smash@frametitle#1
@smash@frametitle

Distributed Sketching in Machine Learning: Relative Efficiency and Randomized Hadamard Transform

From "text.txt"

May 12, 2025

Table of Contents

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) or focus on the accurate expectation of approximation error for the OLS (ordinary least square) estimation (Derezinski, 2023).
- 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).

Introduction (Continued)

- 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.
- 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 the weight accordingly from the result of Eq. (12).
- 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 satisfies $\lim_{n\to\infty} \frac{p}{n} = 0$, which

Introduction (Continued)

- We could also intepretate this idea differently, that is the difference of $\text{tr}[(\tilde{\mathbf{X}}_i^\top \tilde{\mathbf{X}}_i)^{-1}]$ and $\text{tr}[(\frac{1}{K}\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1}]$ converges to 0 at the rate of $\mathcal{O}(\sqrt{p^3 \frac{\log n}{n^3}})$ (Eq. (7)) as n tends to be approaching infinity for p satisfying $\lim_{n\to\infty} \frac{p}{n} = 0$, 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 and the paper of Cherapanamjeri.
- And we utilize many probability inequalities from the book of Vershynin, the papers of Tropp, mainly the Bernstein Inequality of sub-exponential tail bounds and Matrix Chernoff Inequality to prove the result of Corollary 1.

Main Result: Randomized Hadamard Transform

- Our main contribution here is that under the research target of Dobriban and Sheng 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)

Main Result: Randomized Hadamard Transform

- Our main contribution here is that under the research target of Dobriban and Sheng 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)

Normalized Hadamard Matrix: 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.

Main Result (Continued)

- 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 (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)

Main Result (Continued)

- 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 (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)

Linear Regression Setting:

- General linear regression: $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I}_n)$.
- Quality measured by expected mean square error (MSE).
- $\beta \in \mathbb{R}^p$: true, unknown parameter vector.
- $\mathbf{X} \in \mathbb{R}^{n \times p}$: predictor variables.
- $\mathbf{Y} \in \mathbb{R}^n$: response variables.
- $\epsilon \in \mathbb{R}^n$: noise term, $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$ (often $\sigma^2 = 1$).

Main Result (Continued) - Distributed Sketching

- OLS estimation can be computationally burdensome for large n.
- Distributed sketching: allocate sub training sets to local machines, train on each, then aggregate estimators.
- Our focus: partitioned machines for regression.
- Averaging estimator: $\hat{\beta}_{dist} = \sum_{i=1}^{K} w_i \hat{\beta}_i$, with weights w_i for machine i.
- MSE for distributed estimator: $M(\hat{\beta}_{dist}) = \mathbb{E}||\beta - \hat{\beta}_{dist}||^2$.
- Relative efficiency (Lemma 1): $E(\mathbf{X}_1, \ldots, \mathbf{X}_K) = \frac{M(\hat{\beta})}{M(\hat{\beta}_{dist})}$.
- Discussion points:
  - Why uniform partition?
  - RHT before uniform sampling $\implies$ no need to adjust w_i, just use $\frac{1}{K}$ (Lemma 2).
  - What if no RHT? What $E(\mathbf{I}_p, \mathbf{X}_1, ..., \mathbf{X}_K)$ can be achieved?
- Note: This paper discusses high dimension, where $n \to \infty$ and p satisfies $\lim_{n\to\infty} \frac{p}{n} = 0$. ($p$ can also $\to \infty$ but slower than n).

Corollary 1

Corollary (Main Result)

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 satisties* $\lim_{n \to \infty} \frac{p}{n} = 0$.

$\tilde{\mathbf{X}} = \mathbf{H}_n \mathbf{D} \mathbf{X}$, *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 Eq. (3) *result of Lemma 1, we have the following corollary:*

$$\lim \ \mathbb{E}(\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_K) = 1$$

Our Focus on Proving Better Relative Efficiency

We first give the finite sample results of Dobriban and Sheng from section 3.2.

Lemma (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

The Reason for Desired Equal Partition for RHT to Function Well

From Eq. (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

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

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

The Reason for Desired Equal Partition (Continued)

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

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

- ★ Lemma 2 here majorly serves as a new perspective because from Lemma 2 Eq. (12), 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 Eq. (12) and then we find ways to achieve the dream efficiency of 1 by realizing Lemma 2 Eq. (12).
- ★ This is why we introduce RHT in this paper and prove it as our main

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

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*

The Intuition Behind GMM Distribution

One of our main focus here on the finite result of relative efficiency from Dobriban and Sheng (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.

- The chosen GMM distribution is composed of a mixture of two independent Multivariate Gaussian Distributions.
- $100a_1\%$ proportion of data rows sampled from first Multivariate Gaussian Distribution: mean $0 \in \mathbb{R}^p$, covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$.
- $100a_2\%$ proportion of data rows sampled from second Multivariate Gaussian Distribution: mean $\mu_2 \in \mathbb{R}^p$, covariance matrix $c\Sigma \in \mathbb{R}^{p \times p}$.
- $0 \in \mathbb{R}^p$: scalar value of each index in the vector is 0.

The Chosen Gaussian Mixture Model (GMM) Distribution

Definition

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

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

$$f_2(x) = \frac{1}{(2\pi)^{p/2} |c\Sigma|^{1/2}} \exp\Big\{-\tfrac{1}{2} (x-\mu_2)^\top (c\Sigma)^{-1} (x-\mu_2)\Big\},$$

The mixture parameters satisfy

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

Lemma 3

Lemma

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:*

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

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

Proof of Corollary 1 (MAIN RESULT)

Proof

To prove this corollary rigorously, we need to show that as $n \to \infty$ *and* p *satisfies* $\lim_{n\to\infty} \frac{p}{n} = 0$, *we have:*

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

where $\tilde{\mathbf{X}} \in \mathbb{R}^{n\times p}$ *is a matrix with growing* n *and* p *satisfies* $\lim_{n\to\infty} \frac{p}{n} = 0$. *and* $\tilde{\mathbf{X}}_i \in \mathbb{R}^{n_i \times p}$ *with* $n_i = \frac{n}{K}$. *By inverse decomposition* $A^{-1} - B^{-1} = B^{-1}(B - A)A^{-1}$ *for* $A, B \in \mathbb{R}^{p\times p}$. *Using* $|\operatorname{tr}(M)| \le \sqrt{p}\|M\|_F$*:*

$$\begin{aligned}
\left|\operatorname{tr}(A^{-1}) - \operatorname{tr}(B^{-1})\right| &= \left|\operatorname{tr}\left(A^{-1}(A - B)B^{-1}\right)\right| \\
&\le \sqrt{p}\left\| A^{-1}(A - B)B^{-1}\right\|_F \\
&\le \sqrt{p}\|A^{-1}\|_2\|A - B\|_F\|B^{-1}\|_2
\end{aligned}$$

