end{pmatrix} \mathbb{W}_n^\top \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} = \frac{1}{n} \begin{pmatrix} S_{1,n}\Lambda_n^2 S_{1,n}^\top & 0 \\ 0 & 0 \end{pmatrix}. \tag{16}$$

Since X has an l.s.d., we get that the e.s.d. of $\begin{pmatrix} \Lambda_n^2 & 0 \\ 0 & 0 \end{pmatrix}$ converges to some fixed distribution F_Λ, and we know that the e.s.d. of $\begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$ converges to $F_\xi = \xi\delta_1 + (1-\xi)\delta_0$. Then according to Hachem (2008) or Theorem 4.11 of Couillet and Debbah (2011), the e.s.d. of C_n converges to a distribution F_C, whose η-transform η_C is the unique solution of the following system of equations, defined for all $z \in \mathbb{C}^+$:

$$\begin{aligned}
\eta_C(z) &= \int \frac{1}{z\gamma(z)t+1} dF_\xi(t) = \frac{\xi}{z\gamma(z)+1} + (1-\xi), \\
\gamma(z) &= \int \frac{t}{\eta_C(z) + z\delta(z)t} dF_\Lambda(t), \\
\delta(z) &= \int \frac{t}{z\gamma(z)t+1} dF_\xi(t) = \frac{\xi}{z\gamma(z)+1}.
\end{aligned}$$

Moreover, we note that if the support of F_Λ outside of the point mass at zero is bounded away from the origin, then the same is also true for F_C. Indeed, this follows directly from the form of $\Lambda_n S_{1,n}^\top S_{1,n} \Lambda_n$, as its smallest eigenvalue can be bounded below as

$$\lambda_{\min}(\Lambda_n S_{1,n}^\top S_{1,n} \Lambda_n) \geqslant \lambda_{\min}(\Lambda_n)^2 \lambda_{\min}(S_{1,n}^\top S_{1,n}).$$

Moreover, by assumption $\lambda_{\min}(\Lambda_n) > c > 0$ for some universal constant c, and clearly $\lambda_{\min}(S_{1,n}^\top S_{1,n}) = 1$, as $S_{1,n}$ is a partial orthogonal matrix. This ensures that we can use the Stieltjes transform as a tool to calculate the limiting traces of the inverse.

Returning to our equations, using the first and the third equations to solve for $\delta(z)$ and $\gamma(z)$ in terms of $\eta_C(z)$, substituting them in the second equation, we get the following fixed point equation

$$\eta_C(z) = \eta_\Lambda(z(1 + \frac{\xi - 1}{\eta_C(z)}). \tag{17}$$

According to the definition of η-transform (4), for any distribution F with a point mass $f_F(0)$ at zero, we have

$$\eta_F(z) = \int_{t \neq 0} \frac{1}{1+zt} dF(t) + f_F(0).$$

Note that $f_C(0) = f_\Lambda(0) = 1 - \gamma$. Since the l.s.d. of X is compactly supported and bounded away from the origin, we know $\inf[supp(f_\Lambda) \cap \mathbb{R}^*]$ and $\inf[supp(f_\Lambda) \cap \mathbb{R}^*]$ are greater than zero, thus $\frac{1}{t}$ is integrable on the set $\{t > 0\}$ w.r.t. F_Λ and F_C. Since $|\frac{z}{1+tz}| < \frac{1}{t}$ when $z > 0, t > 0$, by the dominated convergence theorem we have

$$\begin{aligned}
\lim_{z\to\infty} \int_{t\neq 0} \frac{z}{1+tz} dF_C(t) &= \int_{t\neq 0} \frac{1}{t} dF_C(t), \\
\lim_{z\to\infty} \int_{t\neq 0} \frac{z}{1+tz} dF_\Lambda(t) &= \int_{t\neq 0} \frac{1}{t} dF_\Lambda(t),
\end{aligned}$$

and hence

$$\int_{t\neq 0} \frac{1}{t} dF_C(t) = \lim_{z\to\infty} z(\eta_C(z) - (1-\gamma)), \tag{18}$$

$$\int_{t\neq 0} \frac{1}{t} dF_\Lambda(t) = \lim_{z\to\infty} z(\eta_\Lambda(z) - (1-\gamma)), \tag{19}$$

and

$$\begin{aligned}\lim_{z\to\infty} \eta_C(z) &= \lim_{z\to\infty} \int_{t\neq 0} \frac{1}{1+zt} dF_C(t) + (1-\gamma) \\ &= \int_{t\neq 0} \lim_{z\to\infty} \frac{1}{1+zt} dF_C(t) + (1-\gamma) \\ &= 1-\gamma. \end{aligned} \tag{20}$$

Subtracting $1-\gamma$ from both sides of (17), multiplying by $z(1+\frac{\xi-1}{\eta_C(z)})$, letting $z\to\infty$, we obtain

$$\lim_{z\to\infty} z(1+\frac{\xi-1}{\eta_C(z)})[\eta_C(z)-(1-\gamma)] = \lim_{z\to\infty} z(1+\frac{\xi-1}{\eta_C(z)})[\eta_\Lambda(z(1+\frac{\xi-1}{\eta_C(z)})) - (1-\gamma)].$$

Note that RHS equals $\int_{t\neq 0} \frac{1}{t} dF_\Lambda(t)$ by (19), and

$$\begin{aligned} LHS &= \lim_{z\to\infty} z(1+\frac{\xi-1}{\eta_C(z)})[\eta_C(z)-(1-\gamma)] \\ &= \lim_{z\to\infty} z[\eta_C(z)-(1-\gamma)](1+\frac{\xi-1}{1-\gamma}) \\ &= \int_{t\neq 0} \frac{1}{t} dF_C(t) \frac{\xi-\gamma}{1-\gamma}, \end{aligned}$$

where the second and the third equations follow from (20) and (19). This shows that

$$\int_{t\neq 0} \frac{1}{t} dF_\Lambda(t) = \frac{\xi-\gamma}{1-\gamma} \int_{t\neq 0} \frac{1}{t} dF_C(t),$$

therefore we have proved that as $n\to\infty$,

$$\frac{\mathrm{tr}[(\Lambda S_1^\top S_1^\top \Lambda)^{-1}]}{\mathrm{tr}(\Lambda^{-2})} \to \frac{\int_{t\neq 0} \frac{1}{t} dF_C(t)}{\int_{t\neq 0} \frac{1}{t} dF_\Lambda(t),} = \frac{1-\gamma}{\xi-\gamma},$$

thus

$$\lim_{n\to\infty} VE(\hat{\beta}_s, \hat{\beta}) = \frac{1-\gamma}{\xi-\gamma}.$$

This finishes the evaluation of VE.

Next, to evaluate of PE, we argue as follows: In the definition of C_n in (16), replace Λ_n by the identity matrix. Since the results do not depend the l.s.d. of Λ_n, it follows directly that

$$PE = \frac{\mathrm{tr}[(X^\top S^\top S X)^{-1} X^\top X]}{p} = \frac{\mathrm{tr}[(S_1^\top S_1)^{-1}]}{\mathrm{tr}(I_p)} \to \frac{1-\gamma}{\xi-\gamma}.$$

Next, to evaluate the limit of OE, we use the additional assumption on X, that is, $X = Z\Sigma^{1/2}$, where Z has iid entries of zero mean, unit variance and finite fourth moment.

