The New England Journal of Statistics in Data Science Volume 26, 1–15 (2025)
DOI: https://doi.org/1339

The merits of Randomized Hadamard Transform in distributed regression via sampled machines

Zhixiang ZHANG* and Yishu YANG[1]

Dedicated to all researchers in the field.

Abstract

Distributed linear regression is widely used to process large datasets by splitting them across multiple machines and combining local estimates. This paper explores the relative efficiency of the Ordinary Least Squares (OLS) estimator in such settings, defined as the ratio of the mean squared error of the global OLS estimator to that of the distributed OLS estimator obtained by averaging local estimates. We identify a Gaussian Mixture Model (GMM) distribution for the data matrix, where rows are drawn from two multivariate Gaussian distributions with distinct means and covariances, causing uniform sampling to yield poor relative efficiency due to uneven variance across partitions. To address this, we propose applying the Randomized Hadamard Transform (RHT) before uniform sampling. Our key finding is that, with RHT and equal weighting of local estimators (each weighted $\frac{1}{K}$ for K machines), the relative efficiency approaches 1 as the number of rows n grows to infinity, with the number of features p fixed but potentially large. This occurs because RHT stabilizes local Gram matrices, reducing the difference between $\mathrm{tr}[(\tilde{\mathbf{X}}_i^T\tilde{\mathbf{X}}_i)^{-1}]$ and $\mathrm{tr}[(\frac{1}{K}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}]$ to zero at a rate of $\mathcal{O}(p\sqrt{\frac{\log n}{n^3}})$. We prove this using concentration inequalities, including the Bernstein Inequality and Matrix Chernoff Inequality, ensuring statistical rigor. This approach enhances computational efficiency and robustness in distributed regression, especially for heterogeneous data.

1. INTRODUCTION

In the era of big scale data analysis, distributed sketching in machine learning has been a common tool for efficiently training large datasets for most linear regression tasks. This means we would partition our full datasets into K blocks and then do the regression on each block separately, which is followed by an averaged process of all the learned local parameters to get an overall parameter. This has been studied much by researchers in the field of distributed machine learning, however, many researchers focus on the optimized absolute sketching errors or proportions of biases in distributed sketching (Wang, 2018[9]) or focus on the accurate expectation of approximation error for the OLS (ordinary least square) estimation (Derezinski, 2023[2]). Not much work has been done to compare the relative efficiency of different sampling methods in distributed sketching, like Subsampled Randomized Hadamard Transform Sampling (SRHT) or Leverage-based Sampling (LBS). However it is good to see that Dobriban and Liu[3] have discussed much results of distributed sketching in OLS estimation via all kinds of efficiencies, which I think would be a further study regarding our research here.

In this article, our main research contribution is to identify the specific distribution of the data matrix $\mathbf{X}$ where the relative efficiency of OLS estimation, i.e. the error ratio of global OLS estimator to distributed OLS estimator by averaging local OLS estimators, would decrease in a drastic way when the sampling method is just the equally partitioned uniform sampling from the $Section 3.2$ Finite sample results of the paper by Dobriban and Sheng[4]. And we contribute to show that the Randomized Hadamard Transform (RHT) here could improve the relative efficiency of OLS estimation by flattening the variance of the local gram matrix for each machine with equal weight for averaging, thus helping us save the time for adjusting th weight accordingly from the result of 2.6. Finally, our most important contribution is to rigorouly proving that the relative efficiency of OLS estimation with equal-weighted partitions after Randomized Hadamard Transform (RHT) could achieve the dream efficiency of 1 when the number of rows n tends to infinity and the number of columns p is fixed and could be as large as any large number. We could also intepretate this idea differently, that is the difference of $\mathrm{tr}[(\tilde{\mathbf{X}}_i^\top\tilde{\mathbf{X}}_i)^{-1}]$ and $\mathrm{tr}[(\frac{1}{K}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}})^{-1}]$ converges to 0 at the rate of $\mathcal{O}(p\sqrt{\frac{\log n}{n^3}})$4.5 as n tends to be large for fixed p being as large as possible, meaning that the local gram matrix after Randomized Hadamard Transform (RHT) is well stablized.

In short, we skip the dirty work of adjusting the weight for each local OLS estimator and just take the average of all local OLS estimators with equal weight $\frac{1}{K}$ for K machines by utilizing the Randomized Hadamard Transform (RHT) to flatten the variance of the local gram matrix for each machine.

Our main intuition of randomized hadamard transfrom RHT is from the paper of Tropp[6] and the paper of Cherapanamjeri[1].

*Supervisor.
[1]Student.

And we utilize many probability inequalities from the book of Vershynin[8], the papers of Tropp[6] and [7], mainly the Bernstein Inequality of sub-exponential tail bounds and Matrix Chernoff Inequality to prove the result of Corollary 1.

2. MAIN RESULT: RANDOMIZED HADAMARD TRANSFORM ON EQUAL PARTITIONS FOR REGRESSION TRAINING.

Our main contribution here is that under the research target of Dobriban and Sheng [4] for the finite sampling results of relative efficiency of global OLS estimator to distributed OLS estimator, we focus on what kinds of distribution would let uniform sampling partitions decrease the relative efficiency drastically and RHT could help resolve it. (two methods in Remark 1)

The idea of normalized Hadamard matrix is as follows: A square matrix $\mathbf{H}$ of dimension $n \times n$, possibly complex, is a Hadamard matrix if $\mathbf{H}/\sqrt{n}$ is orthogonal and $|\mathbf{H}_{ij}| = 1$ for all $i, j = 1, \ldots, n$. An example is the Walsh-Hadamard matrix, defined recursively as:

$$\mathbf{H}_n = \begin{pmatrix} \mathbf{H}_{n/2} & \mathbf{H}_{n/2} \\ \mathbf{H}_{n/2} & -\mathbf{H}_{n/2} \end{pmatrix},$$

where $\mathbf{H}_{n/2}$ is the Walsh-Hadamard matrix of order $n/2$ and $\mathbf{H}_1 = 1$. The Hadamard matrix is orthogonal, meaning $\mathbf{H}_n\mathbf{H}_n^T = n\mathbf{I}_n$, where $\mathbf{I}_n$ is the identity matrix of order n.

We chose a distribution called Gaussian Mixture Model (GMM) distribution (defined in Definition 1), and found that uniform sampling would perform bad in this case in terms of the finite sample results of relative efficiency of OLS estimation between global and distributed estimators proposed by Dobriban and Sheng [4] (Lemma 1) for equal-weight averaging.

Then our major contribution is that we show RHT could flatten the variance so that we could achieve a perfect relative efficiency $\mathbb{E}(\mathbf{I}_p, \tilde{\mathbf{X}}_1, ..., \tilde{\mathbf{X}}_K) = 1$ when n tends to infinity. (See Main Result: Corollary 1)

Here the setting is the most general linear regression on the given matrix $\mathbf{X}$ and $\mathbf{Y}$ where the true relationship is $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I}_n)$, and we do ordinary least square (OLS) regression on the given matrix $\mathbf{X}$ and $\mathbf{Y}$. Then we could measure the quality of linear regression by doing the expected mean square error (MSE) on the given matrix $\mathbf{X}$ and $\mathbf{Y}$.

In the standard linear regression model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, $\beta \in \mathbb{R}^p$ represents the true, unknown parameter vector we aim to estimate. The matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ contains the predictor variables (data), $\mathbf{Y} \in \mathbb{R}^n$ contains the response variables, and $\epsilon \in \mathbb{R}^n$ represents the noise term, typically assumed to have zero mean and identity covariance matrix ($\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_n)$, often with $\sigma^2 = 1$ for simplicity in theoretical analysis).

The Ordinary Least Squares (OLS) estimator, denoted by $\hat{\beta}$, provides an estimate of β based on the observed data $(\mathbf{X}, \mathbf{Y})$. It is calculated by minimizing the sum of squared differences between the observed responses $\mathbf{Y}$ and the predicted responses $\mathbf{X}b$:

$$\hat{\beta} = \arg\min_{b\in\mathbb{R}^p} ||\mathbf{Y} - \mathbf{X}b||_2^2$$

Assuming $\mathbf{X}$ has full column rank (i.e., $\mathbf{X}^\top\mathbf{X}$ is invertible), the unique solution is given by the normal equations:

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

The quality of this estimator is often assessed by its Mean Squared Error (MSE), which measures the expected squared Euclidean distance between the estimator $\hat{\beta}$ and the true parameter β. For a fixed design matrix $\mathbf{X}$,

we denote as $M(\hat{\beta}) = \mathbb{E}||\beta - \hat{\beta}||^2$ where $\hat{\beta}$ is the OLS estimator of β. Since the matrix $\mathbf{X}$ and $\mathbf{Y}$ could have infinity many number of rows, i.e., n could tend to infinity, OLS estimation would be of much computational burden with exponential training time.

So the general ideas to do the regression here to prevent computational burden is by distributed sketching, that we allcoate sub training sets into many local machines and then do training on each sub machines as well as an aggregated averaging estimators. Here our main part is using partitioned machines to do the regression.

The averaging estimator is given by $\hat{\beta}_{dist} = \sum_{i=1}^K w_i\hat{\beta}_i$, where the ratio associated to it is w_i for each machine $i = 1, 2, \ldots, K$. We denote as $M(\hat{\beta}_{dist}) = \mathbb{E}||\beta - \hat{\beta}_{dist}||^2$ where $\hat{\beta}_{dist}$ is the distributed estimator of β. Thus the relative efficiency is defined by Lemma 1 as $E(\mathbf{X}_1, \ldots, \mathbf{X}_K) = \frac{M(\hat{\beta})}{M(\hat{\beta}_{dist})}$.

Firstly, we will talk about why we choose to uniformly partition these machines, then we will discuss that if we use RHT before doing uniform sampling in partitions, we should not adjust w_i and just take average by $\frac{1}{K}$. (Lemma 2) Finally, we discuss what is the case if we don't do RHT here, and what kind of $E(\mathbf{I}_p, \mathbf{X}_1, ..., \mathbf{X}_K)$ we could achieve.

Note: this paper discuss the situation of high dimention, but not infinity for the number of columns. By before, we could say n could be infinity large but p, the number of columns is fixed, but could be as high dimension as $p = 1000$.

Corollary 1. *Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be the assumed full rank matrix such that each row of $\mathbf{X}$ is i.i.d. sampled from the proposed Gaussian Mixture Model (GMM) distribution (Definition 1) where there are totally n rows and number of features p is fixed.*

$\tilde{\mathbf{X}} = \mathbf{H}_n\mathbf{D}\mathbf{X}$, where $\mathbf{H}_n$ is the normalized Hadamard matrix of order n, and $\mathbf{D}$ is the Rademacher matrix of order n with diagnoal entries $d_1, d_2, \ldots, d_n$ to be 1 or -1 with equal probability.

And $\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \ldots, \tilde{\mathbf{X}}_K$ are the partitioned machine submatrices after being Randomzied Hadamard Transformed (RHT) from Remark 1.

Then, based on the 2.3 result of Lemma 1, we have the following corollary:

$$\lim_{n\to\infty} \mathbb{E}(\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_K) = 1$$

as long as the condition of 2.8 is satisfied.

2.1 Our focus on proving better relative efficiency[main focus]

we first give the finite sample results of Dobriban and Sheng [4] from section 3.2 in the beginning.

Lemma 1 (Relative Efficiency of distributed linear regression in partitions)**.** *Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be the assumed full rank matrix and $M(\hat{\beta}) = \mathbb{E}||\beta - \hat{\beta}||^2$ be the expected Mean Square Error of OLS estimation.*

Here the relative efficiency is defined as

$$E(\mathbf{X}_1, \ldots, \mathbf{X}_K) = \frac{M(\hat{\beta})}{M(\hat{\beta}_{dist})},$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_K$ are the partitioned submatrices as described in Remark 1.