The last line uses $\|PQ\|_F \le \|P\|_2\|Q\|_F$ *and* $\|PQ\|_F \le \|Q\|_2\|P\|_F$

Proof of Corollary 1: Coordinate-wise Concentration

Proof

(Continued)

Focus on an element (j,k) *of the local Gram matrix* A*. Let* $\mathbf{Y}_r := \left(\tilde{\mathbf{X}}\right)_{r,*} \quad (r = 1, \ldots, n)$*. The entry of* $\mathbf{A}$ *is* $\mathbf{A}_{jk} = \sum_{r \in P_i} (\mathbf{Y}_r)_j (\mathbf{Y}_r)_k$*, where* P_i *is the set of row indices for machine* i*. The global gram matrix scaled by* $\frac{1}{K}$ *is* $\mathbf{B} = \frac{1}{K}\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}}$ *with entry* $\mathbf{B}_{jk} = \frac{1}{K}\sum_{r=1}^{n} (\mathbf{Y}_r)_j (\mathbf{Y}_r)_k$*. Indicator:* $\mathbf{I}_r := \mathbf{1}_{\left\{\text{row } r \text{ is assigned to machine } i\right\}}$*, so* $\sum_{r=1}^{n} \mathbf{I}_r = n/K$*. Then* $\mathbf{A}_{jk} = \sum_{r=1}^{n} \mathbf{I}_r (\mathbf{Y}_r)_j (\mathbf{Y}_r)_k$*. Difference:* $\mathbf{A}_{jk} - \mathbf{B}_{jk} = \sum_{r=1}^{n} \left(\mathbf{I}_r - \frac{1}{K}\right)(\mathbf{Y}_r)_j (\mathbf{Y}_r)_k$*. Easy to see:* $\mathbb{E}[\mathbf{I}_r] = \frac{1}{K}$ *(uniform sampling independence). This is a sum of independent mean-zero random variables conditioned on* $\tilde{\mathbf{X}}$*. Denote* $\mathbf{Z}_r := \left(\mathbf{I}_r - \frac{1}{K}\right)(\mathbf{Y}_r)_j (\mathbf{Y}_r)_k$ *(independent zero-mean).*

Proof of Corollary 1: Bernstein Tail Bound

Proof

(Continued)

Theorem (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 $t \geq 0$

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

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

Goal: Derive sub-exponential Bernstein bound for $\mathbf{\Delta}_{jk}^{(i)} := \mathbf{A}_{jk} - \mathbf{B}_{jk}$.

Rewrite: $\mathbf{\Delta}_{jk}^{(i)} = \sum_{r=1}^n \left(\mathbf{I}_r - \frac{1}{K}\right) (\mathbf{Y}_r)_j (\mathbf{Y}_r)_k = \sum_{r=1}^n \left(\mathbf{I}_r - \frac{1}{K}\right) \mathbf{z}_r$, where $\mathbf{z}_r = (\mathbf{Y}_r)_j (\mathbf{Y}_r)_k$ and $\mathbf{I}_r \sim$ Bernoulli$(\frac{1}{K})$ independent. Conditioning is on $\mathbf{D}$ (random Rademacher matrix), as $\mathbf{H}$ and $\mathbf{X}$ (sampled from GMM) are

Proof of Corollary 1: Sub-Gaussian Properties

Proof

(Continued)

Lemma (Gaussian Distribution is sub-Gaussian (Vershynin 2.5.8a))

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: $\|Z_1\|_{\psi_2} \leq \gamma\sqrt{\mathbf{\Sigma}_{jj}}$, $\|Z_2 - \mu_2\|_{\psi_2} \leq \omega\sqrt{c\mathbf{\Sigma}_{jj}}$. This implies $\|Z_2\|_{\psi_2} \leq \omega\sqrt{c\mathbf{\Sigma}_{jj}} + \|\mu_2\|_{\psi_2}$. Define $\kappa = \max\{\gamma\sqrt{\mathbf{\Sigma}_{jj}}, \omega\sqrt{c\mathbf{\Sigma}_{jj}} + \|\mu_2\|_{\psi_2}\}$, so $\|\mathbf{X}_{rj}\|_{\psi_2} \leq \kappa$.
$(\mathbf{Y}_r)_j = \sum_{k=1}^{n} h_{rk} d_k \mathbf{X}_{kj}$. Let $\eta_k := h_{rk} d_k \mathbf{X}_{kj}$. Each η_k is independent sub-Gaussian, mean zero, sub-gaussian norm bound $\frac{\kappa}{\sqrt{n}}$ (assuming $h_{rk} = \pm 1/\sqrt{n}$). The text says κ/n which might be a typo if h_{ri} are entries of normalized Hadamard. Let's assume normalized Hadamard was used. So norm is $\frac{\kappa}{\sqrt{n}}$.

Proof of Corollary 1: L2 Norm of Inverse Gram Matrix (Matrix Chernoff)

Proof

(Continued)

Theorem (Matrix Chernoff (Tropp, Thm 2.2))

Let $\{\mathbf{X}_l\}$ *be a finite sequence of independent, random, self-adjoint* $p \times p$ *matrices satisfying* $\mathbf{X}_l \succeq 0$ *and* $\lambda_{\max}(\mathbf{X}_l) \leq R$ *a.s. Define* $\mu_{\min} = \lambda_{\min}\big(\sum_l \mathbb{E}[\mathbf{X}_l]\big), \mu_{\max} = \lambda_{\max}\big(\sum_l \mathbb{E}[\mathbf{X}_l]\big)$. *Then for* $\epsilon \in [0,1]$,

$$\Pr\Big\{\lambda_{\min}\big(\textstyle\sum_l \mathbf{X}_l\big) \leq (1-\epsilon)\,\mu_{\min}\Big\} \;\leq\; p\left[\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right]^{\mu_{\min}/R}, \text{ and for } \epsilon \geq 0,$$

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

Corollary (Simplified lower-tail bound)

Proof of Corollary 1: Applying Matrix Chernoff

Proof

(Continued)