Note that (with convergence below always meaning almost sure convergence)

$$\mathbb{E}\left[x_t^\top (X^\top X)^{-1} x_t)\right] \to \frac{\gamma}{1-\gamma},$$

which has been proved in Section 5.3, and

$$1 + \mathbb{E}\left[x_t^\top (X^\top X)^{-1} x_t\right] \to 1 + \frac{\gamma}{1-\gamma} = \frac{1}{1-\gamma}.$$

On the other hand,

$$\begin{aligned}
&\mathbb{E}\left[x_t^\top (X^\top S^\top S X)^{-1} x_t\right] = \operatorname{tr}(\mathbb{E}\left[X^\top S^\top S X\right]^{-1} \mathbb{E}\left[x_t x_t^\top\right]) \\
&= \operatorname{tr}(\mathbb{E}\left[(\Sigma^{1/2} Z^\top S^\top S Z \Sigma^{1/2})^{-1}\right] \Sigma) = \operatorname{tr}(\mathbb{E}\left[Z^\top S^\top S Z\right]^{-1}).
\end{aligned}$$

Define $C_n = \frac{1}{n} Z^\top S^\top S Z$, then the e.s.d. of C_n converges to a distribution F_C, whose Stieltjes transform $m(z) = m_C(z), z \in \mathbb{C}^+$ is given by (Bai and Silverstein, 2010)

$$m(z) = \frac{1}{\int \frac{s}{1+\gamma s e} dF_{S^\top S}(s) - z} = \frac{1}{\frac{\xi}{1+\gamma e} - z},$$

where

$$e = \frac{1}{\int \frac{s}{1+\gamma s e} dF_{S^\top S}(s) - z} = \frac{1}{\frac{\xi}{1+\gamma e} - z}.$$

And here $F_{S^\top S}$ is the l.s.d. of $S^\top S$, which is $\xi\delta_1 + (1-\xi)\delta_0$. Solving these equations gives

$$m(z) = e(z) = \frac{\xi - \gamma - z + \sqrt{(\xi - \gamma - z)^2 - 4z\gamma}}{2z\gamma}.$$

Therefore

$$\lim_{z \to 0} m(z) = \frac{-1 - \frac{2(\gamma-\xi)-4\gamma}{2(\xi-\gamma)}}{2\gamma} = \frac{-1 + \frac{\xi+\gamma}{\xi-\gamma}}{2\gamma} = \frac{1}{\xi - \gamma}.$$

Thus

$$\operatorname{tr}((Z^\top S^\top S Z)^{-1}) = \frac{1}{n} \operatorname{tr}((\frac{1}{n} Z^\top S^\top S Z)^{-1}) \xrightarrow{a.s.} \gamma m_C(0) = \frac{\gamma}{\xi - \gamma}.$$

Therefore

$$1 + \mathbb{E}\left[x_t^\top (X^\top S^\top S X)^{-1} x_t\right] \to 1 + \frac{\gamma}{\xi - \gamma} = \frac{1}{1 - \gamma/\xi},$$

and we have proved

$$\lim_{n \to \infty} OE(\hat{\beta}_s, \hat{\beta}) = \lim_{n \to \infty} \frac{1 + \mathbb{E}\left[x_t^\top (X^\top S^\top S X)^{-1} x_t\right]}{1 + \mathbb{E}\left[x_t^\top (X^\top X)^{-1} x_t\right]} = \frac{1-\gamma}{1-\gamma/\xi}.$$

This finishes the proof.

5.5.1 Checking the free multiplicative convolution property

Recall that the S-transform of a distribution F is defined as the solution to the equation

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

For more references, see for instance Voiculescu et al. (1992); Hiai and Petz (2006); Nica and Speicher (2006); Anderson et al. (2010).

Since $m(\frac{z+1}{zS(z)}) = -zS(z), \eta(z) = \frac{1}{z}m(-\frac{1}{z})$, we have

$$-zS(z) = m(\frac{z+1}{zS(z)}) = -\frac{zS(z)}{z+1}\eta(-\frac{zS(z)}{z+1}),$$

where $S(z)$ is the S-transform. Therefore

$$\eta_\Lambda(-\frac{zS_\Lambda(z)}{z+1}) = z+1,\ \eta_C(-\frac{zS_C(z)}{z+1}) = z+1.$$

Let $x = -\frac{z}{z+1}S_C(z)$, then $\eta_C(x) = z+1$ and (17) gives

$$\begin{aligned} z+1 = \eta_C(x) &= \eta_\Lambda(x(1+\frac{\xi-1}{\eta_C(x)})) = \eta_\Lambda(-\frac{z}{z+1}S_C(z)(1+\frac{\xi-1}{z+1})) \\ &= \eta_\Lambda(-\frac{z}{z+1}S_C(z)\frac{z+\xi}{z+1}) = \eta_\Lambda(-\frac{z}{z+1}S_\Lambda(z)). \end{aligned}$$

Therefore $S_\Lambda = \frac{z+\xi}{z+1}S_C(z)$, and equivalently $S_C(z) = S_\Lambda(z)\frac{z+1}{z+\xi}$. Let $S_0(z) = \frac{z+1}{z+\xi}$ be the S-transform of some distribution F_0, then the corresponding Stieltjes transform is $m_0(z) = \frac{\xi}{1-z} + \frac{1-\xi}{-z}$, which is the Stieltjes transform for $F_0 = \xi\delta_1 + (1-\xi)\delta_0$. This shows that F_C is a freely multiplicative convolution of F_Λ and $\xi\delta_1 + (1-\xi)\delta_0$.

5.6 Proof of Theorem 2.4

Note that B, H and D are all symmetric matrices satisfying

$$B^2 = B,\ H^2 = I_n,\ D^2 = I_n,$$

and P is also an orthogonal matrix, therefore

$$\begin{aligned} S^\top S &= P^\top DHBHDP \\ (S^\top S)^2 &= P^\top DHBHDPP^\top DHBHDP \\ &= P^\top DHBHDP = S^\top S. \end{aligned}$$

By Proposition 5.2, we only need to find

$$\mathrm{tr}[(X^\top S^\top SX)^{-1}] = \mathrm{tr}[(X^\top P^\top DHBHDPX)^{-1}], \tag{21}$$

and

$$\mathrm{tr}[(X^\top S^\top SX)^{-1}X^\top X] = \mathrm{tr}[(X^\top P^\top DHBHDPX)^{-1}X^\top X]. \tag{22}$$

We first have the following observation.