We have the following results of Expected MSE for global OLS linear regression, partitioned OLS linear regression and Relative Efficiency:

1. *The mean–squared error of the global OLS estimator is*

$$M(\hat{\beta}) = \mathrm{tr}\big[(\mathbf{X}^\top \mathbf{X})^{-1}\big] \tag{2.1}$$

2. *Partition the data into K blocks $\mathbf{X}_1, \ldots, \mathbf{X}_K$, compute local OLS estimates $\hat{\beta}_i$, and aggregate via $\hat{\beta}_{\mathrm{dist}}(w) = \sum_{i=1}^K w_i \hat{\beta}_i$ with weights satisfying $\sum_{i=1}^K w_i = 1$. Then*

$$M\big(\hat{\beta}_{\mathrm{dist}}\big) = \sum_{i=1}^K w_i^2 \ \mathrm{tr}\big[(\mathbf{X}_i^\top \mathbf{X}_i)^{-1}\big] \tag{2.2}$$

3. *The choice $w_i \propto 1/\,\mathrm{tr}\big[(\mathbf{X}_i^\top \mathbf{X}_i)^{-1}\big]$ minimises the risk, yielding optimal efficiency*

$$E(\mathbf{X}_1, \ldots, \mathbf{X}_K) = \mathrm{tr}\big[(\mathbf{X}^\top \mathbf{X})^{-1}\big] \sum_{i=1}^K \frac{1}{\mathrm{tr}\big[(\mathbf{X}_i^\top \mathbf{X}_i)^{-1}\big]} \tag{2.3}$$

Now we will follow this path and discuss the reason for equal partition, and why it is reasonable for us to achieve the dream efficiency after RHT. and we will propose our GMM distribution for uniform sampling to fall.

2.2 The reason for desired equal partition for RHT to function well

From 2.3, we know that here $\mathbf{X}^\top \mathbf{X} = \sum_{i=1}^K \mathbf{X}_i^\top \mathbf{X}_i$, and we denote $\mathbf{M}_i = \mathbf{X}_i^\top \mathbf{X}_i$.

We could also denote $g(\mathbf{M}) := \frac{1}{\mathrm{tr}[\mathbf{M}^{-1}]}$. It is easy to see that $g(\mathbf{M})$ outputs scalar results for each matrix $\mathbf{M}$.

Since our denotation of matrix $\mathbf{M}_i$ is the Gram matrix, we could see that $\mathbf{M}_i$ is positive definite and symmetric. Thus we have $\mathrm{tr}[\mathbf{M}^{-1}]$ is convex since trace function is linear. Thus finally we could conclude that function $g(\mathbf{M})$ is concave.

We use this concave effects to show that ideally uniform partition would achieve the dream Efficiency.

Lemma 2. *Let $g(\mathbf{M})$ be a concave function of the positive definite matrix $\mathbf{M}$. Then*

$$g\left(\sum_{i=1}^K w_i \mathbf{M}_i\right) \geq \sum_{i=1}^K w_i g(\mathbf{M}_i) \tag{2.4}$$

for any $w_i > 0$ with $\sum_{i=1}^K w_i = 1$

Since trace function is linear and scalar ratio inside trace could take out safely, then it is trivial to see that:

$$g\left(\sum_{i=1}^K w_i \mathbf{M}_i\right) = \sum_{i=1}^K w_i g(\mathbf{M}_i) \tag{2.5}$$

when $w_i = \frac{1}{K}$ for all $i = 1, 2, \ldots, K$.

In the Appendix of proof of Lemma 1, we noticed that we have the following rule for deriving the relative efficiency.

$$w_i^\star = \frac{1/a_i}{\sum_{j=1}^K 1/a_j}, \qquad i = 1, \ldots, K, \tag{2.6}$$

★ The Lemma 2 here majorly serves as a new perspective because from the Lemma 2 2.6, it is easy to see that in order to garantee E approach 1 or maximizing E, we need adjust w_i with the value of much difficulty.

★ This is because we need to calculate the trace of the inverse of each local gram matrix $\mathbf{M}_i$ and then adjust w_i accordingly.

★ Thus our desire here is that we just take $w_i = \frac{1}{K}$ for all $i = 1, 2, \ldots, K$ in Lemma 2 2.7 and then we find ways to achieve the dream efficiency of 1 by realizing Lemma 2 2.8.

★ This is why we introduce RHT in this paper and prove it as our main results in Section 6.

Now with our Lemma 2 2.5, it is trivial to derive that when we need the relative efficiency to be 1, we need these two conditions:

$$w_i = \frac{1}{K} \text{ for all } i = 1, 2, \ldots, K \tag{2.7}$$

$$a_i := \mathrm{tr}\big[(\mathbf{X}_i^\top \mathbf{X}_i)^{-1}\big] \text{ remains the same} \tag{2.8}$$

Finally we get the small main result here for what we want:

$$E(\mathbf{X}_1, \ldots, \mathbf{X}_K) = 1, \text{when} \quad w_i = \frac{1}{K}, \quad a_i \text{ is fixed.} \tag{2.9}$$

The idea behind this is that, if we partition the data into K blocks with equal size and equal weight, then we could achieve the dream efficiency of 1.

I have to be rigorous that 2.7 is more like a "sufficient condition", if we partitioned like this and set $w_i = \frac{1}{K}$, then we might achieve the dream efficiency of 1. But this must be relied on the RHT sampling as well as let n tends to infinity. Then under such condition 2.8 is garanteed and we could achieve the desired E = 1.

We would prove this as one of our main result.

2.3 Our proposed RHT and the common uniform sampling

Although theotically the above dream efficiency is not difficult to intepretate, I have to say it is not easy to achive in practice as difference of variance between blocks and the leverage scores always differ. Even if we introduce RHT, I have to say the dream efficiency is achived only when n tends to infinity.

This is the main result we would propose in this paper, now we introduce the sampling method comparisons here.

Remark 1. *The whole row space of the matrix* $\mathbf{X}$ *is partitioned into* K *blocks with equal size.*

The two sampling methods are as follows:

1. ***Uniform Sampling Partition****: Each row of the matrix* $\mathbf{X}$ *is randomly assigned to a machine with probability* $\frac{1}{K}$ *without replacement. We employ the Python function* `np.random.shuffle(indices)` *to shuffle the row indices of matrix* $\mathbf{X}$*, subsequently assigning them to each machine by their indices' positions. It is crucial to note that the assignment of rows is independent as there are no sequential assignment for each row. And once a row is assigned to machine* i*, it cannot be reassigned to any other machine.*
2. ***RHT Sampling Partition****: The matrix* $\mathbf{X}$ *is transformed into* $\tilde{\mathbf{X}} = \mathbf{H}_n\mathbf{D}\mathbf{X}$*, after which uniform sampling is performed on* $\tilde{\mathbf{X}}$*.*

3. THE INTUITION BEHIND GMM DISTRIBUTION

One of our main focus here on the finite result of relative efficiency from Dobriban and Sheng [4] (Lemma 1) is identifying the specific distribution of the data matrix $\mathbf{X}$ where the computed relative efficiency has a drastic decrease when the sampling method here is just the equally partitioned uniform sampling. As we have introduced the effects of Randomized Hadamard Transform (RHT) is to flatten the imbalanced leverage scores or biased variance across different rows of the data matrix $\mathbf{X}$. So the chosen distribution of the data matrix $\mathbf{X}$ here must have biased variance difference bewteen different rows. And we choose the Gaussian Mixture Model (GMM) distribution as our target distribution of the data matrix $\mathbf{X}$ here.

Here, the chosen GMM distribution is composed of a mixture of two independent Multivariate Gaussian Distributions, where $100a_1\%$ proportion of the data rows are sampled from the first Multivariate Gaussian Distribution with mean $0 \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$, and $100a_2\%$ proportion of the data rows are sampled from the second Multivariate Gaussian Distribution with mean $\mu_2 \in \mathbb{R}^p$ and covariance matrix $c\Sigma \in \mathbb{R}^{p\times p}$. Here in $0 \in \mathbb{R}^p$ the scalar value of each index in the vector is 0, and in $\mu_2 \in \mathbb{R}^p$ the scalar value of each index in the vector is μ_2. The covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ is a fixed symmetric positive definite matrix. We have μ_2, a_1, a_2 and c are all positive constants where $a_1 + a_2 = 1$, $c > 1$ and $\mu_2 \neq 0$.

By such distribution Randomized Hadamard Transform (RHT) could narrow down the difference of $\mathrm{tr}\big[(\tilde{\mathbf{X}}_i^\top\tilde{\mathbf{X}}_i)^{-1}\big]$ for $i = 1, 2, \ldots, K$. after flattening the biased variance between two groups of data rows. This helps optimize the computed relative efficiency by simply equal partitions without adjusting w_i based on the value of $\mathrm{tr}\big[(\tilde{\mathbf{X}}_i^\top\tilde{\mathbf{X}}_i)^{-1}\big]$

3.1 The chosen Gaussian Mixture Model (GMM) distribution

Definition 1. Let $\mathbf{X}_{j,*} \in \mathbb{R}^p$ be a generic row of the data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$. Assume $\{\mathbf{X}_{j,*}\}_{j=1}^n$ are i.i.d. with mixture density

$$p_{\mathbf{X}}(x) = a_1\, f_1(x) + a_2\, f_2(x), \qquad x \in \mathbb{R}^p,$$

where

$$f_1(x) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}}\ \exp\Big\{-\tfrac{1}{2}\,x^\top\Sigma^{-1}x\Big\},$$

$$f_2(x) = \frac{1}{(2\pi)^{p/2}\,|c\Sigma|^{1/2}}\ \exp\Big\{-\tfrac{1}{2}\,(x-\boldsymbol{\mu}_2)^\top(c\Sigma)^{-1}(x-\boldsymbol{\mu}_2)\Big\},$$

The mixture parameters satisfy

$$a_1, a_2 > 0, \qquad a_1 + a_2 = 1, \qquad c > 1,$$

$\Sigma \in \mathbb{R}^{p\times p}$ (symmetric positive definite), $\qquad \boldsymbol{\mu}_2 = (\mu_2, \ldots, \mu_2)^\top \neq \mathbf{0}$.

Hence each row is drawn from $\mathcal{N}_p(\mathbf{0}, \Sigma)$ with probability a_1 and from $\mathcal{N}_p(\boldsymbol{\mu}_2, c\Sigma)$ with probability a_2.

Now our targeted distribution has been defined, and we will prove in the following Lemma 3 that there is a direct equalization between the expectation of the local gram matrix after RHT Sampling and the expectation of the local gram matrix after Uniform Sampling both for the same partitioned machine index $i = 1, 2, \ldots, K$ where the sampling method again has been discussed in Remark 1.

Lemma 3. *Under Definition 1, suppose* $\mathbf{X}_i$ *is the* i*-th partitioned machine of matrix* $\mathbf{X}$*, and* $\tilde{\mathbf{X}}_i$ *is the* i*-th partitioned machine of matrix* $\tilde{\mathbf{X}}$ *after Randomized Hadamard Transform (RHT). Then we have:*

$$\mathbb{E}[\tilde{\mathbf{X}}_i^\top\tilde{\mathbf{X}}_i] = \mathbb{E}[\mathbf{X}_i^\top\mathbf{X}_i] = \frac{n}{K}\big[(a_1+a_2c)\Sigma+a_2\mu_2\mu_2^\top\big] = \frac{1}{K}\mathbb{E}[\mathbf{X}^\top\mathbf{X}]$$

This means in expectation, the local gram matrix after RHT Sampling is the same as the local gram matrix after Uniform Sampling for the same partitioned machine with index $i = 1, 2, \ldots, K$. And it is interesting to see that both these two expectations are proportional (scaled by the equal weight $\frac{1}{K}$) to the expectation of the global gram matrix. However, when we explore the real case of the conditions of the trace of the inverse of the local gram matrix which has been partitioned, those has been Randomized Hadamard Transformed would remain more stable and closer to the expectation than those haven't been Randomized Hadamard Transformed because as we have discussed as always RHT could spread out the variance uniformly upon all rows of matrix $\tilde{\mathbf{X}}$ and thus the trace of the inverse of the local gram matrix for each machine would be relatively close to each other when the rows of matrix $\tilde{\mathbf{X}}$ are equally partitioned in size.

Now to prove our propositions and ideas rigorously in statistics, we will mainly focus on the rigorous proof of Corollary 1 here with

the help of Bernstein Inequality of sub-exponential tail bounds and Matrix Chernoff Inequality. Thanks to Prof. Tropp for his research on Subsampled Randomized Hadamard Transform by exploring the applications of Matrix Chernoff Inequality on SRHT.[6] And thanks to Prof. Vershynin for his research on Bernstein Inequality of sub-exponential tail bounds[8] as well as other propositions and lemmas in that book. These contents have helpfully connected sub-exponential norms and sub-gaussian norms.

1. We will prove that $\lim_{n\to\infty} \mathbb{E}(\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_K) = 1$ (Corollary 1), when the condition of 2.7 is satisfied, that is the weight for each local OLS estimator is simply $\frac{1}{K}$ for each machine with index $i = 1, 2, \ldots, K$.