Denote $\mathbf{W}_r = \mathbf{Y}_r \mathbf{Y}_r^T \succeq 0$. $\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} = \sum_{r=1}^n \mathbf{W}_r$. *Need* $\lambda_{\max}(\mathbf{W}_r) \leq R$ *(a.s. or asymptotically). Equivalent to* $\|\mathbf{Y}_r\|_2^2 \leq R$. *From before,* $\|(\mathbf{Y}_r)_j\|_{\psi_2}^2 \leq C\,\kappa$. *By Vershynin Lemma 2.7.7:* $\|(\mathbf{Y}_r)_j{}^2\|_{\psi_1} \leq \|(\mathbf{Y}_r)_j\|_{\psi_2}^2 \leq C\,\kappa = b$. *So* $\max_j \|(\mathbf{Y}_r)_j{}^2\|_{\psi_1} \leq b$. *Let* $\mathbf{R}_r := \|\mathbf{Y}_r\|_2^2 = \sum_{j=1}^p (\mathbf{Y}_r)_j{}^2$. *Sum of p sub-exponential terms.* $\sum_{j=1}^p \|(\mathbf{Y}_r)_j{}^2\|_{\psi_1}^2 \leq pb^2$. *Using Bernstein (Thm 1) for* $\mathbf{R}_r - \mathbb{E}[\mathbf{R}_r]$ *(need* $\mathbb{E}[\mathbf{R}_r]$ *first). Text takes* $t = C_0 \log n$. *(Assuming p is small compared to n or fixed).* $\Pr\{\max_{1\leq r\leq n} \mathbf{R}_r > C_0 \log n\} \leq 2\,n^{1-\frac{c'\,C_0}{b}}$ *(This should be* $2pn^{\cdots}$ *or* $2n \cdot n^{\cdots}$ *by union bound over r). The text has* $2n^{1-cC_0/b}$.

$$\Pr\{\lambda_{\max}(\mathbf{W}_r) > C_0 \log n\} \leq 2\,n^{1-\frac{c'\,C_0}{b}}, \quad C_0 \geq \frac{2b}{c'}$$

So $R = C_0 \log n$. $\mathbb{E}[\mathbf{W}_r] = \mathbb{E}[\mathbf{Y}_r \mathbf{Y}_r^T] = \left((a_1 + a_2 c)\,\boldsymbol{\Sigma} + a_2\,\boldsymbol{\mu}_2 \boldsymbol{\mu}_2^\top\right) := \boldsymbol{\Sigma}^\star$.

Proof of Corollary 1: Final Result

Proof

(Continued)

Combining bounds from Eq. (4), (6), (6) *via union law: With probability at least* $1 - 2p\exp\left(-\frac{\iota n}{\log n}\right) - 4\,n^{1-\frac{c'\,C_0}{b}} - 2p^2\,n^{-\frac{c'\,\chi^2}{b^2}}$ *:*

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

Using Eq. (1)*:*

$|\operatorname{tr}(\mathbf{A}^{-1}) - \operatorname{tr}(\mathbf{B}^{-1})| \leq \sqrt{p} \cdot \frac{K^2\chi p}{(1-\sigma)^2(\lambda^\star)^2}\sqrt{\frac{\log n}{n^3}} = \frac{K^2\chi p\sqrt{p}}{(1-\sigma)^2(\lambda^\star)^2}\sqrt{\frac{\log n}{n^3}}$ *Thus,*

$$\Pr\left\{|\operatorname{tr}(\mathbf{A}^{-1}) - \operatorname{tr}(\mathbf{B}^{-1})| \leq \frac{K^2\chi p\sqrt{p}}{(1-\sigma)^2(\lambda^\star)^2}\sqrt{\frac{\log n}{n^3}}\right\}$$
$$\geq 1 - 2p\exp\left(-\frac{\iota n}{\log n}\right) - 4\,n^{1-\frac{c'\,C_0}{b}} - 2p^2\,n^{-\frac{c'\,\chi^2}{b^2}}$$

Uniform Sampling: Energy Estimation

Consider $\mathbf{x} \in \mathbb{R}^{n\times 1}$, a column vector. "Energy": $E(\mathbf{x}) = \|\mathbf{x}\|_2^2 = \sum_{i=1}^n \mathbf{x}_i^2$. Sample set $\mathbf{S}$ of indices, $|\mathbf{S}| = \ell$. Sum of squared entries in sample: $\mathbf{Sum} = \sum_{j\in\mathbf{S}} \mathbf{x}_j^2$. Estimator of energy: $\hat{E}(\mathbf{x}) = \frac{n}{\ell} \sum_{j\in\mathbf{S}} \mathbf{x}_j^2$. This can be biased if entries $\mathbf{x}_j$ are not comparable in magnitude (missing large entries is harmful).

Uniform Sampling: Energy Estimation

Consider $\mathbf{x} \in \mathbb{R}^{n\times 1}$, a column vector. "Energy": $E(\mathbf{x}) = \|\mathbf{x}\|_2^2 = \sum_{i=1}^n \mathbf{x}_i^2$. Sample set $\mathbf{S}$ of indices, $|\mathbf{S}| = \ell$. Sum of squared entries in sample: $\mathbf{Sum} = \sum_{j\in\mathbf{S}} \mathbf{x}_j^2$. Estimator of energy: $\hat{E}(\mathbf{x}) = \frac{n}{\ell} \sum_{j\in\mathbf{S}} \mathbf{x}_j^2$. This can be biased if entries $\mathbf{x}_j$ are not comparable in magnitude (missing large entries is harmful). **Expectation of the estimator**

$\pi_j = \Pr(j \in \mathbf{S}) = \frac{\binom{n-1}{\ell-1}}{\binom{n}{\ell}} = \frac{\ell}{n}$. Indicator $\delta_j = 1$ if $j \in \mathbf{S}$, 0 otherwise.

$\Pr(\delta_j = 1) = \frac{\ell}{n}$. $M(\mathbf{x}, \mathbf{S}) = \frac{n}{\ell} \sum_{j=1}^n \delta_j \mathbf{x}_j^2$.

$\mathbb{E}[M(\mathbf{x}, \mathbf{S})] = \frac{n}{\ell} \sum_{j=1}^n \mathbb{E}[\delta_j]\mathbf{x}_j^2 = \frac{n}{\ell} \sum_{j=1}^n \frac{\ell}{n}\mathbf{x}_j^2 = \sum_{j=1}^n \mathbf{x}_j^2 = E(\mathbf{x})$. Uniform sampling is unbiased for energy estimation. However, unbiased doesn't mean it's a good estimator.