Lemma 5.9. *For a uniformly distributed permutation matrix P, diagonal matrix B with iid diagonal entries of distribution $\mu_B = \frac{r}{n}\delta_1 + (1-\frac{r}{n})\delta_0$, diagonal matrix D with iid sign random variables, equal to ± 1 with probability one half, and Hadamard matrix H, we have the following equation in distribution*

$$X^\top (P^\top DH)B(HDP)X \stackrel{d}{=} X^\top (P^\top DHDP)B(P^\top DHDP)X.$$

This is true, because we are simply permuting the diagonal matrix of iid Bernoullis in the middle term; but see the end of this section for a formal proof. We call DP the signed-permutation matrix and $W = P^\top DHDP$ the bi-signed-permutation Hadamard matrix. Thus by equations (21), (22), and Lemma 5.9,

$$\begin{aligned}
\mathbb{E}\left[\mathrm{tr}[(X^\top S^\top SX)^{-1}]\right] &= \mathbb{E}\left[\mathrm{tr}[(X^\top (P^\top DHDP)B(P^\top DHDP)X)^{-1}]\right] \\
&= \mathbb{E}\left[\mathrm{tr}[(X^\top WBWX)^{-1}]\right], \\
\mathbb{E}\left[\mathrm{tr}[(X^\top S^\top SX)^{-1}X^\top X]\right] &= \mathbb{E}\left[\mathrm{tr}[(X^\top (P^\top DHDP)B(P^\top DHDP)X)^{-1}X^\top X]\right] \\
&= \mathbb{E}\left[\mathrm{tr}[(X^\top WBWX)^{-1}X^\top X]\right].
\end{aligned}$$

Since $X^\top WBWX$ has the same nonzero eigenvalues as $BWXX^\top WB$, we first find the l.s.d. of

$$C_n = \frac{1}{n} B_n W_n X_n X_n^\top W_n B_n.$$

The following lemma states the asymptotic freeness regarding Hadamard matrix, which will be used to find the l.s.d. of C_n. For more references on free probability, see for instance Voiculescu et al. (1992); Hiai and Petz (2006); Nica and Speicher (2006); Anderson et al. (2010).

Lemma 5.10. *(Freeness of bi-signed-permutation Hadamard matrix) Let X_n, B_n, W_n be defined above, that is, X_n is an $n \times n$ deterministic matrix with uniformly bounded spectral norm and has l.s.d. μ_X, B_n is a diagonal matrix with iid diagonal entries, and W_n is a bi-signed-permutation matrix. Then*

$$\{B_n, \frac{1}{n} W_n X_n X_n^\top W_n\}$$

are asymptotically free in the limit of the non-commutative probability spaces of random matrices, as described in Section 5.1. The law of

$$C_n = \frac{1}{n} B_n W_n X_n X_n^\top W_n B_n$$

converges to the freely multiplicative convolution of μ_B and μ_X, that is, C_n has l.s.d. $\mu_C = \mu_B \boxtimes \mu_X$.

This follows directly from Corollaries 3.5, 3.7 of Anderson and Farrell (2014).

We use μ_B and μ_X to denote the elements in the limiting non-commutative probability space, their laws, and their corresponding probability distributions interchangeably. Since $\mu_B = \xi\delta_1 + (1-\xi)\delta_0$, we have $S_{\mu_B} = \frac{z+1}{z+\xi}$. From the aymptotic freeness, it follows that the S-transform of μ_C is the product of that of μ_B, μ_X, so that

$$S_{\mu_C}(z) = S_{\mu_X}(z) S_{\mu_B}(z) = S_{\mu_X}(z) \frac{z+1}{z+\xi}.$$

We will now simplify this relation. First, note that by the definition of the S-transform, we have

$$\eta_{\mu_C}(-\frac{z}{z+1}S_{\mu_C}(z)) = z+1.$$

Letting $y = -\frac{z}{z+1}S_{\mu_C}$, we have $\eta_{\mu_C}(y) = z+1$. In addition, we can simplify the original relation as

$$\begin{aligned} S_{\mu_X} &= \frac{z+\xi}{z+1}S_{\mu_C}(z) = -\frac{z+\xi}{z}y, \\ z+1 &= \eta_{\mu_X}(-\frac{z}{z+1}S_{\mu_X}(z)) = \eta_{\mu_X}(\frac{z+\xi}{z+1}y) \\ &= \eta_{\mu_X}((1+\frac{\xi-1}{z+1})y) = \eta_{\mu_X}((1+\frac{\xi-1}{\eta_{\mu_C}(y)})y) = \eta_{\mu_C}(z). \end{aligned}$$

So we have obtained

$$\eta_{\mu_X}((1+\frac{\xi-1}{\eta_{\mu_C}(y)})y) = \eta_{\mu_C}(y).$$

This is the same equation as what we obtained in (17) in the proof of Haar projection. Therefore as $n \to \infty$, we have as required

$$\lim_{n\to\infty} VE(\hat{\beta}_s, \hat{\beta}) = \frac{1-\gamma}{\xi-\gamma}.$$

Next we consider

$$\mathbb{E}\left[\mathrm{tr}[(X^\top WBWX)^{-1}X^\top X]\right].$$

Since X has the SVD $X = U\Lambda V^\top$, we have

$$\mathbb{E}\left[\mathrm{tr}[(X^\top WBWX)^{-1}X^\top X]\right] = \mathbb{E}\left[\mathrm{tr}[(U^\top WBWU)^{-1}]\right].$$

Thus we can repeat the above reasoning, except that we replace X by U. Since the result does not depend on X, we have

$$\begin{aligned} \lim_{n\to\infty} PE(\hat{\beta}_s, \hat{\beta}) &= \lim_{n\to\infty} \frac{\mathbb{E}\left[\mathrm{tr}[(X^\top S^\top SX)^{-1}X^\top X]\right]}{p} \\ &= \lim_{n\to\infty} \frac{\mathbb{E}\left[\mathrm{tr}[(U^\top WBWU)^{-1}]\right]}{\mathrm{tr}[U^\top U]} \\ &= \lim_{n\to\infty} VE(\hat{\beta}_s, \hat{\beta}) = \frac{1-\gamma}{\xi-\gamma}. \end{aligned}$$

For OE, since S satisfies $(S^\top S)^2 = S^\top S$ and the e.s.d. of $S^\top S$ converges to $\xi\delta_1 + (1-\xi)\delta_0$, the same reasoning as in Theorem 2.3 also holds in this case for Hadamard projection. This finishes the proof.

Proof. (Proof of Lemma 5.9) Note that both B and D are diagonal matrices whose diagonal entries are iid random variables, and P is a permutation matrix. Define $\tilde{B} = PBP^\top$ and $\tilde{D} = P^\top DP$, then we have

$$\tilde{B} \overset{d}{=} B, \quad \tilde{D} \overset{d}{=} D$$

and

$$DP = P\tilde{D}, \quad P^\top D = \tilde{D}P^\top. \tag{23}$$

Hence

$$\begin{aligned} X^\top P^\top DHDPBP^\top DHDPX &= X^\top P^\top DHP\tilde{D}B\tilde{D}P^\top HDPX \\ &= X^\top P^\top DHPB\tilde{D}^2P^\top HDPX \\ &= X^\top P^\top DHPBP^\top HDPX \\ &= X^\top P^\top DH\tilde{B}HDPX \\ &\overset{d}{=} X^\top P^\top DHBHDPX, \end{aligned}$$

where the first equation follows from (23), the second equation holds because $\tilde{D}$ and B are diagonal entries so they commute, while the third equation holds because $\tilde{D}^2 = I_n$. □

5.7 Proof of Theorem 2.5

We can take

$$S = \begin{pmatrix} s_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & s_n \end{pmatrix},$$

which is an $n \times n$ diagonal matrix and s_i-s are iid random variables with $\mathbb{P}[s_i = 1] = \frac{r}{n}$ and $\mathbb{P}[s_i = 0] = 1 - \frac{r}{n}$. Since $s_i^2 = s_i$, we have $S^2 = S$, hence

$$\begin{aligned} VE(\hat{\beta}_s, \hat{\beta}) &= \frac{\mathbb{E}\left[\mathrm{tr}[(X^\top SX)^{-1}]\right]}{\mathrm{tr}[(X^\top X)^{-1}]} \\ PE(\hat{\beta}_s, \hat{\beta}) &= \frac{\mathbb{E}\left[\mathrm{tr}[(X^\top SX)^{-1}X^\top X]\right]}{p}. \end{aligned}$$

Since X is unitarily invariant and S is a diagonal matrix independent from X, $\{S, X, X^\top\}$ are almost surely asymptotically free in the non-commutative probability space by Theorem 4.3.11 of Hiai and Petz (2006). Since the law of S converges to $\mu_S = \xi\delta_1 + (1-\xi)\delta_0$, the law of X converges to μ_X, thus the law of $SXX^\top S$ converges to the freely multiplicative convolution $\mu_S \boxtimes \mu_X$. The rest of the proof is the same as that in the proof of Theorem 2.4.