4. PROOF OF COROLLARY 1 (MAIN RESULT)

Proof. To prove this corollary rigorously, we need to show that as $n \to \infty$ and p is fixed and could be as large as 1000, we have:

$$\lim_{n\to\infty} \operatorname{tr}\big[(\tilde{\mathbf{X}}_i^\top \tilde{\mathbf{X}}_i)^{-1}\big] = \operatorname{tr}\big[\big(\tfrac{1}{K}(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})\big)^{-1}\big]$$

where $\tilde{\mathbf{X}} \in \mathbb{R}^{n\times p}$ is a matrix with growing number of rows n and fixed number of columns p. and $\tilde{\mathbf{X}}_i \in \mathbb{R}^{n_i \times p}$ is also a matrix with growing number of rows $n_i = \frac{n}{K}$ and fixed number of columns p.

By inverse decomposition between two matrix A and B, we have:

$$A^{-1} - B^{-1} = B^{-1}(B - A)A^{-1}$$

where $A, B \in \mathbb{R}^{p\times p}$.

Now we utilize a simple inequality that the absolute value of the trace of a matrix M is less than or equal to its Frobenius norm, i.e., $|\operatorname{tr}(M)| \le \|M\|_F$, hence we have:

$$\begin{aligned} \big|\operatorname{tr}(A^{-1}) - \operatorname{tr}(B^{-1})\big| &= \big|\operatorname{tr}\big(A^{-1}(A - B)B^{-1}\big)\big| \\ &\le \big\| B^{-1}(B - A)A^{-1}\big\|_F \\ &\le \|B^{-1}\|_2 \|B - A\|_F \|A^{-1}\|_2 \end{aligned} \tag{4.0}$$

The last line of inequality is derived by the fact inequality that $\|PQ\|_F \le \|P\|_2 \|Q\|_F$.

Now we make a very important denotation here, that is:

$$\mathbf{B} = \frac{1}{K}\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} \qquad \mathbf{A} = \tilde{\mathbf{X}}_i^\top \tilde{\mathbf{X}}_i.$$

Since for each row $(\mathbf{X})_{r,*}$ of matrix $\mathbf{X}$, $(\mathbf{X})_{r,*}$ is i.i.d. sampled from the GMM distribution. Thus we could make reasonable assumption that $\mathbf{X}$ has full column rank, which implies data covariance has full rank (there exists a minimum eigenvalue).

This means the Gram matrix remains well conditioned as n grows.

4.1 Coordinate-wise Concentration

We randomly choose a coordinate pair $(j, k) \in \{1, \ldots, p\}^2$ and focus on the element of the local Gram matrix A on that coordinate pair position. Every row of $\tilde{\mathbf{X}}$ that lands on machine i contributes an outer product, so if we set

$$\mathbf{Y}_r := \big(\tilde{\mathbf{X}}\big)_{r,*} \qquad (r = 1, \ldots, n),$$

the corresponding entry of $\mathbf{A}$ can be written as

$$\mathbf{A}_{jk} \;=\; \sum_{r\in P_i} (\mathbf{Y}_r)_j (\mathbf{Y}_r)_k,$$

where P_i is the set of row indices assigned to local machine i.

By contrast, the global gram matrix scaled by $\frac{1}{K}$ is $\mathbf{B} = \frac{1}{K}\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$ has entry

$$\mathbf{B}_{jk} \;=\; \frac{1}{K}\sum_{r=1}^{n} (\mathbf{Y}_r)_j \, (\mathbf{Y}_r)_k,$$

To unity $\mathbf{A}_{jk}$ and $\mathbf{B}_{jk}$ for their changing indices in the big sum, we introduce a proposed indicator:

$$\mathbf{I}_r := \mathbf{1}_{\{\text{row } r \text{ is assigned to machine } i\}}, \qquad \text{so} \quad \sum_{r=1}^{n} \mathbf{I}_r = n/K.$$

With this introduced indicator, the local entry $\mathbf{A}_{jk}$ becomes

$$\mathbf{A}_{jk} \;=\; \sum_{r=1}^{n} \mathbf{I}_r \, (\mathbf{Y}_r)_j \, (\mathbf{Y}_r)_k.$$

Subtracting $\mathbf{B}_{jk}$ we find that the difference is a sum of independent, centred terms:

$$\mathbf{A}_{jk} - \mathbf{B}_{jk} \;=\; \sum_{r=1}^{n} \Big(\mathbf{I}_r - \tfrac{1}{k}\Big)(\mathbf{Y}_r)_j \, (\mathbf{Y}_r)_k.$$

It is easy to see that:

$$\mathbb{E}[\mathbf{I}_r] = \frac{1}{K}$$

This is because each row has probability $\frac{1}{K}$ of being assigned to any of the K machines, and we have mentioned that our uniform sampling partitions are independent as we just use the numpy.shuffle() to randomized the indices.

Thus we have: this is a sum of independent mean-zero random variables conditioned on the transformed matrix $\tilde{\mathbf{X}}$.

Here we denote $\mathbf{Z}_r$ is an independent zero-mean random variable.:

$$\mathbf{Z}_r := \Big(\mathbf{I}_r - \frac{1}{K}\Big)(\mathbf{Y}_r)_j \, (\mathbf{Y}_r)_k$$

Follow the book[8], we navigate to the theorem of Bernstein's inequality.

Theorem 1 (Bernstein tail bound for sub–exponential summands)**.** *Let $X_1, \ldots, X_N$ be independent, centred random variables that are all sub–exponential. Then for every $t \geq 0$*

$$\Pr\Big\{ \big|\textstyle\sum_{i=1}^N X_i\big| \geq t \Big\} \leq 2 \exp\left[- c \min\left(\frac{t^2}{\sum_{i=1}^N \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}} \right) \right],$$

where $c > 0$ is a universal constant and $\|\cdot\|_{\psi_1}$ denotes the sub–exponential norm.

Our primary goal here is to derive a direct sub-exponential Bernstein bound for $\boldsymbol{\Delta}_{jk}^{(i)} := \mathbf{A}_{jk} - \mathbf{B}_{jk}$.

Now we rewrite:

$$\boldsymbol{\Delta}_{jk}^{(i)} = \sum_{r=1}^{n} \Big(\mathbf{I}_r - \tfrac{1}{K} \Big) (\mathbf{Y}_r)_j (\mathbf{Y}_r)_k = \sum_{r=1}^{n} \Big(\mathbf{I}_r - \tfrac{1}{K} \Big) \mathbf{z}_r$$

where we denote $\mathbf{z}_r = (\mathbf{Y}_r)_j (\mathbf{Y}_r)_k$ here with $\mathbf{I}_r \sim \text{Bernoulli}(\frac{1}{K})$ for any row r, which is independent to each other.

Here it should be noted that although we have conditioned on $\tilde{\mathbf{X}} = \mathbf{H}_n \mathbf{D} \mathbf{X}$, $\mathbf{H}_n$ is a fixed hadamard matrix for order n and $\mathbf{X}$ is also a fixed regression matrix, which is formed after the sampling of n data rows in Definition 1. The conditioning is actually applied only to the matrix $\mathbf{D}$ to fix the randomization of d_i for $i = 1, 2, \ldots, n$.

And now we are only exploring the boundedness of variance and sub-gaussian norm of $\mathbf{X}_{rj}$, we don't have to think about conditioning very early here.

When we are sampling data rows $(\mathbf{X})_{r,*}$ for any row index $r = 1, 2, \ldots, n$, there is a probability of a_1 that this row is drawn from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, and there is a probability of a_2 that this row is drawn from $\mathcal{N}_p(\boldsymbol{\mu}_2, c\boldsymbol{\Sigma})$. Thus we have the following for the j-th element $\mathbf{X}_{rj}$ of the row $(\mathbf{X})_{r,*}$:

$$\mathbf{X}_{rj} = \begin{cases} Z_1, & \text{with prob. } a_1, \\ Z_2, & \text{with prob. } a_2, \end{cases} \quad Z_1 \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{jj}),\ Z_2 \sim \mathcal{N}(\mu_2,\, c\boldsymbol{\Sigma}_{jj}).$$

In our case, $\mathbf{X}$ is fixed after the sampling of n data rows in Definition 1, so either Z_1 or Z_2 is determined for each row. We introduce our lemma 2.5.8. (a) here from the book [8]:

Lemma 4 (Gaussian Distribution is sub–Gaussian)**.** *Let $X \sim \mathcal{N}(0, 1)$. Then X is sub–Gaussian and there exists an absolute constant $C > 0$ such that*

$$\|X\|_{\psi_2} \leq C.$$

More generally, if $X \sim \mathcal{N}(0, \sigma^2)$ for some $\sigma > 0$, then

$$\|X\|_{\psi_2} \leq C\sigma.$$

From Lemma 4 above, we have $\|Z_1\|_{\psi_2} \leq \gamma \sqrt{\boldsymbol{\Sigma}_{jj}}$ for some absolute constant $\gamma > 0$, and we have $\|Z_2 - \mu_2\|_{\psi_2} \leq \omega \sqrt{c\boldsymbol{\Sigma}_{jj}}$ for some absolute constant $\omega > 0$, this implies that we have $\|Z_2\|_{\psi_2} \leq \omega \sqrt{c\boldsymbol{\Sigma}_{jj}} + \|\mu_2\|_{\psi_2}$.

Then we define $\kappa = \max\big\{ \gamma \sqrt{\boldsymbol{\Sigma}_{jj}}, \omega \sqrt{c\boldsymbol{\Sigma}_{jj}} + \|\mu_2\|_{\psi_2} \big\}$ and have:

$$\|\mathbf{X}_{rj}\|_{\psi_2} \leq \kappa$$

Now again by ideas before, here we rewrite again:

$$(\mathbf{Y}_r)_j = \sum_{i=1}^{n} h_{ri} d_i \mathbf{X}_{ij}$$

And we denote:

$$\eta_i := h_{ri} d_i \mathbf{X}_{ij}$$

Hence, each η_i is an indepedent sub-Gaussian random variable with mean zero and a sub-gaussian norm bound of $\frac{\kappa}{n}$.

Follow the book of [8], we navigate to the theorem of sum of independent sub–Gaussian random variables in proposition 2.6.1.

Theorem 2 (Sum of independent sub–Gaussian random variables)**.** *Let $X_1, \ldots, X_N$ be independent, mean–zero, sub–Gaussian random variables. Then the partial sum $S := \sum_{i=1}^N X_i$ is itself sub–Gaussian and satisfies*

$$\|S\|_{\psi_2}^2 \leq C \sum_{i=1}^{N} \|X_i\|_{\psi_2}^2,$$

where $C > 0$ is an absolute constant.

Applying Theorem 2 to the independent η_i gives

$$\|(\mathbf{Y}_r)_j\|_{\psi_2}^2 \leq C \sum_{i=1}^{n} \|\eta_i\|_{\psi_2}^2 \leq C\,\kappa$$

$$\Longrightarrow \qquad \|(\mathbf{Y}_r)_j\|_{\psi_2} \leq \sqrt{C\,\kappa},$$

where C is the absolute constant in the Theorem 2.

Thus by the lemma of the book[8] which is Lemma 2.7.7, we have:

$$\|\mathbf{z}_r\|_{\psi_1} = \|(\mathbf{Y}_r)_j (\mathbf{Y}_r)_k\|_{\psi_1} \leq C\,\kappa = b$$

as we denote $b := C\,\kappa$, which is a new absolute constant here.

By the fact that $\|\mathbf{Z}_r\|_{\psi_1} \leq \|\mathbf{z}_r\|_{\psi_1} \leq b$, we have the following inequality for the sum of sub-exponential norm of $\mathbf{Z}_r$ where we denote as $\boldsymbol{\nu}^2$:

$$\boldsymbol{\nu}^2 := \sum_{r=1}^{n} \|\mathbf{Z}_r\|_{\psi_1}^2 \leq nb^2$$

Thus we could plug this $\boldsymbol{\nu}^2$ and $\max_r \|\mathbf{Z}_r\|_{\psi_1} \leq b$ into Theorem 1.