Uniform Sampling: Failure Under Two-Cluster GMM

From GMM (Definition 1):

- Cluster 1: $\mathbf{X}_{j,*} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ (prob a_1)
- Cluster 2: $\mathbf{X}_{j,*} \sim \mathcal{N}_p(\boldsymbol{\mu_2}, c\mathbf{\Sigma})$ (prob a_2), $c > 1$.

Cluster 2 rows have higher variance $\implies$ larger norms (inflated by c).

These are high leverage-score observations (outliers). Leverage score: $h_j = \mathbf{X}_{j,*}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_{j,*}$. Large $\|\mathbf{X}_{j,*}\|_2$ or sparse direction $\implies$ large h_j.

Uniform Sampling: Failure Under Two-Cluster GMM

From GMM (Definition 1):

- Cluster 1: $\mathbf{X}_{j,*} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$ (prob a_1)
- Cluster 2: $\mathbf{X}_{j,*} \sim \mathcal{N}_p(\boldsymbol{\mu_2}, c\mathbf{\Sigma})$ (prob a_2), $c > 1$.

Cluster 2 rows have higher variance $\implies$ larger norms (inflated by c). These are high leverage-score observations (outliers). Leverage score: $h_j = \mathbf{X}_{j,*}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}_{j,*}$. Large $\|\mathbf{X}_{j,*}\|_2$ or sparse direction $\implies$ large h_j.

Uneven Distribution of High-Leverage Rows

Equally partitioned uniform sampling ($n_i = n/K$) can lead to imbalanced partitions: some machines get many Cluster 2 rows, others few. This biases $\mathrm{tr}[(\mathbf{X}_i^\top \mathbf{X}_i)^{-1}]$. m_i: number of Cluster 2 rows in machine i. $m_i \sim \mathrm{Binomial}(n/K, a_2)$ for large n. $\mathbb{E}[m_i] = \frac{n}{K} a_2$. Fluctuations are expected. Chernoff/Hoeffding bound on deviation of m_i: For $\delta_n \in (0, 1)$, $\Pr\left\{\left|m_i - a_2 \frac{n}{K}\right| \geq \delta_n a_2 \frac{n}{K}\right\} \leq 2 \exp\left(-\frac{\delta_n^2 a_2 \frac{n}{K}}{3}\right)$. For all K machines: $\Pr\left\{\max_i \left|m_i - a_2 \frac{n}{K}\right| \leq \delta_n a_2 \frac{n}{K}\right\} \geq 1 - 2K \exp\left(-\frac{\delta_n^2 a_2 \frac{n}{K}}{3}\right)$. Take $\delta_n = \sqrt{\frac{K \log n}{a_2 n}}$: $\Pr\left\{\max_i \left|m_i - a_2 \frac{n}{K}\right| \leq \sqrt{a_2 \frac{n}{K} \log n}\right\} \geq 1 - \frac{2K}{n^{\frac{1}{3}}}$. Deviation

Uniform Sampling: Deviation of Local Gram Matrix

$\mathbf{M}_i = \mathbf{X_i}^\top \mathbf{X_i}$. From Lemma 3: $\mathbb{E}[\mathbf{M_i}] = \frac{n}{K}\left((a_1 + a_2 c)\,\boldsymbol{\Sigma} + a_2\,\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top\right)$. Ideally, $\mathbf{M}_i$ should be close to $\mathbb{E}[\mathbf{M_i}]$. Uneven distribution of Cluster 2 rows $\implies$ $\mathbf{M}_i$ can differ greatly. Decompose $\mathbf{M}_i = \mathbf{M}_i^{(1)} + \mathbf{M}_i^{(2)}$ (contributions from Cluster 1 and 2). High variance of m_i $\implies$ $\mathbf{M}_i^{(2)}$ has much variance across i.

- If $m_i \approx 0$ (e.g., $\frac{n}{K}a_2 - \sqrt{a_2 \frac{n}{K} \log n} \leq 0$): $\mathbf{M}_i \approx \mathbf{M}_i^{(1)} \sim \mathcal{W}_p(\frac{n}{K}, \boldsymbol{\Sigma})$. Smaller scale in $\boldsymbol{\mu}_2$ direction. Partition "misses" Cluster 2.
- If $m_j \gg \frac{n}{K}a_2$: $\mathbf{M}_j^{(2)} = m_j \boldsymbol{\mu}_2 \boldsymbol{\mu}_2^T + \sum_{r=1}^{m_j} \boldsymbol{\epsilon}_r\,\boldsymbol{\epsilon}_r^T + \text{cross terms}$, where $\boldsymbol{\epsilon}_r \sim \mathcal{N}(\mathbf{0}, c\boldsymbol{\Sigma})$. $\mathbf{M}_j$ has larger variance along $\boldsymbol{\mu}_2$ direction.

Matrix Bernstein bound (Tropp, Thm 1.6):

$\Pr\left\{\|\mathbf{M}_i - \mathbb{E}[\mathbf{M}_i]\| \leq \epsilon \frac{n}{K}\right\} \geq 1 - 2p \exp\left(-\frac{3\frac{n}{K}\epsilon^2}{6\sigma_{\max}^2 + 2\epsilon R}\right)$, where $\sigma_{\max}^2 \sim c\lambda_{\max}(\boldsymbol{\Sigma}) + |\boldsymbol{\mu}_2|^2$, $R \sim C \log(\frac{n}{K})$ (max sq row norm). For fixed ϵ, deviation bound $\epsilon\frac{n}{K}$ can be large.

Merits of Randomized Hadamard Transform (RHT)

RHT can flatten vector variance / redistribute energy. Normalized Hadamard: $\mathbf{H}_n = \frac{1}{\sqrt{n}}\mathbf{H}$, entries $h_{ij} = \pm\frac{1}{\sqrt{n}}$. Orthogonal: $\mathbf{H}_n^\top \mathbf{H}_n = \mathbf{I}_n$. Rademacher matrix: $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$, $d_j = \pm 1$ with prob $\frac{1}{2}$. $\mathbb{E}[d_j] = 0, \mathrm{Var}(d_j) = 1$. Orthogonal: $\mathbf{D}^\top \mathbf{D} = \mathbf{I}_n$. $\mathbf{y} = \mathbf{H}_n \mathbf{D} \mathbf{x}$ is more "flattened" than $\mathbf{x}$.