5.8 Proof of Theorem 2.6

Define

$$S = \begin{pmatrix} s_1 & & \\ & \ddots & \\ & & s_n \end{pmatrix}, \quad W = \begin{pmatrix} w_1 & & \\ & \ddots & \\ & & w_n \end{pmatrix},$$

where the s_i-s are independent and $s_i|\pi_i \sim Bernoulli(\pi_i)$. S is independent of Z because π_i is independent of z_i, by the assumption. W has l.s.d. F_w. According to Proposition 5.1, the values of

VE, PE are determined by $\mathrm{tr}[(X^\top X)^{-1}]$, $\mathrm{tr}[Q_1(S,X)] = \mathrm{tr}[(X^\top SX)^{-1}]$, and $\mathrm{tr}[Q_2(S,X)]$. Note that under the elliptical model $X = WZ\Sigma^{1/2}$, we have

$$\begin{aligned}\mathrm{tr}[(X^\top X)^{-1}] &= \mathrm{tr}[(\Sigma^{1/2}Z^\top W^2 Z\Sigma^{1/2})^{-1}],\\ \mathrm{tr}[Q_1(S,X)] &= \mathrm{tr}[(\Sigma^{1/2}Z^\top WSWZ\Sigma^{1/2})^{-1}],\\ \mathrm{tr}[Q_2(S,X)] &= \mathrm{tr}[(Z^\top WSWZ)^{-1}Z^\top W^2 Z].\end{aligned}$$

Note that the e.s.d. of Σ converges in distribution to some probability distribution F_Σ, and the e.s.d. of WSW converges in distribution to F_{sw^2}, the limiting distribution of $s_i w_i^2$, $i = 1,\dots,n$. Again from the results of Zhang (2007) or Paul and Silverstein (2009), with probability 1, the e.s.d. of $C_n = \frac{1}{n}\Sigma^{1/2}Z^\top WSWZ\Sigma^{1/2}$ converges to a probability distribution function F_C, whose Stieltjes transform $m_C(z)$, for $z \in \mathbb{C}^+$ is given by

$$m_C(z) = \int \frac{1}{t\int \frac{u}{1+\gamma e_C u}dF_{sw^2}(u) - z}dF_\Sigma(t),$$

where $e_C = e_C(z)$ is the unique solution in $\mathbb{C}^+$ of the equation

$$e_C = \int \frac{t}{t\int \frac{u}{1+\gamma e_C u}dF_{sw^2}(u) - z}dF_\Sigma(t).$$

Similarly, the e.s.d. of $D_n = \frac{1}{n}\Sigma^{1/2}Z^\top W^2 Z\Sigma^{1/2}$ converges to a probability distribution F_D, whose Stieltjes transform $m_D(s)$, for $z \in \mathbb{C}^+$ is given by

$$m_D(z) = \int \frac{1}{t\int \frac{u}{1+\gamma e_D u}dF_{w^2}(u) - z}dF_\Sigma(t),$$

where $e_D = e_D(z)$ is the unique solution in $\mathbb{C}^+$ of the equation

$$e_D = \int \frac{t}{t\int \frac{u}{1+\gamma e_D u}dF_{w^2}(u) - z}dF_\Sigma(t).$$

Since F_C and F_D have no point mass at the origin, we can set $z = 0$. See Dobriban and Sheng (2018) for a detailed argument building on the results of Couillet and Hachem (2014). Therefore

$$m_C(0) = \frac{1}{\int \frac{u}{1+\gamma e_C(0) u}dF_{sw^2}(u)}\int \frac{1}{t}dF_\Sigma(t), \quad e_C(0) = \frac{1}{\int \frac{u}{1+\gamma e_C(0) u}dF_{sw^2}(u)}.$$

Note also that

$$e_C(0) = \frac{\gamma e_C(0)}{\int \frac{\gamma e_C(0) u}{1+\gamma e_C(0) u}dF_{sw^2}(u)} = \frac{\gamma e_C(0)}{1-\eta_{sw^2}(\gamma e_C(0))},$$

thus $\eta_{sw^2}(\gamma e_C(0)) = 1-\gamma$, and

$$m_C(0) = e_C(0)\int \frac{1}{t}dF_\Sigma(t) = \frac{\eta_{sw^2}^{-1}(1-\gamma)}{\gamma}\int \frac{1}{t}dF_\Sigma(t). \tag{24}$$

Similarly,

$$m_D(0) = e_D(0) \int \frac{1}{t} dF_\Sigma(t) = \frac{\eta_{w^2}^{-1}(1-\gamma)}{\gamma} \int \frac{1}{t} dF_\Sigma(t).$$

Hence, again by the same argument as we have seen several times before, the traces have limits that can be evaluated in terms of Stieltjes transforms, and we have

$$\begin{aligned} VE(\hat{\beta}_s, \hat{\beta}) &= \frac{\mathrm{tr}[Q_1(S,X)]}{\mathrm{tr}[(X^\top X)^{-1}]} = \frac{\mathrm{tr}[(\Sigma^{1/2} Z^\top W S W Z \Sigma^{1/2})^{-1}]}{\mathrm{tr}[(\Sigma^{1/2} Z^\top W^2 Z \Sigma^{1/2})^{-1}]} \\ &\to \frac{m_C(0)}{m_D(0)} = \frac{\eta_{sw^2}^{-1}(1-\gamma)}{\eta_{w^2}^{-1}(1-\gamma)}, \end{aligned}$$

and the result for VE follows.

We then deal with PE. Note that

$$PE(\hat{\beta}_s, \hat{\beta}) = \frac{\mathbb{E}\left[\mathrm{tr}[(Z^\top W S W Z)^{-1} Z^\top W^2 Z]\right]}{p}.$$

We first assume that Z has iid $\mathcal{N}(0,1)$ entries. Denote $T_1 = WSW$, $T_2 = W(I-S)W$. Since S is a diagonal matrix whose diagonal entries are 1 or 0, W is also a diagonal matrix, T_1 and T_2 are both diagonal matrices and the set of their nonzero entries is complementary. So $Z^\top T_1 Z$ and $Z^\top T_2 Z$ are independent from each other and $T_1 + T_2 = W^2$. We have

$$\begin{aligned} \mathbb{E}\left[\mathrm{tr}[(Z^\top W S W Z)^{-1} Z^\top W^2 Z]\right] &= \mathbb{E}\left[\mathrm{tr}[(Z^\top T_1 Z)^{-1} Z^\top (T_1 + T_2) Z]\right] \\ &= \mathbb{E}\left[\mathrm{tr}[I_p + (Z^\top T_1 Z)^{-1} Z^\top T_2 Z]\right] \\ &= p + \mathrm{tr}[\mathbb{E}\left[(Z^\top T_1 Z)^{-1}\right] \mathbb{E}\left[Z^\top T_2 Z\right]]. \end{aligned}$$