After the augment of the bound at right hand side by these two inequalities plugged in, We have the following inequality:

$$\Pr\big\{ |\boldsymbol{\Delta}_{jk}^{(i)}| \geq t \big\} \leq 2 \exp\left[- c \min\left(\frac{t^2}{nb^2}, \frac{t}{b} \right) \right].$$

To let the second term dominate, we choose:

$$t = \chi\sqrt{n\log n} \qquad \text{for some } \chi > 0.$$

Thus we have:

$$\Pr\{|\mathbf{\Delta}_{jk}^{(i)}| \geq \chi\sqrt{n\log n}\} \;\leq\; 2\,\exp\big[-\,c\,\tfrac{\chi^2}{b^2}(\log n)\big] \;=\; 2\,n^{-c\,\chi^2/b^2}$$

Because there are at most p^2 entries for all the (j,k) entries, hence by the law of probability union we have:

$$\Pr\Big\{\max_{j,k}|\mathbf{\Delta}_{jk}^{(i)}| \;\geq\; \chi\sqrt{n\,\log n}\Big\} \;\leq\; 2\,p^2\,n^{-\frac{c\,\chi^2}{b^2}}.$$

Thus we have :

$$\|\mathbf{A}-\mathbf{B}\|_F^2 \;=\; \textstyle\sum_{j,k}(|\mathbf{\Delta}_{jk}^{(i)}|)^2 \;\leq\; \sum_{j,k}\big(\max_{j,k}|\mathbf{\Delta}_{jk}^{(i)}|\big)^2 \;=\; p^2\,\chi^2\,n\,\log n$$

With probability at least $1-2\,p^2\,n^{-\frac{c\,\chi^2}{b^2}}$.
Hence we have our finding here:

$$\Pr\Big\{\;\|\mathbf{A}-\mathbf{B}\|_F \leq p\,\chi\sqrt{n\,\log n}\;\Big\} \;\geq\; 1-2\,p^2\,n^{-\frac{c\,\chi^2}{b^2}}. \quad (4.1)$$

4.2 L2 norm of inverse of gram matrix

Here we introduce the theorem proposed by Prof. Tropp from his paper on the analysis of subsampled randomized hadamard transform (SRHT)[6] with is marked as Theorem 2.2 in that paper. This Matrix Chernoff Theorem is a very useful tool to bound the minimum eigenvalues of $\mathbf{A}$ and $\mathbf{B}$ in our case, then to bound $\|\mathbf{A}^{-1}\|_2$ and $\|\mathbf{B}^{-1}\|_2$.

Theorem 3 (Matrix Chernoff)**.** *Let $\{\mathbf{X}_l\}$ be a finite sequence of independent, random, self–adjoint $p\times p$ matrices satisfying $\mathbf{X}_l \succeq 0$ and $\lambda_{\max}(\mathbf{X}_l)\leq R$ almost surely. Define*

$$\mu_{\min} = \lambda_{\min}\Big(\sum_l \mathbb{E}[\mathbf{X}_l]\Big), \quad \mu_{\max} = \lambda_{\max}\Big(\sum_l \mathbb{E}[\mathbf{X}_l]\Big).$$

Then for any $\epsilon\in[0,1]$,

$$\Pr\Big\{\lambda_{\min}\Big(\sum_l \mathbf{X}_l\Big) \leq (1-\epsilon)\,\mu_{\min}\Big\} \;\leq\; p\left[\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right]^{\mu_{\min}/R},$$

and for any $\epsilon\geq 0$,

$$\Pr\Big\{\lambda_{\max}\Big(\sum_l \mathbf{X}_l\Big) \geq (1+\epsilon)\,\mu_{\max}\Big\} \;\leq\; p\left[\frac{e^{\,\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right]^{\mu_{\max}/R}.$$

Corollary 2 (Simplified lower-tail bound)**.** *Under the same hypotheses as Theorem 3, for any $\epsilon\in[0,1]$ one has*

$$\Pr\Big\{\;\lambda_{\min}\big(\textstyle\sum_l \mathbf{X}_l\big) \leq (1-\epsilon)\,\mu_{\min}\Big\} \;\leq\; p\,\exp\Big(-\tfrac{\epsilon^2\,\mu_{\min}}{2R}\Big).$$

Proof. Starting from the bound:

$$\Pr\Big\{\lambda_{\min}\Big(\sum_l \mathbf{X}_l\Big) \leq (1-\epsilon)\,\mu_{\min}\Big\} \;\leq\; p\left[\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right]^{\mu_{\min}/R},$$

We use the standard inequality here that is: $\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}} \leq e^{-\epsilon^2/2}$, which is valid for $0\leq\epsilon\leq 1$.

Substituting this inequality gives the right hand side exponential result:

$$\big[\exp\big(-\tfrac{\epsilon^2}{2}\big)\big]^{\mu_{\min}/R} = \exp\Big(-\tfrac{\epsilon^2\,\mu_{\min}}{2R}\Big).$$

Then this is claimed of the simplified lower-tail bound.

Thus we have the probability inequality of the minimum eigenvalue of the sum of independent random matrices becomes:

$$\Pr\Big\{\lambda_{\min}\Big(\sum_l \mathbf{X}_l\Big) \leq (1-\epsilon)\,\mu_{\min}\Big\} \;\leq\; p\,\exp\Big(-\tfrac{\epsilon^2\,\mu_{\min}}{2R}\Big)$$

□

Now we denote $\mathbf{W}_r = \mathbf{Y}_r\mathbf{Y}_r^T$ is positive semidefinite, and we have $\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} = \sum_{r=1}^n \mathbf{W}_r$, where $\mathbf{Y}_r$ is each row of matrix $\tilde{\mathbf{X}}$ that has been randomized hadamard transformed.

To satisfy the condition of Corollary 2, we need to show that:

$$\lambda_{\max}(\mathbf{W}_r) \leq R$$

where R is a corresponding bound, it should be noted that in the Theorem 3, the condition is satisfied almostly surely. This means some asymptotical bound also meets the condition here.

To prove the above inequality, it is equivalent to show that $\|\mathbf{Y}_r\|_2^2 \leq R$ for all r simultaneously.

From before we have shown that:

$$\|(\mathbf{Y}_r)_j\|_{\psi_2}^2 \leq C\sum_{i=1}^n \|\eta_i\|_{\psi_2}^2 \leq C\,\kappa$$

Thus we have again by the lemma of 2.7.7. from the book[8]:

$$\|(\mathbf{Y}_r)_j{}^2\|_{\psi_1} \leq \|(\mathbf{Y}_r)_j\|_{\psi_2}\|(\mathbf{Y}_r)_j\|_{\psi_2} = \|(\mathbf{Y}_r)_j\|_{\psi_2}^2 \leq C\,\kappa = b$$

It is then trivial to have that:

$$max_j\|(\mathbf{Y}_r)_j{}^2\|_{\psi_1} \leq b$$

Now for simplicity, we denote $\mathbf{R}_r$ as the sum of $(\mathbf{Y}_r)_j{}^2$:

$$\mathbf{R}_r := \|\mathbf{Y}_r\|_2^2 = \sum_{j=1}^p (\mathbf{Y}_r)_j{}^2$$

Since $\|(\mathbf{Y}_r)_j{}^2\|_{\psi_1} \leq b$, we have the following Inequality:

$$\sum_{j=1}^p \|(\mathbf{Y}_r)_j{}^2\|_{\psi_1}^2 \leq \sum_{j=1}^p b^2 = pb^2$$

Now we could see that the two inequality results here are very similar to before, only infinite n is replaced by finite p.

Again, we use the Theorem 1 here to derive the bound of $\|\mathbf{Y}_r\|_2^2$.

We take $t = C_0 \log n$ here, because the two denominators here are both constants that don't grow with n due to our assumption here that p could be very large but is fixed. This could ensure that first term dominants, and $\frac{t}{max_j \|(\mathbf{Y}_r)_j{}^2\|_{\psi_1}}$ grow with n.

Again by The Law of Union of Probability and plugging with $t = C_0 \log n$, we have:

$$\Pr\left\{\max_{1\le r\le n} \mathbf{R}_r > C_0 \log n\right\} \;\le\; 2\, n^{\,1-\frac{c\,C_0}{b}} \tag{4.2}$$

By the illustration above, the $\lambda_{\max}$ of $\mathbf{W}_r$ is bounded by $C_0 \log n$ with probability at least $1 - 2n^{\,1-\frac{c\,C_0}{b}}$.

That is:

$$\Pr\left\{\lambda_{\max}(\mathbf{W}_r) > C_0 \log n\right\} \;\le\; 2\, n^{\,1-\frac{c\,C_0}{b}}$$

In order to make this probability $2\, n^{\,1-\frac{c\,C_0}{b}}$ be exponentially small as n tends to infinity, we only need to choose the absolute constant C_0 such that $C_0 \ge \frac{2b}{c}$.

So we have the following as our result of the probability inequality of the bound of $\lambda_{\max}(\mathbf{W}_r)$:

$$\Pr\left\{\lambda_{\max}(\mathbf{W}_r) > C_0 \log n\right\} \;\le\; 2\, n^{\,1-\frac{c\,C_0}{b}}, \quad C_0 \ge \frac{2b}{c}$$

Now we have garantee the bounding effect of $\lambda_{\max}$ of $\mathbf{W}_r$.

And now we could utilize the Theorem 3 to bound the minimum eigenvalue of $\mathbf{B}$ by denoting the expectation of $\mathbf{W}_r$ as $\mathbf{\Sigma}^\star$ first.

$$\mathbb{E}[\mathbf{W}_r] = \mathbb{E}[\mathbf{Y}_r \mathbf{Y}_r^T] = \left((a_1 + a_2 c)\,\mathbf{\Sigma} + a_2\, \boldsymbol{\mu}_2 \boldsymbol{\mu}_2^\top\right) := \mathbf{\Sigma}^\star$$

Then we denote the minimum eigenvalue of $\mathbf{\Sigma}^\star$ as $\lambda^\star$, and we denote the $\mu^\star$ as the minimum eigenvalue of the sum of $\mathbb{E}[\mathbf{W}_r]$:

$$\lambda^\star = \lambda_{\min}(\mathbf{\Sigma}^\star), \mu^\star = \mu_{\min} = \lambda_{\min}\Big(\sum_r \mathbb{E}[\mathbf{W}_r]\Big) = \frac{n}{K}\lambda^\star$$

Then we have by the Corollary 2:

$$\Pr\{\lambda_{\min}\left(\frac{\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}}{K}\right) < \frac{(1-\sigma) n \lambda^\star}{K}\} \le p \exp\Big(-\frac{\iota n}{\log n}\Big)$$

where $\iota = \frac{\sigma^2 \lambda^\star}{2KC_0}$ is a constant and σ is an absolute constant from 0 to 1.

Now we consider the worst case that the two probabilities here, the probability of $\lambda_{\max}(\mathbf{W}_r)$ is larger than its bound and the probability of $\lambda_{\min}(\frac{\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}}{K})$ is larger than its bound, have no intersection. Then with probability at least:

$$1 - p \exp\Big(-\frac{\iota n}{\log n}\Big) - 2\, n^{\,1-\frac{c\,C_0}{b}}$$

We have

$$\lambda_{\min}(\mathbf{B}) = \lambda_{\min}(\frac{\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}}{K}) \ge \frac{(1-\sigma) n \lambda^\star}{K}$$

This implies the important result about the probability inequality of the bound of $\|\mathbf{B}^{-1}\|_2$ here we desire. Due to the fact that $\|\mathbf{B}^{-1}\|_2 = \frac{1}{\lambda_{\min}(\mathbf{B})}$, we have:

$$\Pr\left\{\|\mathbf{B}^{-1}\|_2 \le \tfrac{K}{(1-\sigma)\lambda^\star} \cdot \tfrac{1}{n}\right\} \ge 1 - p \exp\Big(-\frac{\iota n}{\log n}\Big) - 2\, n^{\,1-\frac{c\,C_0}{b}} \tag{4.3}$$

Now we consider the case of $\mathbf{A} = \tilde{\mathbf{X}}_i^\top \tilde{\mathbf{X}}_i = \sum_{r\in \mathbf{I}_i} \mathbf{Y}_r \mathbf{Y}_r^T$ where $\mathbf{I}_i$ is the index set of the rows in the i-th machine. It should be noted that the result here for boundedness of $\lambda_{\min}(\mathbf{A})$ is exactly the same to the result of $\lambda_{\min}(\mathbf{B})$ above. The reason is that from Lemma 3 we know $\mathbb{E}[\mathbf{A}] = \mathbb{E}[\mathbf{B}]$ due to the fact that $\mathbb{E}[\mathbf{X}^\top \mathbf{X}] = \mathbf{E}[\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}]$. And the result of the boundedness of $\lambda_{\min}(\mathbf{W}_r)$ is exactly the same for both $\mathbf{A}$ and $\mathbf{B}$ for any row index r.

Now we write down the same result but here for $\|\mathbf{A}^{-1}\|_2$:

$$\Pr\left\{\|\mathbf{A}^{-1}\|_2 \le \tfrac{K}{(1-\sigma)\lambda^\star} \cdot \tfrac{1}{n}\right\} \ge 1 - p \exp\Big(-\frac{\iota n}{\log n}\Big) - 2\, n^{\,1-\frac{c\,C_0}{b}} \tag{4.4}$$

Again, only make sure $C_0 \ge \frac{2b}{c}$.