Merits of Randomized Hadamard Transform (RHT)

RHT can flatten vector variance / redistribute energy. Normalized Hadamard: $\mathbf{H}_n = \frac{1}{\sqrt{n}}\mathbf{H}$, entries $h_{ij} = \pm\frac{1}{\sqrt{n}}$. Orthogonal: $\mathbf{H}_n^\top \mathbf{H}_n = \mathbf{I}_n$.
Rademacher matrix: $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$, $d_j = \pm 1$ with prob $\frac{1}{2}$.
$\mathbb{E}[d_j] = 0, \mathrm{Var}(d_j) = 1$. Orthogonal: $\mathbf{D}^\top \mathbf{D} = \mathbf{I}_n$. $\mathbf{y} = \mathbf{H}_n\mathbf{D}\mathbf{x}$ is more "flattened" than $\mathbf{x}$. **Bounding effect**
$\mathbf{D}$ introduces randomness, $\mathbf{H}_n$ averages unevenness. (Intuition from Tropp)

Lemma

Let $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$. *Let* $\mathbf{y} := \mathbf{H}_n\mathbf{D}\mathbf{x} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T$. *For any* $i = 1, \ldots, n$, $\mathbf{y}_i = \sum_{j=1}^n h_{ij} d_j \mathbf{x}_j$. (??) $\Pr\{|\mathbf{y}_i| \geq t\} \leq 2\exp\left(-\frac{nt^2}{2\|\mathbf{x}\|_2^2}\right)$. (??)
(Sub-Gaussian by Hoeffding)

Take $t = \sqrt{\frac{\log(n)}{n}}\|\mathbf{x}\|_2$: $\Pr\left\{|\mathbf{y}_i| \geq \frac{\sqrt{\log(n)}}{\sqrt{n}}\|\mathbf{x}\|_2\right\} \leq 2n^{-\frac{1}{2}} \to 0$ as $n \to \infty$.
This is the bounding effect.

RHT: Asymptotic Normality

Mean and Variance of $\mathbf{y}_i$: $\mathbf{y}_i = \sum_{j=1}^{n} h_{ij}\, d_j\, \mathbf{x}_j$.

$\mathbb{E}[\mathbf{y}_i] = \sum_{j=1}^{n} h_{ij}\, \mathbf{x}_j\, \mathbb{E}[d_j] = 0$. $\mathrm{Var}(\mathbf{y}_i) = \mathbb{E}\left[\left(\sum_{j=1}^{n} h_{ij}\, d_j\, \mathbf{x}_j\right)^2\right] =$
$\sum_{j=1}^{n} h_{ij}^2\, \mathbf{x}_j^2\, \mathbb{E}[d_j^2] = \sum_{j=1}^{n} (\frac{1}{n}) \mathbf{x}_j^2 = \frac{1}{n}\|\mathbf{x}\|_2^2$. Let $\eta_j = h_{ij}\, d_j\, \mathbf{x}_j$. These are independent (not i.i.d. if $\mathbf{x}_j$ differ). $\mathbb{E}[\eta_j] = 0$, $\mathrm{Var}[\eta_j] = \mathbf{x}_j^2/n$.

Theorem (CLT - Lindeberg form)

Let $X_1, X_2, \ldots$ be independent r.v. with $\mathbb{E}[X_k] = \mu_k, \mathrm{Var}(X_k) = \sigma_k^2 < \infty$. 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

Let $\mathbf{y} := \mathbf{H}_n\mathbf{D}\mathbf{x}$. Under Lindeberg CLT for η_j: $\mathbf{y}_i \xrightarrow{d} \mathcal{N}\left(0, \frac{\|\mathbf{x}\|_2^2}{n}\right)$ (9)
(Original paper uses (??) for this) as long as

RHT: Eigenvalue Preservation

$\mathbf{H}_n$ is orthogonal, $\mathbf{D}$ is orthogonal.

Lemma

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

$$\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} = (\mathbf{H}_n\mathbf{D}\mathbf{X})^\top(\mathbf{H}_n\mathbf{D}\mathbf{X}) = \mathbf{X}^\top\mathbf{D}^\top\mathbf{H}_n^\top\mathbf{H}_n\mathbf{D}\mathbf{X} = \mathbf{X}^\top\mathbf{D}^\top\mathbf{I}_n\mathbf{D}\mathbf{X} = \mathbf{X}^\top\mathbf{I}_n\mathbf{X} = \mathbf{X}$$

Thus, $\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}}$ and $\mathbf{X}^\top\mathbf{X}$ have the same eigenvalues.

Experimental Results

Comparisons of RHT Sampling partitions vs Uniform Sampling partitions. Metric: Relative efficiency (Global MSE / Distributed MSE with equal weights). Simulation: $n = 8192, p = 30$ (models $p/n \to 0$, large n).

Experimental Results

Comparisons of RHT Sampling partitions vs Uniform Sampling partitions. Metric: Relative efficiency (Global MSE / Distributed MSE with equal weights). Simulation: $n = 8192, p = 30$ (models $p/n \to 0$, large n).

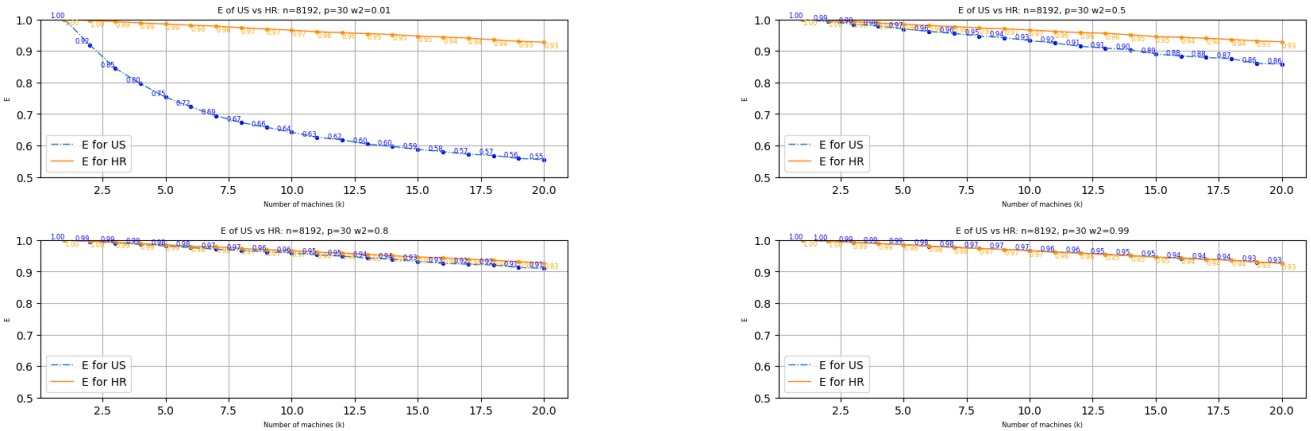


Figure: Comparison of Relative Efficiency with varying a_2. Settings: $c = 100, \boldsymbol{\mu}_2 = (5, \ldots, 5)^\top, p = 30$.