Note that

$$\mathbb{E}\left[(Z^\top T_2 Z)_{ij}\right] = \sum_{k=1}^n \mathbb{E}\left[z_{ki} T_{2,kk} z_{kj}\right] = \sum_{k=1}^n T_{2,kk} \delta_{ij}$$

thus

$$\begin{aligned} \mathbb{E}\left[Z^\top T_2 Z\right] &= \mathbb{E}\left[\mathrm{tr}(T_2)\right] I_p, \\ \mathbb{E}\left[\mathrm{tr}[(Z^\top W S W Z)^{-1} Z^\top W^2 Z]\right] &= p + \mathbb{E}\left[\mathrm{tr}(T_2)\right] \mathrm{tr}[\mathbb{E}\left[(Z^\top T_1 Z)^{-1}\right]]. \end{aligned}$$

Note that $\frac{1}{n} Z^\top W S W Z$ is equal to C_n with Σ replaced by the identity. Thus by (24),

$$\begin{aligned} \frac{1}{p} \mathrm{tr}[(\frac{1}{n} Z^\top W S W Z)^{-1}]] &\xrightarrow{a.s.} \frac{\eta_{sw^2}^{-1}(1-\gamma)}{\gamma}, \\ \mathrm{tr}[(Z^\top W S W Z)^{-1}]] &\xrightarrow{a.s.} \eta_{sw^2}^{-1}(1-\gamma), \end{aligned}$$

thus

$$\begin{aligned} \lim_{n\to\infty} PE(\hat{\beta}_s, \hat{\beta}) &= 1 + \frac{1}{p} \mathrm{tr}(T_2) \eta_{sw^2}^{-1}(1-\gamma) \\ &= 1 + \frac{1}{\gamma} \mathbb{E}\left[w^2(1-s)\right] \eta_{sw^2}^{-1}(1-\gamma) \end{aligned}$$

Then we use a similar Lindeberg swapping argument as in Theorem 2.3 to show extend this to Z with iid entries of zero mean, unit variance and finite fourth moment. This finishes the proof for PE. For the last claim, for OE, note that

$$\begin{aligned}\mathbb{E}\left[x_t^\top (X^\top X) x_t\right] &= \mathbb{E}\left[w^2\right] \mathbb{E}\left[z_t^\top (Z^\top W^2 Z)^{-1} z_t\right] \\ &= \mathbb{E}\left[w^2\right] \mathbb{E}\left[\operatorname{tr}[(Z^\top W^2 Z)^{-1}]\right] \\ &\to \mathbb{E}\left[w^2\right] \eta_{w^2}^{-1}(1-\gamma),\end{aligned}$$

and that

$$\begin{aligned}\mathbb{E}\left[x_t^\top (X^\top S^\top S X) x_t\right] &= \mathbb{E}\left[w^2\right] \mathbb{E}\left[z_t^\top (Z^\top W S W Z)^{-1} z_t\right] \\ &= \mathbb{E}\left[w^2\right] \mathbb{E}\left[\operatorname{tr}[(Z^\top W S W Z)^{-1}]\right] \\ &\to \mathbb{E}\left[w^2\right] \eta_{sw^2}^{-1}(1-\gamma).\end{aligned}$$

Thus

$$\lim_{n\to\infty} OE(\hat{\beta}_s, \hat{\beta}) = \lim_{n\to\infty} \frac{1 + \mathbb{E}\left[x_t^\top (X^\top S^\top S X)^{-1} x_t\right]}{1 + \mathbb{E}\left[x_t^\top (X^\top X)^{-1} x_t\right]} = \frac{1 + \mathbb{E}\left[w^2\right] \eta_{sw^2}^{-1}(1-\gamma)}{1 + \mathbb{E}\left[w^2\right] \eta_{w^2}^{-1}(1-\gamma)},$$

This finishes the proof.

5.8.1 Proof of Corollary 2.7

It suffices to show that leverage score sampling that samples the i-th row with probability $\min(\frac{r}{p} h_{ii}, 1)$ is equivalent to sample with probability $\min\left[\frac{r}{p}\left(1 - \frac{1}{1+w^2\eta_{w^2}^{-1}(1-\gamma)}\right), 1\right]$. Given that the latter probability is independent from z_i, the statement of the corollary will then follow directly from Theorem 2.6.

To see this equivalence, first note that

$$\begin{aligned}h_{ii} &= x_i^\top (\sum_{j\neq i} x_j x_j^\top + x_i x_i^\top)^{-1} x_i = x_i^\top (\sum_{j\neq i} x_j x_j^\top)^{-1} x_i - \frac{(x_i^\top (\sum_{j\neq i} x_j x_j^\top)^{-1} x_i)^2}{1 + x_i^\top (\sum_{j\neq i} x_j x_j^\top)^{-1} x_i} \\ &= \frac{x_i^\top (\sum_{j\neq i} x_j x_j^\top)^{-1} x_i}{1 + x_i^\top (\sum_{j\neq i} x_j x_j^\top)^{-1} x_i}\end{aligned}$$

and

$$\begin{aligned}\frac{1}{1-h_{ii}} &= 1 + x_i^\top (\sum_{j\neq i} x_j x_j^\top)^{-1} x_i = 1 + w_i^2 z_i^\top \Sigma^{1/2} (\sum_{j\neq i} x_j x_j^\top)^{-1} \Sigma^{1/2} z_i \\ &= 1 + w_i^2 z_i^\top (\sum_{j\neq i} w_j^2 z_j z_j^\top)^{-1} z_i.\end{aligned}$$

Denote $R = \sum_{j=1}^n w_j^2 z_j z_j^\top$, $R_{(i)} = \sum_{j\neq i} w_j^2 z_j z_j^\top$, so that $\frac{1}{1-h_{ii}} = 1 + w_i^2 z_i^\top R_{(i)}^{-1} z_i$.

Since z_i and $R_{(i)}$ are independent for each $i = 1, \ldots, n$, while z_i has iid entries of zero mean and unit variance and bounded moments of sufficienctly high order, then by the concentration of quadratic forms lemma 5.11 cited below, we have

$$\frac{1}{n} z_i^\top R_{(i)}^{-1} z_i - \frac{1}{n} \operatorname{tr}(R_{(i)}^{-1}) \xrightarrow{a.s.} 0.$$

Lemma 5.11 (Concentration of quadratic forms, consequence of Lemma B.26 in Bai and Silverstein (2010)). *Let $x \in \mathbb{R}^p$ be a random vector with iid entries and $\mathbb{E}[x] = 0$, for which $\mathbb{E}\left[(\sqrt{p}x_i)^2\right] = \sigma^2$ and $\sup_i \mathbb{E}\left[(\sqrt{p}x_i)^{4+\eta}\right] < C$ for some $\eta > 0$ and $C < \infty$. Moreover, let A_p be a sequence of random $p \times p$ symmetric matrices independent of x, with uniformly bounded eigenvalues. Then the quadratic forms $x^\top A_p x$ concentrate around their means at the following rate*

$$P(|x^\top A_p x - p^{-1}\sigma^2 \operatorname{tr} A_p|^{2+\eta/2} > C) \leqslant Cp^{-(1+\eta/4)}.$$

To use lemma 5.11, we only need to guarantee that the smallest eigenvalue of $R_{(i)}$ is uniformly bounded below. For this, it is enough that the smallest eigenvalue of R is uniformly bounded below. Since w_i are bounded away from zero, this property follows from the corresponding one for the sample covariance matrix of z_i, which is just the well-known Bai-Yin law (Bai and Silverstein, 2010).