Now we do the final conclude here:

From 4.1, 4.3, 4.4, we have the following result by the union law of probability:

With probability at least:

$$1 - 2p \exp\Big(-\frac{\iota n}{\log n}\Big) - 4\, n^{\,1-\frac{c\,C_0}{b}} - 2p^2\, n^{-\frac{c\,\chi^2}{b^2}}$$

We have:

$$\|\mathbf{A}^{-1}\|_2 \cdot \|\mathbf{B}^{-1}\|_2 \cdot \|\mathbf{A} - \mathbf{B}\|_F \le \frac{K^2 p \chi}{(1-\sigma)^2 (\lambda^\star)^2} \sqrt{\frac{\log n}{n^3}}$$

Then we have the following result by the inequality of 4.0 which has been discussed before:

$$|\operatorname{tr}(\mathbf{A}^{-1}) - \operatorname{tr}(\mathbf{B}^{-1})| \le \frac{K^2 p \chi}{(1-\sigma)^2 (\lambda^\star)^2} \sqrt{\frac{\log n}{n^3}}$$

with the probability that has been discussed above.

Thus the final result here is that:

$$\begin{aligned}\Pr\left\{|\operatorname{tr}(\mathbf{A}^{-1}) - \operatorname{tr}(\mathbf{B}^{-1})| \le \frac{K^2 p \chi}{(1-\sigma)^2 (\lambda^\star)^2} \sqrt{\frac{\log n}{n^3}}\right\} \\ \ge 1 - 2p \exp\Big(-\frac{\iota n}{\log n}\Big) - 4\, n^{\,1-\frac{c\,C_0}{b}} - 2p^2\, n^{-\frac{c\,\chi^2}{b^2}}\end{aligned} \tag{4.5}$$

Result 4.5 implies that $\operatorname{tr}(\mathbf{A}^{-1})$ converges to $\operatorname{tr}(\mathbf{B}^{-1})$ with probability 1 as $n \to \infty$. This is the result we desire to show in the very beginning:

$$\lim_{n\to\infty} \operatorname{tr}\big[(\tilde{\mathbf{X}}_i^\top \tilde{\mathbf{X}}_i)^{-1}\big] = \operatorname{tr}\big[\big(\tfrac{1}{K}(\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})\big)^{-1}\big]$$

We complete the proof of main result here for the dreamming E = 1 in section 4.

□

5. UNIFORM SAMPLING

Consider $\mathbf{x} \in \mathbb{R}^{n\times 1}$, where $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$, which is a column vector with entries to be sampled randomly. Now we define the term "energy" as the squared Euclidean norm of $\mathbf{x}$, i.e., $E(\mathbf{x}) = \|\mathbf{x}\|_2^2 = \sum_{i=1}^n \mathbf{x}_i^2$, measuring strength of the magnitude. We use the sum of squared entries in the sampling set, i.e., $\mathbf{Sum} = \sum_{j\in\mathbf{S}} \mathbf{x}_j^2$, where $|\mathbf{S}| = \ell$, which is the set of the numbers of indices sampled.

Thus, we could denote the estimator of the energy as $\hat{E}(\mathbf{x}) = \frac{n}{\ell}\sum_{j\in\mathbf{S}} \mathbf{x}_j^2$. This raises problems as when the $\mathbf{x}_j$ entry are not comparable in magnitude, i.e. some entries are much larger than others and their number is very small, such estimator would be biased because missing large entries would be harmful to the estimation.

5.1 Expectation of the estimator

We define $M(\mathbf{x}, \mathbf{S})$ to be the estimator of ℓ_2 squared norm of the vector $\mathbf{x}$ and $\mathbf{S}$ is the subsample of $\{1, 2, \ldots, n\}$ to be uniform sampled. We calculate π_j for index j to be sampled in $\mathbf{S}$:

$$\pi_j = \frac{\binom{n-1}{\ell-1}}{\binom{n}{\ell}} = \frac{\ell}{n} \tag{5.1}$$

Thus we have:

$$\Pr(\delta_j = 1) = \frac{\ell}{n}, \quad \Pr(\delta_j = 0) = 1 - \frac{\ell}{n}, \quad j = 1, 2, \ldots, n \tag{5.2}$$

where δ_j is the indicator function for j to be sampled in $\mathbf{S}$. We have:

$$\delta_j = \begin{cases} 1, & \text{if } j \in \mathbf{S}, \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

Now $M(\mathbf{x}, \mathbf{S}) = \frac{n}{\ell}\sum_{j=1}^n \delta_j \mathbf{x}_j^2$, where $\mathbf{x}_j$ is the jth entry of the vector $\mathbf{x}$. Thus we have the expectation of $M(\mathbf{x}, \mathbf{S})$:

$$\mathbb{E}[M(\mathbf{x}, \mathbf{S})] = \frac{n}{\ell}\sum_{j=1}^n \mathbb{E}[\delta_j]\mathbf{x}_j^2 = \frac{n}{\ell}\sum_{j=1}^n \frac{\ell}{n}\mathbf{x}_j^2 = \sum_{j=1}^n \mathbf{x}_j^2 = E(\mathbf{x}) \tag{5.4}$$

since $\mathbb{E}[\delta_j] = \Pr(j \in \mathbf{S}) \cdot 1 + \Pr(j \notin \mathbf{S}) \cdot 0 = \Pr(j \in \mathbf{S}) = \frac{\ell}{n}$.

The result here is that uniform sampling is unbiased for the energy of the vector $\mathbf{x}$. However, unbiased doen't mean it is a good estimator. We will explore when uniform sampling would fall in estimating the energy of the vector $\mathbf{x}$.

5.2 Failure of Uniform Row-Wise Sampling Partitioning Under a Two-Cluster GMM

From our chosen Gaussian Mixture Model (GMM) distribution in Definition 1, we define two clusters of separate multivariate Gaussian distributions as follows: Cluster 1: $\mathbf{X}_{j,*} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ with sampling probability a_1 and Cluster 2: $\mathbf{X}_{j,*} \sim \mathcal{N}_p(\boldsymbol{\mu_2}, c\mathbf{\Sigma})$ with sampling probability a_2, where $a_1 + a_2 = 1, a_1 > 0, a_2 > 0$ and $c > 1$ for inflating the covariance matrix. Here $\mathbf{X}_{j,*}$ is any row of the matrix $\mathbf{X}$.

Such mixture construction of the distribution is to ensure that Cluster 2 rows have higher variance than Cluster 1 rows, which means Cluster 2 rows typically have larger norms due to inflated covariance matrix factor c. These data rows mainly serve as the high leverage-scores observations, which are the outliers of $\mathbf{X}$ that might influence the Ordinary Least Square regression in our research. It should be noted that the leverage score of a row could be expressed as $h_j = \mathbf{X}_{j,*}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}_{j,*}$, so large $\|\mathbf{X}_{j,*}\|_2$ or $\mathbf{X}_{j,*}$ lying in a sparsely sampled direction [5] would lead to large leverage score h_j.

We will now argue that under our chosen GMM distribution, using equally partitioned uniform sampling ($n_i = \frac{n}{K}$), as well as each row partitioned into any machine with probability $\frac{1}{K}$ for distributed OLS regression would lead to imbalanced partitions, some machines might get many high-leverage, high-variance data rows from Cluster 2 while others might get few. This would cause biased $\mathrm{tr}[(\mathbf{X}_i^\top\mathbf{X}_i)^{-1}]$ for each partitioned machine matrix $\mathbf{X}_i$.

5.3 Uneven Distribution of High-Leverage Rows under Uniform Sampling

Because the row assignments of all the data rows $\mathbf{X}_{j,*}$, $j = 1, 2, \ldots, n$ are independent and identically distributed (i.i.d.), we could see that the number of Cluster 2 sampled rows in each machine follows an approximate binomial distribution or hypergeometric distribution. Thus we denote that m_i to be the number of Cluster 2 data rows in the i-th machine, and $m_i \sim \text{Binomial}(n/K, a_2)$ when n is large enough. Then it is easy to see that $\mathbb{E}[m_i] = \frac{n}{K}a_2$. However in high probability under uniform sampling each machine with index i would not have exactly $\frac{n}{K}a_2$ rows from Cluster 2, but rather a fluctuated number.

In fact, a standard Chernoff/Hoeffding bound states the following inequality for any partition i about the deviation of m_i from its expectation $\mathbb{E}[m_i]$:

$$\Pr\left\{|m_i - a_2\tfrac{n}{K}| \geq \delta_n a_2 \tfrac{n}{K}\right\} \leq 2\exp\left(-\tfrac{\delta_n^2 a_2 \frac{n}{K}}{3}\right), \quad \delta_n \in (0,1)$$

This means:

$$\Pr\left\{|m_i - a_2\tfrac{n}{K}| \leq \delta_n a_2 \tfrac{n}{K}\right\} \geq 1 - 2\exp\left(-\tfrac{\delta_n^2 a_2 \frac{n}{K}}{3}\right), \quad \delta_n \in (0,1)$$

Now we consider the case of all the K machines with each machine i have such probability inequality convergence, we have the following result here:

$$\Pr\left\{\max_i |m_i - a_2\tfrac{n}{K}| \leq \delta_n a_2 \tfrac{n}{K}\right\} \geq 1 - 2K\exp\left(-\tfrac{\delta_n^2 a_2 \frac{n}{K}}{3}\right), \quad \delta_n \in (0,1)$$

Now we take $\delta_n = \sqrt{\frac{\log n}{a_2 \frac{n}{K}}}$, it is easy to see that this satisfies the condition of $\delta_n \in (0,1)$ when n is large enough, thus we have:

$$\Pr\left\{\max_i |m_i - a_2\tfrac{n}{K}| \leq \sqrt{a_2 \tfrac{n}{K}\log n}\right\} \geq 1 - \tfrac{2K}{n^{\frac{1}{3}}} \tag{5.5}$$

Hence from the probability inequality of maximum deviation of m_i from its expectation $\mathbb{E}[m_i]$, we could see that the absolute bound of such deviation could be as large as $\sqrt{a_2 \frac{n}{K} \log n}$, which means the value of m_i could be as large as $\frac{n}{K} a_2 + \sqrt{a_2 \frac{n}{K} \log n}$ or as small as $\frac{n}{K} a_2 - \sqrt{a_2 \frac{n}{K} \log n}$. This just means that we have shown the uneven distribution of high-leverage data rows from Cluster 2 in all the K machines, which leads to the biased estimation of $\mathrm{tr}[(\mathbf{X}_i^\top \mathbf{X}_i)^{-1}]$ for each machine i due to biased compositions of two Clusters of data rows in each machines for index i.