- Gap large for small a_2 (e.g., 0.01). Reasonable: small $a_2 \implies$ inflated cluster hard to catch; RHT flattens variance.
- RHT efficiency line nearly horizontal at 1, matches theory ($n \to \infty$).

Experimental Results (Continued)

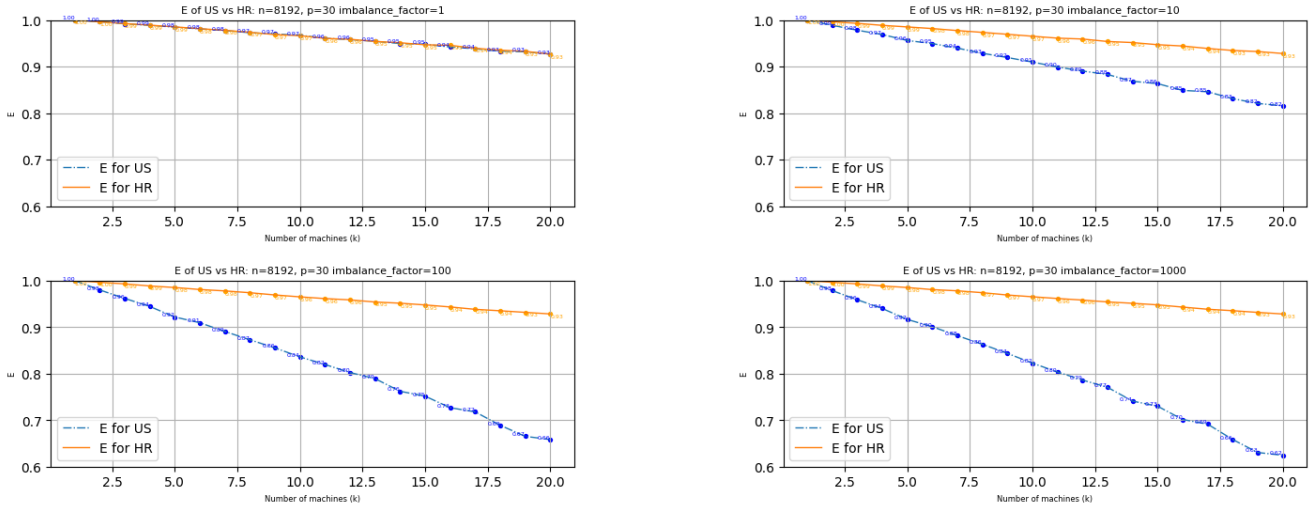


Figure: Comparison of Relative Efficiency with varying c. Settings: $a_2 = 0.2, \boldsymbol{\mu}_2 = (5, \ldots, 5)^\top, p = 30$.

- Gap increases drastically with c, matches theory.
- RHT efficiency line nearly horizontal at 1.

Experimental Results (Continued)

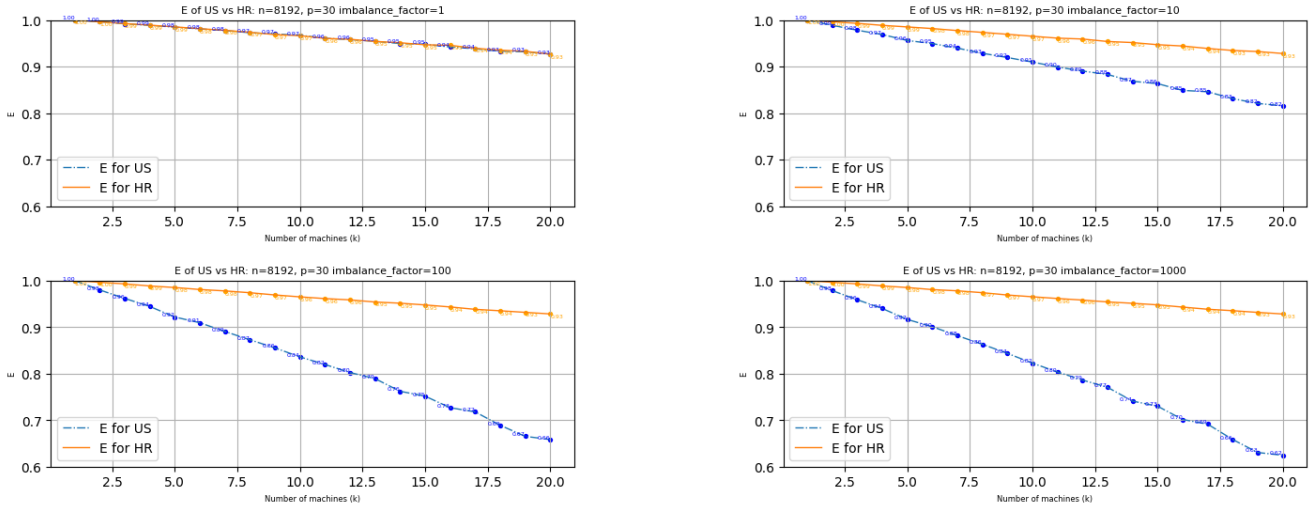
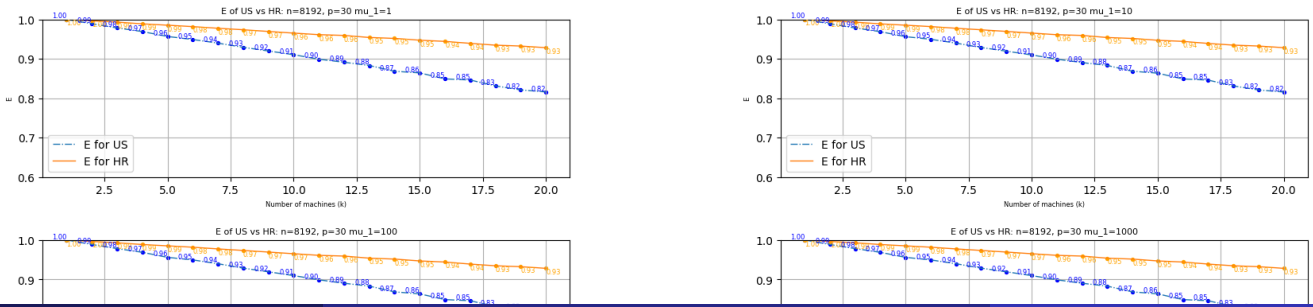


Figure: Comparison of Relative Efficiency with varying c. Settings: $a_2 = 0.2, \boldsymbol{\mu}_2 = (5, \ldots, 5)^\top, p = 30$.