Continuing with our argument, by the standard rank-one-perturbation argument (Bai and Silverstein, 2010), we have $\lim_{n\to\infty} \frac{1}{n}\operatorname{tr}[R_{(i)}^{-1}] - \frac{1}{n}\operatorname{tr}[R^{-1}] = 0$, since $R_{(i)}$ is a rank-one perturbation of R. Recall that Z has iid entries satisfying $\mathbb{E}[Z_{ij}] = 0$,$\mathbb{E}\left[Z_{ij}^2\right] = 1$. Moreover, it is easy to see that by the $4+\eta$-th moment assumption we have for each $\delta > 0$ that

$$\frac{1}{\delta^2 np}\sum_{i,j}\mathbb{E}\left[Z_{ij}^2 I_{[|Z_{ij}|>\delta\sqrt{n}]}\right] \to 0, \text{ as } n \to \infty.$$

Also, the e.s.d. of W^2 converges weakly to the distribution of w^2. By the results of Zhang (2007) or Paul and Silverstein (2009), with probability 1, the e.s.d. of $B_n = n^{-1}Z^\top W^2 Z$ converges in distribution to a probability distribution F_B whose Stieltjes transform satisfies

$$m_B(z) = \frac{1}{\int \frac{s}{1+\gamma e_B s} dF_{w^2}(s) - z},$$

where for $z \in \mathbb{C}^+$, $e_B = e_B(z)$ is the unique solution in $\mathbb{C}^+$ to the equation

$$e_B = \frac{1}{\int \frac{s}{1+\gamma e_B s} dF_{w^2}(s) - z}.$$

Also, by the same reasoning as in the proof of the Haar matrix case, the l.s.d. is supported on an interval bounded away from zero. This means that we can find the almost sure limits of the traces in terms of the Stieltjes transform of the l.s.d. at zero, or equivalently in terms of the inverse eta-transform:

$$\frac{1}{p}\operatorname{tr}(\frac{1}{n}R^{-1}) = \frac{n}{p}\operatorname{tr}[(Z_w^\top Z_w)^{-1}] \to \frac{\eta_{w^2}^{-1}(1-\gamma)}{\gamma},$$

and therefore $\operatorname{tr}(R^{-1}) \xrightarrow{a.s.} \eta_{w^2}^{-1}(1-\gamma)$. Thus, from the expression of h_{ii} given at the beginning, we also have

$$|h_{ii} - 1 + \frac{1}{1 + w_i^2 \eta_{w^2}^{-1}(1-\gamma)}| \xrightarrow{a.s.} 0.$$

Thus as n goes to infinity, leverage-based sampling is equivalent to sampling x_i with probability

$$\pi_i = \min\left(\frac{r}{p}(1 - \frac{1}{1 + w_i^2 \eta_{w^2}^{-1}(1-\gamma)}), 1\right), \tag{25}$$

in the sense that $|\min(\frac{r}{p}h_{ii}, 1) - \pi_i| \xrightarrow{a.s.} 0$. Therefore, it is not hard to see that the performance metrics we study have the same limits for leverage sampling and for sampling with respect to π_i. We argue for this in more detail below. Let S^* be the sampling matrix based on the leverage scores, with diagonal entries $s_i^* \sim Bernoulli(\min(r/nh_{ii}, 1))$. This is the original sampling mechanism to which the theorem refers. Now, we have shown that $\|S - S^*\|_{op} \to 0$ almost surely. Because of this, one can check that $\mathrm{tr}[Q_1(S, X)] - \mathrm{tr}[Q_1(S^*, X)] \to 0$ almost surely. This follows by a simple matrix calculation expressing $A^{-1} - B^{-1} = -A^{-1}(B - A)B^{-1}$, and bounding the trace using Lemma 5.7.

5.9 Details of example for leverage sampling, Section 2.5.2

We assume that $\mathbb{P}[w_i = \pm d_1] = \mathbb{P}[w_i = \pm d_2] = \frac{1}{4}$. Then

$$\eta_{w^2}(z) = \frac{1}{2}\frac{1}{1 + zd_1^2} + \frac{1}{2}\frac{1}{1 + zd_2^2},$$
$$\eta_{sw^2}(z) = (1 - \frac{1}{2}\min(\pi_1, 1) - \frac{1}{2}\min(\pi_2, 1)) + \frac{1}{2}\min(\pi_1, 1)\frac{1}{1 + d_1^2 z} + \frac{1}{2}\min(\pi_2, 1)\frac{1}{1 + d_2^2 z},$$

where

$$\pi_1 = \frac{\xi}{\gamma}(1 - \frac{1}{1 + d_1^2 \eta_{w^2}^{-1}(1-\gamma)}),\ \pi_2 = \frac{\xi}{\gamma}(1 - \frac{1}{1 + d_2^2 \eta_{w^2}^{-1}(1-\gamma)}).$$

It is easy to see that

$$\pi_1 + \pi_2 = 2\xi,$$

and

$$\eta_{F_{w^2}}^{-1}(1-\gamma) = \frac{1}{2d_1^2 d_2^2}(-d_1^2 - d_2^2 + \frac{d_1^2 + d_2^2}{2(1-\gamma)} + \sqrt{(d_1^2 + d_2^2 - \frac{d_1^2 + d_2^2}{2(1-\gamma)}) + \frac{4d_1^2 d_2^2 \gamma}{1-\gamma}}),$$

If we use the r rows of X with the largest leverage scores, the truncated distribution $\tilde{w}$ in Theorem 2.8 can be written as

$$F_{\tilde{w}^2}(t) = \begin{cases} \delta_{d_2^2}, & 0 < \frac{r}{n} \le \frac{1}{2} \\ (1 - \frac{n}{2r})\delta_{d_1^2} + \frac{n}{2r}\delta_{d_2^2}, & \frac{1}{2} < \frac{r}{n} \le 1. \end{cases}$$

Therefore

$$\eta_{\tilde{w}^2}(z) = \begin{cases} \frac{1}{1+d_2^2 z}, & 0 < \frac{r}{n} \le \frac{1}{2} \\ (1 - \frac{n}{2r})\frac{1}{1+d_1^2 z} + \frac{n}{2r}\frac{1}{1+d_2^2 z}, & \frac{1}{2} < \frac{r}{n} \le 1, \end{cases}$$

thus

$$\eta^{-1}_{F_{\tilde{w}^2}}(1-\frac{\gamma}{\xi}) = \begin{cases} \frac{\gamma}{d_2^2(\xi-\gamma)}, & 0 < \frac{r}{n} \leq \frac{1}{2} \\ \frac{1}{2d_1^2 d_2^2}[-b + \sqrt{b^2 + \frac{4d_1^2 d_2^2 \gamma}{\xi-\gamma}}], & \frac{1}{2} < \frac{r}{n} \leq 1. \end{cases}$$

Here

$$b = d_1^2 + d_2^2 - \frac{(2\xi-1)d_2^2 + d_1^2}{2(\xi-\gamma)}.$$

We use bisection method to find $\eta^{-1}_{sw^2}(1-\gamma)$. The Python codes for the implementation and Figure 2 can be found at `https://github.com/liusf15/Sketching-lr`.