5.4 Deviation of local gram matrix from its expectation

As we have defined $\mathbf{X_i}$ as the i-th partitioned local matrix and $\mathbf{M_i} = \mathbf{X_i}^\top \mathbf{X_i}$ as the i-th local gram matrix, we have the expectation of $\mathbf{M_i}$ from Lemma 3 as:

$$\mathbb{E}[\mathbf{M_i}] = \frac{n}{K} \left((a_1 + a_2 c) \, \mathbf{\Sigma} + a_2 \, \boldsymbol{\mu}_2 \boldsymbol{\mu}_2^\top \right)$$

Ideally, for a good result of consistent distributed ordinary least square (OLS) regression as we have discussed in Lemma 1, each i-th local gram matrix $\mathbf{M_i}$ should be close to its expectation $\mathbb{E}[\mathbf{M_i}]$. However, due to the uneven distribution of high-leverage data rows from Cluster 2 in all the K machines as we have discussed above in the last subsection, we could see that each local gram matrix $\mathbf{M_i}$ can differ greatly. Now we could decompose $\mathbf{M}_i = \mathbf{M}_i^{(1)} + \mathbf{M}_i^{(2)}$ into contributions from Cluster 1 and Cluster 2, which are sums of $\mathbf{X}_{j,*} \mathbf{X}_{j,*}^\top$ in both two Clusters respectively. As we have discussed the high variance of Cluster 2 row number m_i in each machine i above, the Cluster 2 contribution $\mathbf{M}_i^{(2)}$ is also of much variance across i. We take some reasonable instances as follows:

- As we have discussed that the value of m_i could be as small as $\frac{n}{K} a_2 - \sqrt{a_2 \frac{n}{K} \log n}$, assume in a practical case that $\frac{n}{K} a_2 - \sqrt{a_2 \frac{n}{K} \log n} \leq 0$. In such case the real worst case is that partition i gets no Cluster 2 points ($m_i = 0$). Thus we have $\mathbf{M}_i = \mathbf{M}_i^{(1)}$ is a sample covariance matrix of $\frac{n}{K}$ Cluster 1 rows from $N_p(\mathbf{0}, \mathbf{\Sigma})$, which follows that $\mathbf{M}_i$ is approximately a Wishart distribution with $\frac{n}{K}$ degrees of freedom and covariance matrix $\mathbf{\Sigma}$: $\mathbf{M}_i \sim \mathcal{W}_d(\frac{n}{K}, \mathbf{\Sigma})$. This $\mathbf{M}_i$ would be of much smaller scale in $\boldsymbol{\mu}_2$ direction as there is no c inflated Cluster 2 rows. Essentially, partition i "misses" the high-variance Cluster 2, so its covariance estimate underrepresents the directions of $\mathbf{\Sigma}$ that Cluster 2 emphasizes.
- As we have discussed that the value of m_j could be as large as $\frac{n}{K} a_2 + \sqrt{a_2 \frac{n}{K} \log n}$, we could assume here that $m_j \gg \frac{n}{K} a_2$ in such case of any j-th partitioned machine. Then we could rewrite as follows:

$$\mathbf{M}_j^{(2)} = \sum_{r=1}^{m_j} \mathbf{X}_{j,r} \, \mathbf{X}_{j,r}^T = m_j (\boldsymbol{\mu}_2 \boldsymbol{\mu}_2^T) + \sum_{r=1}^{m_j} \boldsymbol{\epsilon}_r \, \boldsymbol{\epsilon}_r^T + \text{(cross terms)}$$

where $\boldsymbol{\epsilon}_r = \mathbf{X}_{j,r} - \boldsymbol{\mu}_2 \sim \mathcal{N}(\mathbf{0}, c\mathbf{\Sigma})$. Thus we have $\mathbf{M}_j^{(2)}$ is a sum of $m_j \boldsymbol{\mu}_2 \boldsymbol{\mu}_2^\top$ (a low-rank component in the $\boldsymbol{\mu}_2$ direction) and a Wishart-like component of m_j degrees of freedom and covariance matrix $c\mathbf{\Sigma}$. Thus $\mathbf{M}_j$ will have a much larger variance along the $\boldsymbol{\mu}_2$ direction (due to the $\boldsymbol{\mu}_2 \boldsymbol{\mu}_2^\top$ term and the inflated noise $c\mathbf{\Sigma}$) than expected, and overall norm $|\mathbf{M}_j|$ is higher.

Again, similar to our ideas of binomial distribution discussed in the last subsection, we could apply a standard matrix Bernstein bound here from the paper of Tropp [7] for $Theorem 1.6$ in that paper:

$$\begin{aligned} \Pr \left\{ \left\| \mathbf{M}_i - \frac{n}{K} \left((a_1 + a_2 c) \, \mathbf{\Sigma} + a_2 \, \boldsymbol{\mu}_2 \boldsymbol{\mu}_2^\top \right) \right\| \leq \epsilon \frac{n}{K} \right\} \\ \geq 1 - 2p \exp \left(- \frac{3 \frac{n}{K} \epsilon^2}{6 \sigma_{\max}^2 (c, \boldsymbol{\mu_2}) + 2 \epsilon C \log(\frac{n}{K})} \right) \end{aligned} \tag{5.6}$$

where $\sigma_{\max}^2(c, \boldsymbol{\mu}_2)$ is a tail variance proxy (on the order of $c\lambda_{\max}(\mathbf{\Sigma}) + |\boldsymbol{\mu}_2|^2$) and C is an absolute constant here.

To apply the matrix Bernstein bound (Theorem 1.6 in [7]), we identify the variance parameter σ^2 with $\frac{n}{K} \sigma_{\max}^2(c, \boldsymbol{\mu}_2)$ and the matrix norm bound R with $C \log(\frac{n}{K})$. The latter bounds the maximum squared row norm $\max_{1 \leq j \leq \frac{n}{K}} \|\mathbf{X}_{j,*}\|_2^2$, based on an argument similar to that leading to 4.2.

Thus for any fixed $\epsilon \in (0, 1)$ and n is large enough comparable to p, we have the norm of the deviation of $\mathbf{M}_i$ from its expectation $\mathbb{E}[\mathbf{M}_i]$ is bounded by $\epsilon \frac{n}{K}$ with high probability, and this bound could be very large.

6. MERITS OF RANDOMIZED HADAMARD TRANSFORM

The pivotal point of Randomized Hadamard Transform (RHT) is that it could flattern the vector variance or redistribute the energy we have discussed.

Here we define the normalized Hadamard matrix as $\mathbf{H}_n = \frac{1}{\sqrt{n}} \mathbf{H}$, where $\mathbf{H}$ is the Hadamard matrix of order n. Thus it is obvious that $(\mathbf{H}_n)_{i,j} = \frac{1}{\sqrt{n}}$ or $(\mathbf{H}_n)_{i,j} = -\frac{1}{\sqrt{n}}$ depending on the position of entry (i, j) for the Hadamard matrix $\mathbf{H}$. And for the simplicity of notation, we denote $(\mathbf{H}_n)_{i,j} = h_{ij}$ in all this article's work. It should be noted that normalized Hadamard matrix is fixed for a given order n, and it is orthogonal, i.e. $\mathbf{H}_n^\top \mathbf{H}_n = \mathbf{I}_n$.

We define the Rademacher element as $d_j = \pm 1$ with equal probability $\frac{1}{2}$ for $j = 1, 2, \ldots, n$. The Rademacher matrix is given by $\mathbf{D} = \mathrm{diag}(d_1, d_2, \ldots, d_n)$. Clearly, $\mathbf{D}$ is random, and $\mathbb{E}[d_j] = 0$ while $\mathrm{Var}(d_j) = 1$. Additionally, $\mathbf{D}$ is orthogonal, i.e., $\mathbf{D}^\top \mathbf{D} = \mathbf{I}_n$.

Thus for a given vector $\mathbf{x}$, we have $\mathbf{y} = \mathbf{H}_n \mathbf{D} \mathbf{x}$ is more flatten than $\mathbf{x}$ in terms of variance. We would discuss its boundedness and asymptotically normal below. We would also explain the eigenvalue preservation.

6.1 Boundeding effect

The key point of bounding effect here is that matrix $\mathbf{D}$ introduces randomness, and normalized Hadamard matrix helps average unevenness since it is composed of comparable $\frac{1}{\sqrt{n}}$ and $-\frac{1}{\sqrt{n}}$ entries.

We have the intuition from the paper of Tropp [6], and write down our lemma here:

Lemma 5. *Let* $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ *be the original vector. Then we denote* $\mathbf{y} := \mathbf{H}_n\mathbf{D}\mathbf{x}$. *and rewrite it as* $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]^T$

for any $i = 1, 2, \ldots, n$,

$$\mathbf{y}_i = \sum_{j=1}^{n} h_{ij} d_j \mathbf{x}_j \tag{6.1}$$

$$\Pr\{|\mathbf{y}_i| \geq t\} \leq 2\exp\left(-\frac{nt^2}{2\|\mathbf{x}\|_2^2}\right) \tag{6.2}$$

So it is easy to see that for each given order n and a given vector $\mathbf{x}$, we have that each entry of $\mathbf{y}$ after Randomized Hadamard Transform (RHT) follows the distribution of sub-Gaussian, which serves as one of our main ideas in proving the main result of this article.

Now we take $t = \sqrt{\frac{\log(n)}{n}}\|\mathbf{x}\|_2$, and we have:

$$\Pr\left\{|\mathbf{y}_i| \geq \frac{\sqrt{\log(n)}}{\sqrt{n}}\|\mathbf{x}\|_2\right\} \leq 2\exp\left(-\frac{\log(n)}{2}\right) = 2n^{-\frac{1}{2}} \tag{6.3}$$

As $n \to \infty$,

$$\Pr\left\{|\mathbf{y}_i| \geq \frac{\sqrt{\log(n)}}{\sqrt{n}}\|\mathbf{x}\|_2\right\} \to 0 \tag{6.4}$$

This is the so called bounding effect.

6.2 Asymptotic normality

Now we will prove the asymptotically normality of RHT, finding out what kinds of situations that RHT would be asymptotically normal.

For convenient visualization, we calculate the mean and variance of $\mathbf{y}_i$ here for further study.

Because d_j follows a Rademacher distribution, d_j and d_k are independent for $j \neq k$.

$$\mathbf{y}_i = \sum_{j=1}^{n} h_{ij}\, d_j\, \mathbf{x}_j, \qquad d_j \overset{\text{i.i.d.}}{\sim} \text{Rad}(\pm 1),$$

$$\mathbb{E}[\mathbf{y}_i] = \sum_{j=1}^{n} h_{ij}\, \mathbf{x}_j\, \mathbb{E}[d_j] = 0,$$

$$\begin{aligned}
\text{Var}(\mathbf{y}_i) = \mathbb{E}\big[(\mathbf{y}_i - \mathbb{E}[\mathbf{y}_i])^2\big] &= \mathbb{E}\left[\left(\sum_{j=1}^{n} h_{ij}\, d_j\, \mathbf{x}_j\right)^2\right] \\
&= \sum_{j=1}^{n}\sum_{k=1}^{n} h_{ij}h_{ik}\, \mathbf{x}_j\mathbf{x}_k\, \mathbb{E}[d_j d_k] \\
&= \sum_{j=1}^{n} h_{ij}^2\, \mathbf{x}_j^2\, \mathbb{E}[d_j^2] + \sum_{j\neq k} h_{ij}h_{ik}\, \mathbf{x}_j\mathbf{x}_k\, \mathbb{E}[d_j]\, \mathbb{E}[d_k] \\
&= \sum_{j=1}^{n} h_{ij}^2\, \mathbf{x}_j^2 \quad (\text{because } \mathbb{E}[d_j] = 0,\ \mathbb{E}[d_j^2] = 1) \\
&= \frac{1}{n}\|\mathbf{x}\|_2^2.
\end{aligned}$$

Also for $\eta_j = h_{ij}\, d_j\, \mathbf{x}_j$, from the Appendix proof of Lemma 5 we have known that each η_j is strictly i.i.d. with $\mathbb{E}[\eta_j] = 0$ and $\text{Var}[\eta_j] = \mathbf{x}_j^2/n$.

As long as i.i.d. is satisfied for η_j, we could apply the central limit theorem (CLT) in the version of Lindeberg's condition to show that $\mathbf{y}_i$ is asymptotically normal.

Theorem 4 (Central Limit Theorem — Lindeberg form)**.** *Let* $X_1, X_2, \ldots$ *be independent random variables with* $\mathbb{E}[X_k] = \mu_k$ *and* $\text{Var}(X_k) = \sigma_k^2 < \infty$. *Define* $s_n^2 := \sum_{k=1}^{n} \sigma_k^2$. *If for every* $\varepsilon > 0$

$$\frac{1}{s_n^2}\sum_{k=1}^{n} \mathbb{E}\Big[(X_k - \mu_k)^2\, \mathbf{1}_{\{|X_k - \mu_k| > \varepsilon s_n\}}\Big] \xrightarrow[n\to\infty]{} 0,$$

then

$$Z_n := \frac{\sum_{k=1}^{n}(X_k - \mu_k)}{s_n} \xrightarrow{d} \mathcal{N}(0, 1).$$

Lemma 6. *Let* $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]^T$ *be the original vector. Then we denote* $\mathbf{y} := \mathbf{H}_n\mathbf{D}\mathbf{x}$. *and rewrite it as* $\mathbf{y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n]^T$

Under the central limit theorem of Lindeberg form taking on η_j *which is i.i.d., we have:*

$$\mathbf{y}_i \xrightarrow{d} \mathcal{N}\left(0, \frac{\|\mathbf{x}\|_2^2}{n}\right) \tag{6.5}$$

as long as

$$\lim_{n\to\infty}\left(\max_{1\leq j\leq n}\frac{\mathbf{x}_j^2}{\|\mathbf{x}\|_2^2}\right) = 0 \tag{6.6}$$

6.3 Eigenvalue preservation

It should be noted that the normalized Hadamard matrix is orthogonal, i.e. $\mathbf{H}_n^\top \mathbf{H}_n = \mathbf{I}_n$, and the Rademacher matrix is also orthogonal, i.e. $\mathbf{D}^\top \mathbf{D} = \mathbf{I}_n$.