- Gap increases drastically with c, matches theory.
- RHT efficiency line nearly horizontal at 1.



Discussion

- Future work: Further focus on proof of similar main result (Corollary 1).
- Consider the case where p is comparable to n (e.g., $p = c_1 n$). This means $p \to \infty$ at the same level as $n \to \infty$.
- Current result is conditioned on GMM, a specific distribution.
- Merits of RHT should be applied to more general settings of distributions.
- Future exploration: More general distributions for RHT Sampling.

Acknowledgements

- 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: Proof of Lemma 1 I

Appendix: Proof of Lemma 1 II

Proof (Proof of Lemma 1)

Fix noise $\varepsilon \sim \mathcal{N}(0, \mathbf{I}_n)$, *so* $\mathrm{Cov}(\varepsilon) = \mathbf{I}_n$; $\sigma^2 = 1$. *Abbreviate* $a_i := \mathrm{tr}\big[(\mathbf{X}_i^\top \mathbf{X}_i)^{-1}\big] > 0$.

Full-sample estimator: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{Y}$.

$\hat{\beta} - \beta = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \varepsilon$. *Since* $\mathbb{E}[\varepsilon] = \mathbf{0}$, $\mathbb{E}[\hat{\beta}] = \beta$.

$\mathrm{Cov}(\hat{\beta}) = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{I}_n \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} = (\mathbf{X}^\top \mathbf{X})^{-1}$. *For mean-zero random vector* $\mathbf{Z}$, $\mathbb{E}\|\mathbf{Z}\|_2^2 = \mathrm{tr}(\mathrm{Cov}(\mathbf{Z}))$. *Apply to* $\mathbf{Z} = \hat{\beta} - \beta$:

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

Local estimators and aggregation: *Worker* i: $\mathbf{Y}_i = \mathbf{X}_i\beta + \varepsilon_i$, $\varepsilon_i \overset{indep}{\sim} \mathcal{N}(0, \mathbf{I}_{n_i})$. $\hat{\beta}_i = (\mathbf{X}_i^\top \mathbf{X}_i)^{-1}\mathbf{X}_i^\top \mathbf{Y}_i = \beta + (\mathbf{X}_i^\top \mathbf{X}_i)^{-1}\mathbf{X}_i^\top \varepsilon_i$. *Unbiased,* $\mathrm{Cov}(\hat{\beta}_i) = (\mathbf{X}_i^\top \mathbf{X}_i)^{-1}$. *Combine:* $\hat{\beta}_{\mathrm{dist}}(w) := \sum_{i=1}^k w_i \hat{\beta}_i$, *with* $\sum w_i = 1$.

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

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

Choosing weights: *Auxiliary* $u_i = w_i\sqrt{a_i}$, $v_i = 1/\sqrt{a_i}$.

$1 = \big(\sum u_i v_i\big)^2 \le \big(\sum u_i^2\big)\big(\sum v_i^2\big) = \big(\sum w_i^2 a_i\big)\big(\sum \frac{1}{a_i}\big)$ *(Cauchy-Schwarz).*

So $\sum w_i^2 a_i \ge$ [illegible] *Equality if* u, v *linearly dependent* ($w_i \sqrt{a_i}$ [illegible] $/\sqrt{a_i}$).

Appendix: Proof of Lemma 3 I

Appendix: Proof of Lemma 3 II

Proof (Proof of Lemma 3)

Row $\mathbf{X}_{j,*}$ *from GMM.*

1. **Mean***:* $\mathbb{E}[\mathbf{X}_{j,*}] = a_1\,\mathbf{0} + a_2\,\boldsymbol{\mu}_2 = a_2\,\boldsymbol{\mu}_2$.

2. **Second moment***:*

$\mathbb{E}[\mathbf{X}_{j,*}\mathbf{X}_{j,*}^\top] = a_1\,\boldsymbol{\Sigma} + a_2\,(c\boldsymbol{\Sigma} + \boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top) = (a_1 + a_2 c)\,\boldsymbol{\Sigma} + a_2\,\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top$.

3. **Covariance***:*

$\mathrm{Cov}(\mathbf{X}_{j,*}) = \mathbb{E}[\mathbf{X}_{j,*}\mathbf{X}_{j,*}^\top] - \mathbb{E}[\mathbf{X}_{j,*}]\,\mathbb{E}[\mathbf{X}_{j,*}]^\top = (a_1 + a_2 c)\,\boldsymbol{\Sigma} + a_1 a_2\,\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top$.

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

$\mathbb{E}[\mathbf{X}_i^\top\mathbf{X}_i] = \sum_{\ell\in\mathbf{S}} \mathbb{E}[\mathbf{X}_{\ell,*}\mathbf{X}_{\ell,*}^\top] = \frac{n}{K}\left((a_1 + a_2 c)\boldsymbol{\Sigma} + a_2\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top\right)$. (**S** *rows for machine* i*).*

Transformed matrix $\tilde{\mathbf{X}} = \mathbf{H}_n\mathbf{D}\mathbf{X}$*. Row* $\tilde{\mathbf{X}}_{\ell,*} = \mathbf{e}_\ell^\top\tilde{\mathbf{X}} = \sum_{j=1}^n h_{\ell j}\,d_j\,\mathbf{X}_{j,*}$.

1. **Mean of transformed row***:* $\mathbb{E}[\tilde{\mathbf{X}}_{\ell,*}] = \sum_{j=1}^n h_{\ell j}\,\mathbb{E}[d_j]\,\mathbb{E}[\mathbf{X}_{j,*}] = \mathbf{0}$ *(since* $\mathbb{E}[d_j] = 0$*).*

2. **Second moment of transformed row***:* $\mathbb{E}[d_j d_k] = \delta_{jk}$. $\mathbb{E}\big[\tilde{\mathbf{X}}_{\ell,*}\tilde{\mathbf{X}}_{\ell,*}^\top\big] = \sum_{j=1}^n h_{\ell j}^2\,\mathbb{E}[\mathbf{X}_{j,*}\mathbf{X}_{j,*}^\top] = \frac{1}{n}\sum_{j=1}^n \mathbb{E}[\mathbf{X}_{j,*}\mathbf{X}_{j,*}^\top] = (a_1 + a_2 c)\,\boldsymbol{\Sigma} + a_2\,\boldsymbol{\mu}_2\boldsymbol{\mu}_2^\top$. *(*$h_{\ell j}^2 = 1/n$*).*

Appendix: Proof of Lemma 5 I

Appendix: Proof of Lemma 5 II

Proof (Proof of Lemma 5 (Boundedness))

Theorem (Hoeffding's Inequality)

Let $X_1, \ldots, X_n$ be independent r.v. $a_i \leq X_i \leq b_i$ a.s. $S_n = \sum X_i$. Then for $t > 0$, $\Pr(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum (b_i - a_i)^2}\right)$.