5.10 Additional numerical results

5.10.1 Comparison with previous bounds

We also compare our results with the upper bounds given in Raskutti and Mahoney (2016). For sub-Gaussian projections, they showed that if $r \geq c \log n$, then with probability greater than 0.7, it holds that

$$PE \leq 44(1+\frac{n}{r}),\ RE \leq 1 + 44\frac{p}{r}.$$

For Hadamard projection, they showed that if $r \geq cp \log n(\log p + \log\log n)$, then with probability greater than 0.8, it holds that

$$PE \leq 1 + 40\log(np)(1+\frac{p}{r}),\ RE \leq 40\log(np)(1+\frac{n}{r}).$$

In Figure 11, we plot both our theoretical lines and the above upper bounds, as well as the simulation results. It is shown that our theory is much more accurate than these upper bounds.

5.10.2 Computation time

In this section we perform a more rigorous empirical comparison of the running time of sketching. We know that the running time of OLS has order of magnitude $O(np^2)$, while the running time of Hadamard projections is $c_1 np^2 + c_2 np\log(n)$ for some constants c_i. While the cubic term clearly dominates for large n, p, our goal is to understand the performance for finite samples n, p on typical commodity hardware. For this reason, we perform careful timing experiments to determine the approximate values of the constants on a MacBook Pro (2.5 GHz CPU, Intel Core i7).

We obtain the following results. The time for full OLS and Hadamard sketching is approximately

$$t_{full} = 4 \times 10^{-11} np^2,\ t_{Hadamard} = 2 \times 10^{-8} pn\log n + 4 \times 10^{-11} rp^2$$

See Figure 12 for a comparison of the running times for various combinations of n, p. For instance, we show the results for $n = 7 \cdot 10^4$, and $p = 1.4 \cdot 10^4$ with the sampling ratio ranging from 0.2 to 1. We see that we save time if we take $r/n \leqslant 0.6$.

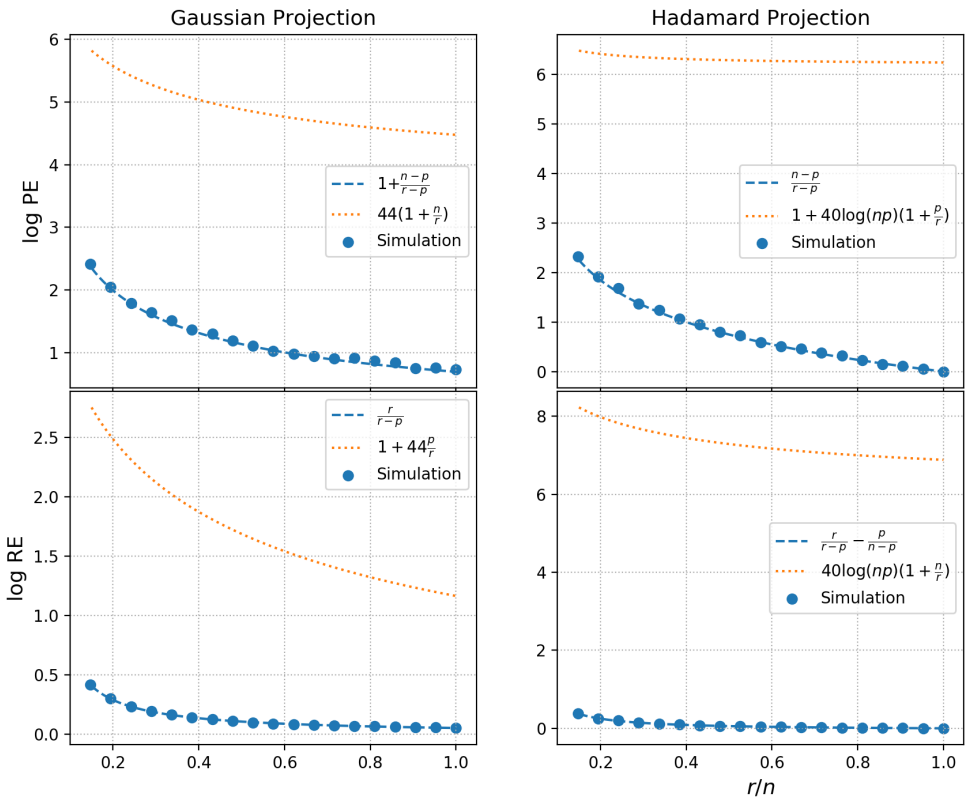

Figure 11: Comparison with prior bounds. In this simulation, we let $n = 2000$, the aspect ratio $\gamma = 0.05$, with r/n ranging from 0.15 to 1. The first column displays the results for PE and RE for Gaussian projection, while the second column shows results for randomized Hadamard projection. The y-axis is on the log scale. The data matrix X is generated from Gaussian distribution and fixed at the beginning, while the coefficient β is generated from uniform distribution and also fixed. At each dimension r we repeat the simulation 50 times and all the relative efficiencies are averaged over 50 simulations. In each simulation, we generate the noise ε as well as the sketching matrix S. The orange dotted lines are drawn according to Section 5.10.1, while the blue dashed lines are drawn according to our Theorem 2.1 and Theorem 2.4.

We can also perform a more quantitative analysis. If we want to reduce the time by a factor of $0 < c < 1$, then we need

$$\frac{2 \times 10^{-8} pn \log n + 4 \times 10^{-11} rp^2}{4 \times 10^{-11} np^2} \leq c$$

or also $r \leq cn - 500\frac{n \log n}{p}$, when $0 < c - 500\frac{\log n}{p} < 1$. Then the out-of-sample prediction efficiency is lower bounded by

$$\begin{aligned} OE(\hat{\beta}_s, \hat{\beta}) &= \frac{r(n-p)}{n(r-p)} \\ \geq &\frac{n-p}{n}\left(1 + \frac{p}{n(c - \frac{500 \log n}{p}) - p}\right) = (1-\gamma)\left(1 + \frac{\gamma}{c - \frac{500 \log n}{p} - \gamma}\right). \end{aligned}$$

This shows how much we lose if we decrease the time by a factor of c.

Similarly, if we want to control the VE, say to ensure that $VE(\hat{\beta}_s, \hat{\beta}) \leq 1 + \delta$, then we need

$$r \geq \frac{n-p}{1+\delta} + p,$$

then the we must spend at least a fraction of the full OLS time given below

$$\frac{r}{n} + \frac{500 \log n}{p} \geq \frac{1-\gamma}{1+\delta} + \gamma + \frac{500 \log n}{p}.$$

References

D. Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281. ACM, 2001.

D. Ahfock, W. J. Astle, and S. Richardson. Statistical properties of sketching algorithms. *arXiv preprint arXiv:1706.03665*, 2017.

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*, pages 557–563. ACM, 2006.

G. W. Anderson and B. Farrell. Asymptotically liberating sequences of random unitary matrices. *Advances in Mathematics*, 255:381–413, 2014.

G. W. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*. Number 118. Cambridge University Press, 2010.

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

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

T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

E. Bingham and H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250. ACM, 2001.

T. I. Cannings and R. J. Samworth. Random-projection ensemble classification. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):959–1035, 2017.