This implies that:

Lemma 7. *If* $\tilde{\mathbf{X}} = \mathbf{H}_n \mathbf{D} \mathbf{X}$, *then*

$$\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} = \mathbf{X}^\top \mathbf{X}. \tag{6.7}$$

where $\tilde{\mathbf{X}} = \mathbf{H}_n \mathbf{D} \mathbf{X}$ and $\mathbf{X}$ is the original matrix.

Thus when you do eigenvalue decomposition on $\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$, you will get the same eigenvalues as $\mathbf{X}^\top \mathbf{X}$.

ACKNOWLEDGMENTS

Thank Prof. Zhixiang Zhang, my supervisor.

Thank Prof. Zhi Liu, my reference recommender.

Thank Prof. Lihu Xu, my reference recommender.

Thank all the dedicated researchers in the field of sketching.

I love you, my grandma, alive or dead.

APPENDIX A. LEMMA PROOF

A.1 proof of Lemma 1

Proof. Throughout the argument we fix the single noise $\varepsilon \sim \mathcal{N}(0, \mathbf{I}_n)$, so that $\mathrm{Cov}(\varepsilon) = \mathbf{I}_n$; the variance parameter σ^2 has been set to 1. For each data block we abbreviate

$$a_i := \mathrm{tr}\big[(\mathbf{X}_i^\top \mathbf{X}_i)^{-1}\big], \qquad i = 1, \dots, k,$$

and note that $a_i > 0$ because $\mathbf{X}_i^\top \mathbf{X}_i$ is positive definite.

The full-sample estimator. Writing $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$ and $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ gives

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

Since the noise is centred, $\mathbb{E}[\hat{\beta}] = \beta$ and the covariance of the estimator is

$$\mathrm{Cov}(\hat{\beta}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{I}_n \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} = (\mathbf{X}^\top \mathbf{X})^{-1}.$$

For any mean-zero random vector $\mathbf{Z}$ we have $\mathbb{E}\|\mathbf{Z}\|_2^2 = \mathrm{tr}(\mathrm{Cov}(\mathbf{Z}))$ because $\mathbb{E}[\mathbf{Z}^\top \mathbf{Z}] = \mathrm{tr}(\mathbb{E}[\mathbf{Z}\mathbf{Z}^\top])$. Applying this identity to $\mathbf{Z} = \hat{\beta} - \beta$ yields the global mean-squared error

$$\mathbb{E}\big\|\hat{\beta} - \beta\big\|_2^2 = \mathrm{tr}\big[(\mathbf{X}^\top \mathbf{X})^{-1}\big]. \tag{2.1}$$

Local estimators and their aggregation. On worker i the response satisfies $\mathbf{Y}_i = \mathbf{X}_i \beta + \varepsilon_i$ with $\varepsilon_i \overset{\text{indep}}{\sim} \mathcal{N}(0, \mathbf{I}_{n_i})$. The corresponding OLS estimate is

$$\hat{\beta}_i = (\mathbf{X}_i^\top \mathbf{X}_i)^{-1} \mathbf{X}_i^\top \mathbf{Y}_i = \beta + (\mathbf{X}_i^\top \mathbf{X}_i)^{-1} \mathbf{X}_i^\top \varepsilon_i,$$

an unbiased vector whose covariance equals $(\mathbf{X}_i^\top \mathbf{X}_i)^{-1}$. We combine the k local estimates using weights $w_1, \dots, w_k$ that satisfy $\sum_{i=1}^k w_i = 1$,

$$\hat{\beta}_{\mathrm{dist}}(w) := \sum_{i=1}^k w_i \hat{\beta}_i.$$

Independence of the noises entails

$$\mathrm{Cov}\big(\hat{\beta}_{\mathrm{dist}}(w)\big) = \sum_{i=1}^k w_i^2 (\mathbf{X}_i^\top \mathbf{X}_i)^{-1},$$

so the distributed mean-squared error is

$$\mathbb{E}\big\|\hat{\beta}_{\mathrm{dist}}(w) - \beta\big\|_2^2 = \sum_{i=1}^k w_i^2 a_i. \tag{2.2}$$

Choosing the weights. Introduce the auxiliary vectors $u_i = w_i \sqrt{a_i}$ and $v_i = 1/\sqrt{a_i}$. Because $\sum_i w_i = 1$,

$$1 = \Big(\sum_{i=1}^k u_i v_i\Big)^2 \le \Big(\sum_{i=1}^k u_i^2\Big)\Big(\sum_{i=1}^k v_i^2\Big) = \Big(\sum_{i=1}^k w_i^2 a_i\Big)\Big(\sum_{i=1}^k \frac{1}{a_i}\Big),$$

where the inequality is the classical Cauchy–Schwarz bound $|\langle u, v\rangle| \le \|u\|_2 \|v\|_2$. Consequently

$$\sum_{i=1}^k w_i^2 a_i \ge \frac{1}{\sum_{i=1}^k 1/a_i},$$

with equality precisely when the vectors u and v are linearly dependent, that is, when $w_i \sqrt{a_i} = \lambda / \sqrt{a_i}$ for some constant λ. Imposing $\sum_i w_i = 1$ leads to the optimal choice

$$w_i^\star = \frac{1/a_i}{\sum_{j=1}^k 1/a_j}, \qquad i = 1, \dots, k,$$

and the minimum attainable error

$$\mathbb{E}\big\|\hat{\beta}_{\mathrm{dist}}(w^\star) - \beta\big\|_2^2 = \frac{1}{\sum_{i=1}^k 1/a_i}.$$

Finite-sample efficiency. Define

$$E(\mathbf{X}_1, \dots, \mathbf{X}_k) := \frac{\mathbb{E}\big\|\hat{\beta} - \beta\big\|_2^2}{\mathbb{E}\big\|\hat{\beta}_{\mathrm{dist}}(w^\star) - \beta\big\|_2^2}.$$

Substituting the two risk expressions gives

$$E(\mathbf{X}_1,\dots,\mathbf{X}_k) := \frac{\mathbb{E}\big\|\hat{\beta}-\beta\big\|_2^2}{\mathbb{E}\big\|\hat{\beta}_{\text{dist}}(w^\star)-\beta\big\|_2^2} = \operatorname{tr}\big[(\mathbf{X}^\top\mathbf{X})^{-1}\big]\sum_{i=1}^{k}\frac{1}{\operatorname{tr}\big[(\mathbf{X}_i^\top\mathbf{X}_i)^{-1}\big]} \tag{2.3}$$

Thus all three relative efficiency formulas are proved for Lemma 1. □

A.2 Proof of Lemma 3

Proof. We first explore mean, covariance matrix and second moment of each row in our chosen Gaussian Mixture Model (GMM) distribution. Suppose that the j-th row of the original matrix $\mathbf{X}$ is $\mathbf{X}_{j,*}$, which is a p-dimensional vector.

1. **Mean**

$$\begin{aligned}\mathbb{E}[\mathbf{X}_{j,*}] &= a_1\,\mathbb{E}[\mathbf{X}_{j,*}\mid \mathbf{X}_{j,*}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})]\\ &\quad + a_2\,\mathbb{E}[\mathbf{X}_{j,*}\mid \mathbf{X}_{j,*}\sim\mathcal{N}(\boldsymbol{\mu}_2, c\boldsymbol{\Sigma})]\\ &= a_1\,\mathbf{0} + a_2\,\boldsymbol{\mu}_2\\ &= a_2\,\boldsymbol{\mu}_2.\end{aligned}$$

2. **Second moment**

$$\begin{aligned}\mathbb{E}[\mathbf{X}_{j,*}\mathbf{X}_{j,*}^\top] &= a_1\,\mathbb{E}\big[\mathbf{X}_{j,*}\mathbf{X}_{j,*}^\top\mid \mathbf{X}_{j,*}\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})\big]\\ &\quad + a_2\,\mathbb{E}\big[\mathbf{X}_{j,*}\mathbf{X}_{j,*}^\top\mid \mathbf{X}_{j,*}\sim\mathcal{N}(\boldsymbol{\mu}_2, c\boldsymbol{\Sigma})\big]\\ &= a_1\left(\boldsymbol{\Sigma}+\mathbf{0}\cdot\mathbf{0}^\top\right) + a_2\left(c\boldsymbol{\Sigma}+\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top\right)\\ &= (a_1+a_2c)\,\boldsymbol{\Sigma} + a_2\,\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top.\end{aligned}$$

3. **Covariance**

$$\begin{aligned}\operatorname{Cov}(\mathbf{X}_{j,*}) &= \mathbb{E}[\mathbf{X}_{j,*}\mathbf{X}_{j,*}^\top] - \mathbb{E}[\mathbf{X}_{j,*}]\,\mathbb{E}[\mathbf{X}_{j,*}]^\top\\ &= \left(a_1+a_2c\right)\boldsymbol{\Sigma} + a_2\,\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top - \left(a_2\,\boldsymbol{\mu}_2\right)\left(a_2\,\boldsymbol{\mu}_2\right)^\top\\ &= (a_1+a_2c)\,\boldsymbol{\Sigma} + a_1a_2\,\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top.\end{aligned}$$

We would found that all these variables are fixed for such GMM distribution, and we could use these fixed elements from each row to sum as the expectation of the sampled local gram matrix.

It is easy to see that:

$$\mathbb{E}\big[\mathbf{X}^\top\mathbf{X}\big] = \sum_{j=1}^{n}\mathbb{E}\big[\mathbf{X}_{j,*}\mathbf{X}_{j,*}^\top\big] = n\left((a_1+a_2c)\,\boldsymbol{\Sigma} + a_2\,\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top\right)$$

Also:

$$\mathbb{E}[\mathbf{X}_i^\top\mathbf{X}_i] = \sum_{\ell\in\mathbf{S}}\mathbb{E}[\mathbf{X}_{\ell,*}\mathbf{X}_{\ell,*}^\top] = \frac{n}{k}\left((a_1+a_2c)\boldsymbol{\Sigma} + a_2\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top\right)$$

Where $\mathbf{S}$ is the set of indices of rows assigned to machine i.

We then explore mean, covariance matrix and second moment of each row of the transformed matrix after RHT is applied to the $\mathbf{X}$ of the original GMM distribution.

For any fixed row index ℓ we can write

$$\tilde{\mathbf{X}}_{\ell,*} = \mathbf{e}_\ell^\top\tilde{\mathbf{X}} = \sum_{j=1}^{n} h_{\ell j}\,d_j\,\mathbf{X}_{j,*},$$

where $h_{\ell j}\in\{\pm n^{-1/2}\}$ are entries of the orthogonal Hadamard matrix $\mathbf{H}_n$, the scalars d_j are i.i.d. Rademacher ($\Pr\{d_j=\pm1\}=1/2$), and each original row $\mathbf{X}_{j,*}$ is drawn from the mixture $a_1\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}) + a_2\mathcal{N}(\boldsymbol{\mu}_2, c\boldsymbol{\Sigma})$.

1. *Mean*. Because $\mathbb{E}[d_j]=0$,

$$\mathbb{E}[\tilde{\mathbf{X}}_{\ell,*}] = \sum_{j=1}^{n} h_{\ell j}\,\mathbb{E}[d_j]\,\mathbb{E}[\mathbf{X}_{j,*}] = \mathbf{0}.$$

2. *Second moment*. Independence of the d_j variables implies $\mathbb{E}[d_jd_k]=\delta_{jk}$, leaving only the diagonal terms:

$$\begin{aligned}\mathbb{E}[\tilde{\mathbf{X}}_{\ell,*}\tilde{\mathbf{X}}_{\ell,*}^\top] &= \sum_{j=1}^{n} h_{\ell j}^2\,\mathbb{E}[\mathbf{X}_{j,*}\mathbf{X}_{j,*}^\top] \qquad \left(h_{\ell j}^2 = n^{-1}\right)\\ &= \tfrac{1}{n}\sum_{j=1}^{n}\mathbb{E}[\mathbf{X}_{j,*}\mathbf{X}_{j,*}^\top]\\ &= (a_1+a_2c)\,\boldsymbol{\Sigma} + a_2\,\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top.\end{aligned}$$

3. *Covariance*. The mean is zero, so the covariance equals the second moment:

$$\operatorname{Cov}(\tilde{\mathbf{X}}_{\ell,*}) = (a_1+a_2c)\,\boldsymbol{\Sigma} + a_2\,\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top.$$

Thus again for the expectation of the local gram matrix after RHT:

$$\mathbb{E}[\tilde{\mathbf{X}}_i^\top\tilde{\mathbf{X}}_i] = \sum_{\ell\in S}\mathbb{E}[\tilde{\mathbf{X}}_{\ell,*}\tilde{\mathbf{X}}_{\ell,*}^\top] = \frac{n}{k}\left[(a_1+a_2c)\boldsymbol{\Sigma} + a_2\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top\right]$$