Recall $\mathbf{y}_i = \sum_{j=1}^n h_{ij} d_j \mathbf{x}_j$. Let $\eta_j = h_{ij} d_j \mathbf{x}_j$. η_j are independent (since d_j are i.i.d. and $h_{ij}, \mathbf{x}_j$ fixed for given i, j). Bound for η_j: $|\eta_j| = |h_{ij} d_j \mathbf{x}_j| = \frac{1}{\sqrt{n}} |\mathbf{x}_j|$ (since $h_{ij} = \pm 1/\sqrt{n}, |d_j| = 1$). So $\frac{-1}{\sqrt{n}} |\mathbf{x}_j| \leq \eta_j \leq \frac{1}{\sqrt{n}} |\mathbf{x}_j|$. Here $a_j = \frac{-1}{\sqrt{n}} |\mathbf{x}_j|$, $b_j = \frac{1}{\sqrt{n}} |\mathbf{x}_j|$. $S_n = \mathbf{y}_i = \sum \eta_j$. $\mathbb{E}[S_n] = \mathbb{E}[\mathbf{y}_i] = 0$.

$\sum_{j=1}^n (b_j - a_j)^2 = \sum_{j=1}^n \left(\frac{2}{\sqrt{n}} |\mathbf{x}_j|\right)^2 = \sum_{j=1}^n \frac{4}{n} \mathbf{x}_j^2 = \frac{4}{n} \|\mathbf{x}\|_2^2$. *(Corrected from text's $2t^2 / \sum (\frac{2}{\sqrt{n}} |\mathbf{x}_j|)^2$ directly to $\frac{nt^2}{2\|\mathbf{x}\|_2^2}$ which implies denominator was $\frac{4\|\mathbf{x}\|_2^2}{n}$, so $2t^2 / (\frac{4\|\mathbf{x}\|_2^2}{n}) = \frac{2t^2 n}{4\|\mathbf{x}\|_2^2} = \frac{nt^2}{2\|\mathbf{x}\|_2^2}$) Plugging into Hoeffding:*

$\Pr\{|\quad - 0| > t\} \leq 2 \exp\left(\quad\right) = 2 \exp\left(\quad\right)$ *This is Eq. (33)*

Appendix: Proof of Lemma 6 I

Appendix: Proof of Lemma 6 II

Proof (Proof of Lemma 6 (Asymptotic Normality))

Using Lindeberg CLT (Theorem 4). $\eta_j = h_{ij} d_j \mathbf{x}_j$, $j = 1, \ldots, n$. *Independent.* $\mathbb{E}[\eta_j] = 0$. $\mathrm{Var}[\eta_j] = h_{ij}^2 \mathbf{x}_j^2 \mathrm{Var}[d_j] = \frac{1}{n}\mathbf{x}_j^2$. $s_n^2 = \sum_{j=1}^n \mathrm{Var}[\eta_j] = \sum_{j=1}^n \frac{1}{n}\mathbf{x}_j^2 = \frac{1}{n}\|\mathbf{x}\|_2^2$. *Lindeberg condition:* $\frac{1}{s_n^2}\sum_{j=1}^n \mathbb{E}\left[\eta_j^2 \mathbf{1}_{\{|\eta_j| > \varepsilon s_n\}}\right] \xrightarrow{n\to\infty} 0$. $|\eta_j| = \frac{1}{\sqrt{n}}|\mathbf{x}_j|$. *Condition becomes* $\mathbf{1}_{\{\frac{1}{\sqrt{n}}|\mathbf{x}_j| > \varepsilon \frac{1}{\sqrt{n}}\|\mathbf{x}\|_2\}} = \mathbf{1}_{\{|\mathbf{x}_j| > \varepsilon\|\mathbf{x}\|_2\}}$. *Term becomes* $\frac{n}{\|\mathbf{x}\|_2^2}\sum_{j=1}^n \mathbb{E}\left[\frac{1}{n}\mathbf{x}_j^2 \mathbf{1}_{\{|\mathbf{x}_j| > \varepsilon\|\mathbf{x}\|_2\}}\right] = \frac{1}{\|\mathbf{x}\|_2^2}\sum_{j=1}^n \mathbf{x}_j^2 \mathbf{1}_{\{|\mathbf{x}_j| > \varepsilon\|\mathbf{x}\|_2\}}$. *(Since* $\mathbf{x}_j$ *are fixed,* $\mathbb{E}$ *over* d_j *only affects* η_j^2*, not the indicator on* $|\mathbf{x}_j|$*). This sum goes to 0 if for any* $\varepsilon > 0$*, the terms* $\mathbf{x}_j^2$ *where* $|\mathbf{x}_j| > \varepsilon\|\mathbf{x}\|_2$ *become negligible relative to* $\|\mathbf{x}\|_2^2$*. This is equivalent to the Lyapunov condition if we consider* $\mathbf{x}_j$ *values. The text states this sum is* $\sum_{i=k}^n \frac{\mathbf{x}_{(i)}^2}{\|\mathbf{x}\|_2^2}$ *if* $\frac{\mathbf{x}_{(k-1)}}{\|\mathbf{x}\|_2} \le \varepsilon < \frac{\mathbf{x}_{(k)}}{\|\mathbf{x}\|_2}$ *(ordered values* $\mathbf{x}_{(i)}$*). The condition for this to go to 0 as* $n \to \infty$ *is that no single* $\mathbf{x}_j^2$ *dominates the sum* $\|\mathbf{x}\|_2^2$*. This is satisfied if* $\lim_{n\to\infty}\left(\max_{j=1,\ldots,n} \frac{\mathbf{x}_j^2}{\|\mathbf{x}\|_2^2}\right) = 0$*. This is Eq. (33). Then*