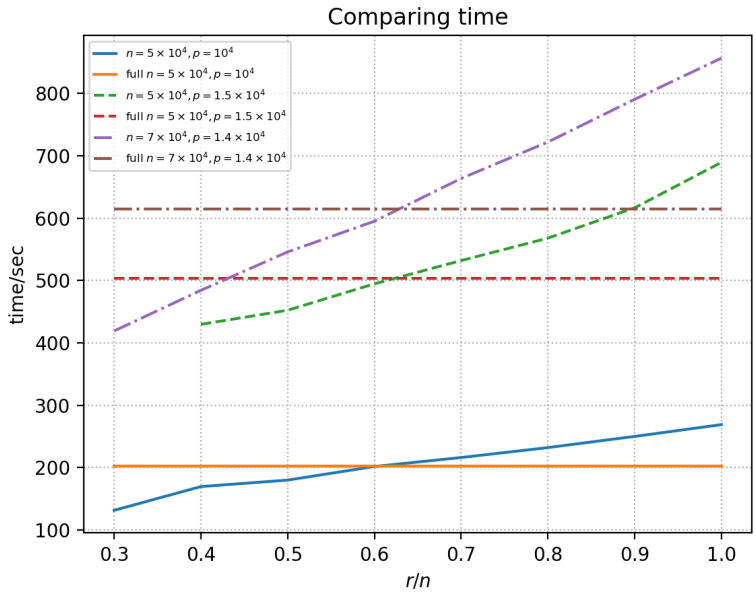

Figure 12: A comparison of the running times for various combinations of n, p.

S. Chatterjee. A generalization of the lindeberg principle. *The Annals of Probability*, 34(6):2061–2076, 2006.

S. Chen, R. Varma, A. Singh, and J. Kovačcević. A statistical perspective of sampling scores for linear regression. In *Information Theory (ISIT), 2016 IEEE International Symposium on*, pages 1556–1560. IEEE, 2016.

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

R. Couillet and W. Hachem. Analysis of the limiting spectral measure of large random matrices of the separable covariance type. *Random Matrices: Theory and Applications*, 3(04):1450016, 2014.

P. Dhillon, Y. Lu, D. P. Foster, and L. Ungar. New subsampling algorithms for fast least squares regression. In *Advances in neural information processing systems*, pages 360–368, 2013.

E. Dobriban. Efficient computation of limit spectra of sample covariance matrices. *Random Matrices: Theory and Applications*, 04(04):1550019, 2015.

E. Dobriban. Permutation methods for factor analysis and PCA. *arXiv preprint arXiv:1710.00479*, 2017a.

E. Dobriban. Sharp detection in PCA under correlations: all eigenvalues matter. *The Annals of Statistics*, 45(4):1810–1833, 2017b.

E. Dobriban and A. B. Owen. Deterministic parallel analysis: An improved method for selecting the number of factors and principal components. *arXiv preprint arXiv:1711.04155*, 2017.

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

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

E. Dobriban, W. Leeb, and A. Singer. Optimal prediction in the linearly transformed spiked model. *arXiv preprint arXiv:1709.03393*, 2017.

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

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

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

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

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

M. M. Fard, Y. Grinberg, J. Pineau, and D. Precup. Compressed least-squares regression on sparse spaces. In *AAAI*, 2012.

X. Z. Fern and C. E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 186–193, 2003.

K. J. Galinsky, G. Bhatia, P.-R. Loh, S. Georgiev, S. Mukherjee, N. J. Patterson, and A. L. Price. Fast principal-component analysis reveals convergent evolution of adh1b in europe and east asia. *The American Journal of Human Genetics*, 98(3):456–472, 2016.

W. Hachem. An expression for $\int log(t/\sigma^2 + 1)\mu \boxtimes \tilde{\mu}(dt)$. *unpublished*, 2008.

N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

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

S. Jassim, H. Al-Assam, and H. Sellahewa. Improving performance and security of biometrics using efficient and stable random projection techniques. In *Image and Signal Processing and Analysis, 2009. ISPA 2009. Proceedings of 6th International Symposium on*, pages 556–561. IEEE, 2009.

A. Kabán. New bounds on compressive linear least squares regression. In *Artificial Intelligence and Statistics*, pages 448–456, 2014.

R. Kannan. Foundations of data science, Simons Foundation, 2018. URL `https://www.youtube.com/watch?v=9GMT3FnQTGM&t=1600s`.

K. Liu, H. Kargupta, and J. Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on knowledge and Data Engineering*, 18(1): 92–106, 2006.

L. T. Liu, E. Dobriban, and A. Singer. *e* pca: High dimensional exponential family PCA. *arXiv preprint arXiv:1611.05550, to appear in the Annals of Applied Statistics*, 2016.

M. Lopes, L. Jacob, and M. J. Wainwright. A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, pages 1206–1214, 2011.

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

Y. Lu, P. Dhillon, D. P. Foster, and L. Ungar. Faster ridge regression via the subsampled randomized hadamard transform. In *Advances in neural information processing systems*, pages 369–377, 2013.

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. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

O. Maillard and R. Munos. Compressed least-squares regression. In *Advances in Neural Information Processing Systems*, pages 1213–1221, 2009.

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

K. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Academic Press, 1979.

S. Ng. Opportunities and challenges: Lessons from analyzing terabytes of scanner data. Technical report, National Bureau of Economic Research, 2017.

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

S. Oymak and J. A. Tropp. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 2017.

D. Papailiopoulos, A. Kyrillidis, and C. Boutsidis. Provable deterministic leverage score sampling. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 997–1006. ACM, 2014.

D. Paul and A. Aue. Random matrix theory in statistics: A review. *Journal of Statistical Planning and Inference*, 150:1–29, 2014.

D. Paul and J. W. Silverstein. No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix. *Journal of Multivariate Analysis*, 100(1):37–57, 2009.

M. Pilanci and M. J. Wainwright. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115, 2015.

M. Pilanci and M. J. Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *The Journal of Machine Learning Research*, 17(1):1842–1879, 2016.

M. Pilanci and M. J. Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017.

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

M. J. Schneider and S. Gupta. Forecasting sales of new and existing products using consumer reviews: A random projections approach. *International Journal of Forecasting*, 32(2):243–256, 2016.

R. Srivastava, P. Li, and D. Ruppert. Raptt: An exact two-sample test in high dimensions using random projections. *Journal of Computational and Graphical Statistics*, 25(3):954–970, 2016.

G.-A. Thanei, C. Heinze, and N. Meinshausen. Random projections for large-scale regression. In *Big and complex data analysis*, pages 51–68. Springer, 2017.

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

D. V. Voiculescu, K. J. Dykema, and A. Nica. *Free random variables*. Number 1. American Mathematical Soc., 1992.

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

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

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

Y. Yang, M. Pilanci, and M. J. Wainwright. Randomized sketches for kernels: Fast and optimal nonparametric regression. *The Annals of Statistics*, 45(3):991–1023, 2017.

J. Yao, Z. Bai, and S. Zheng. *Large Sample Covariance Matrices and High-Dimensional Data Analysis*. Cambridge University Press, New York, 2015.

L. Zhang. *Spectral analysis of large dimentional random matrices*. PhD thesis, 2007.

S. Zhou, L. Wasserman, and J. D. Lafferty. Compressed regression. In *Advances in Neural Information Processing Systems*, pages 1713–1720, 2008.