We found that:

$$\mathbb{E}[\tilde{\mathbf{X}}_i^\top\tilde{\mathbf{X}}_i] = \mathbb{E}[{\mathbf{X}_i}^\top\mathbf{X}_i] = \frac{n}{k}\left[(a_1+a_2c)\boldsymbol{\Sigma} + a_2\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top\right] = \frac{1}{k}\mathbb{E}[\mathbf{X}^\top\mathbf{X}]$$

This means that the final result of Lemma 3 is proved. □

A.3 proof of Lemma 5

Proof. Now we introduce the Hoeffding's inequality here:

Theorem 5 (Hoeffding's Inequality)**.** *Let $X_1,\dots,X_n$ be independent random variables satisfying $a_i\le X_i\le b_i$ almost surely for $i=1,\dots,n$. Define*

$$S_n = \sum_{i=1}^{n} X_i.$$

Then for every $t > 0$,

$$\Pr\big(|S_n - \mathbb{E}[S_n]| \geq t\big) \;\leq\; 2\,\exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Now denote each term from Lemma 5 6.1 as $\eta_j = h_{ij}d_j\mathbf{x}_j$. And trivially we have $\mathbf{y}_i = \sum_{j=1}^{n}\eta_j$. Again since h_{ij} and $\mathbf{x}_j$ are fixed values in the matrix, we can conclude below that:

$$\eta_j \text{ are i.i.d.} \qquad \text{since } d_j \text{ are i.i.d.}$$

Now we find bound for η_j:

$$|\eta_j| = \frac{1}{\sqrt{n}}|\mathbf{x}_j| \quad \Rightarrow \quad \frac{-1}{\sqrt{n}}|\mathbf{x}_j| \leq \eta_j \leq \frac{1}{\sqrt{n}}|\mathbf{x}_j|$$

Then we denote here $a_j = \frac{-1}{\sqrt{n}}|\mathbf{x}_j|$, and $b_j = \frac{1}{\sqrt{n}}|\mathbf{x}_j|$, and we have $a_j \leq \eta_j \leq b_j$.

For the sum of η_j we denote: $S_n = \mathbf{y}_i = \sum_{j=1}^{n}\eta_j$, with before we have $\mathbb{E}[S_n] = 0$.

Finally, plug in all these variables into the Hoeffding's inequality, we get the result of Lemma 5's 6.2.

$$\Pr\{|\mathbf{y}_i - 0| \geq t\} \leq 2\exp\left(-\frac{2t^2}{\sum_{j=1}^{n}\left(\frac{2}{\sqrt{n}}|\mathbf{x}_j|\right)^2}\right)$$

$$\Longrightarrow \; Lemma\ 5's 6.2$$

Hence we have proved the result of Lemma 5's 6.2 here.

□

A.4 proof of Lemma 6

Proof. We conclude our results here to fit in Theorem 4 of Lindeberg's CLT. Let $\eta_j, j = 1, 2, \ldots, n$ be i.i.d. random variables with $\mathbb{E}[\eta_j] = 0$ and $\mathrm{Var}[\eta_j] = \mathbf{x}_j^2/n$. Define $s_n^2 = \sum_{j=1}^{n}\mathrm{Var}[\eta_j] = \|\mathbf{x}\|_2^2/n$. Then, we have:

$$\begin{aligned}
&\frac{1}{s_n^2}\sum_{j=1}^{n}\mathbb{E}\big[(\eta_j - \mathbb{E}[\eta_j])^2\,\mathbf{1}_{\{|\eta_j - \mathbb{E}[\eta_j]| > \varepsilon s_n\}}\big] \\
&= \frac{1}{\|\mathbf{x}\|_2^2/n}\sum_{j=1}^{n}\mathbb{E}\Big[(\eta_j)^2\,\mathbf{1}_{\{|\eta_j| > \varepsilon\|\mathbf{x}\|_2/\sqrt{n}\}}\Big] \\
&= \frac{n}{\|\mathbf{x}\|_2^2}\sum_{j=1}^{n}\mathbb{E}\Big[(\eta_j)^2\,\mathbf{1}_{\{|\eta_j| > \varepsilon\|\mathbf{x}\|_2/\sqrt{n}\}}\Big] \\
&= \frac{n}{\|\mathbf{x}\|_2^2}\sum_{j=1}^{n}\mathbb{E}\big[(\eta_j)^2\,\mathbf{1}_{\{|\mathbf{x}_j| > \varepsilon\|\mathbf{x}\|_2\}}\big] \\
&= \frac{1}{\|\mathbf{x}\|_2^2}\sum_{j=1}^{n}\mathbf{x}_j^2\,\mathbb{E}\big[\mathbf{1}_{\{|\mathbf{x}_j| > \varepsilon\|\mathbf{x}\|_2\}}\big] \\
&= \sum_{i=k}^{n}\frac{\mathbf{x}_{(i)}^2}{\|\mathbf{x}\|_2^2} \\
&\text{if } \frac{\mathbf{x}_{(k-1)}}{\|\mathbf{x}\|_2} \leq \varepsilon < \frac{\mathbf{x}_{(k)}}{\|\mathbf{x}\|_2} \text{ for any } \varepsilon > 0 \\
&\text{where } \mathbf{x}_{(1)}, \mathbf{x}_{(2)}, \ldots, \mathbf{x}_{(n)} \text{ are the ordered values of } \mathbf{x}_j \\
&\text{and } k = 1, 2, \ldots, n.
\end{aligned}$$

Our condition is that the above limit goes to 0 as $n \to \infty$. Then we have:

$$\lim_{n\to\infty}\left(\frac{\mathbf{x}_{(n)}^2}{\|\mathbf{x}\|_2^2}\right) = 0$$

This implies Lemma 6 (6.6) is satisfied.

□

REFERENCES

[1] Yeshwanth Cherapanamjeri and Jelani Nelson. *Uniform Approximations for Randomized Hadamard Transforms with Applications*. 2022. arXiv: 2203.01599 [cs.LG]. URL: https://arxiv.org/abs/2203.01599.

[2] Michał Dereziński. *Algorithmic Gaussianization through Sketching: Converting Data into Sub-gaussian Random Designs*. 2023. arXiv: 2206.10291 [cs.LG]. URL: https://arxiv.org/abs/2206.10291.

[3] Edgar Dobriban and Sifan Liu. *Asymptotics for Sketching in Least Squares Regression*. 2019. arXiv: 1810.06089 [math.ST]. URL: https://arxiv.org/abs/1810.06089.

[4] Edgar Dobriban and Yue Sheng. *Distributed linear regression by averaging*. 2022. arXiv: 1810.00412 [math.ST]. URL: https://arxiv.org/abs/1810.00412.

[5] Harvineet Singh, Christopher Musco, and Rumi Chunara. *Active Linear Regression in the Online Setting via Leverage Score Sampling*. https://realworldml.github.io/files/cr/paper55.pdf. 2023.

[6] Joel A. Tropp. *Improved analysis of the subsampled randomized Hadamard transform*. 2011. arXiv: 1011.1595 [math.NA]. URL: https://arxiv.org/abs/1011.1595.

[7] Joel A. Tropp. "User-Friendly Tail Bounds for Sums of Random Matrices". In: *Foundations of Computational Mathematics* 12.4 (Aug. 2011), pp. 389–434. ISSN: 1615-3383. DOI: 10.1007/s10208-011-9099-z. URL: http://dx.doi.org/10.1007/s10208-011-9099-z.

[8] Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018.

[9] Shusen Wang, Alex Gittens, and Michael W. Mahoney. *Sketched Ridge Regression: Optimization Perspective, Statistical Perspective, and Model Averaging*. 2018. arXiv: 1702.04837 [stat.ML]. URL: https://arxiv.org/abs/1702.04837.

B. EXPERIENTAL RESULTS

In this section, we will show the comparisons of Randomized Hadamard Transform (RHT) Sampling partitions vs Uniform Sampling partitions in terms of the performance of relative efficiency of the ratio of the global mean-squared error (MSE) of the full-sample estimator to the mean-squared error (MSE) of the distributed estimator with equal-weighted partitions.

The first graph's setting is that $c = 10$ for fixed inflated ratio and $\boldsymbol{\mu}_2 = (10, 10, \ldots, 10)^\top \in \mathbb{R}^{1000\times 1}$ for fixed mean vector of second inflated Cluster. Only the a_2 composition proportion is changing from 0.1 to 0.9.

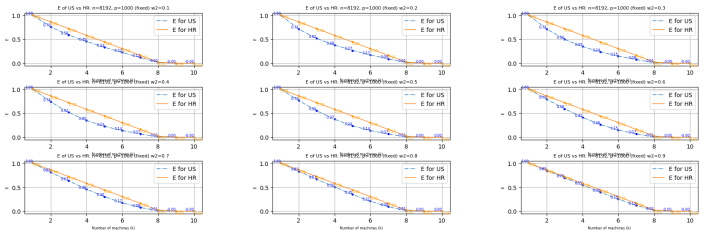

Figure 1: Comparison of Relative Efficiency between RHT Sampling and Uniform Sampling with varying a_2 proportion. Settings: $c = 10$, $\boldsymbol{\mu}_2 = (10, \ldots, 10)^\top$, $p = 1000$.

We could see that there is a regression of the gap of two relative efficiencies when a_2 is increasing, it is interesting to see that the gap reaches the maximum when a_2 is between 0.3 and 0.4, and then dcrease to 0 when a_2 is approaching 1.

The second graph's setting is that $a_2 = 0.2$ for fixed composition proportion and $\boldsymbol{\mu}_2 = (5, 5, \ldots, 5)^\top \in \mathbb{R}^{1000\times 1}$ for fixed mean vector of second inflated Cluster. Only the c inflated ratio is changing from 1 to 500.

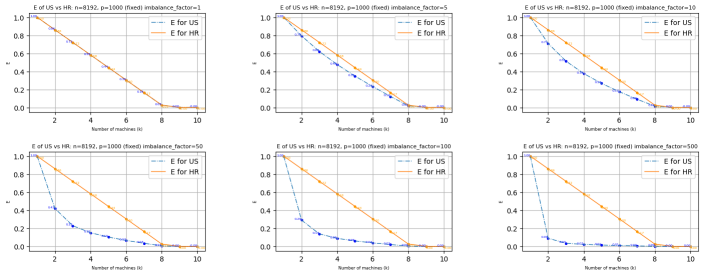

Figure 2: Comparison of Relative Efficiency between RHT Sampling and Uniform Sampling with varying c inflated ratio. Settings: $a_2 = 0.2$, $\boldsymbol{\mu}_2 = (5, \ldots, 5)^\top$, $p = 1000$.

We could see that the gap increases drastically when c is increasing drastically, corresponding to our theoretical results before.

The third graph's setting is that $a_2 = 0.2$ for fixed composition proportion and $c = 10$ for fixed inflated ratio. Only the $\boldsymbol{\mu}_2$ mean vector is changing from $(1, 1, \ldots, 1)^\top$ to $(1000, 1000, \ldots, 1000)^\top$.

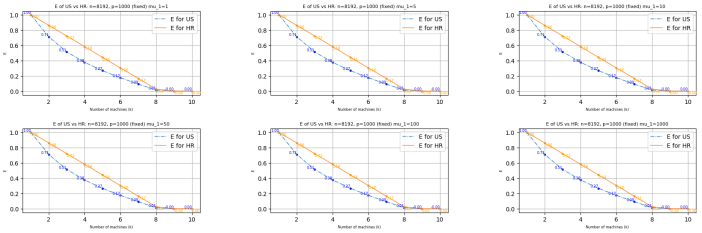

Figure 3: Comparison of Relative Efficiency between RHT Sampling and Uniform Sampling with varying $\boldsymbol{\mu}_2$ mean vector. Settings: $a_2 = 0.2$, $c = 10$, $p = 1000$.

We could see that the gap remains exactly the same when $\boldsymbol{\mu}_2$ is varying, which means the direction of the mean vector does not affect the relative efficiency of the RHT Sampling vs Uniform Sampling. This is an interesting result and would be our future discussion for further research.

C. DISCUSSION

We would further focus on the proof of similar main result of Corollary 1 in the future work, and consider the case that p is comparable to n in the case that $p = c_1 n$ for any constant c_1. This is the case p could also tends to infinity when n is tending to infinity.

Received May 2025

Zhixiang Zhang. Faculty of Science and Technology, University of Macau, Macau.
E-mail address: zhixzhang@um.edu.mo

Yishu Yang. Faculty of Science and Technology, University of Macau, Macau.
E-mail address: dc12828@um.edu.mo