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

Prof. Zhixiang Zhang Yishu Yang

Faculty of Science and Technology, University of Macau

May 20, 2025

Abstract

Distributed sketching has emerged as a pivotal technique for efficiently training linear regression models on massive datasets.

- Investigate the relative efficiency of distributed Ordinary Least Squares (OLS) estimation via partitioned machines.
- Identify the specific distribution of Gaussian Mixture Models (GMM) containing biased variance where uniform sampling would underperform.
- Demonstrate that the application of the Randomized Hadamard Transform (RHT) prior to partitioning can substantially enhance the relative efficiency of OLS estimation

Abstract:Continued

- RHT enables efficient equal-weighted averaging of local OLS estimators, obviating the need for complex weight adjustments.
- rigorously prove that under the condition of sublinear growth of the number of features p with respect to the number of samples n, specifically when $\ln n = o(p(n))$ and $p(n) \ln p(n) = o(n)$, the relative efficiency of OLS estimation with equal-weighted partitions after RHT asymptotically approaches the ideal value of 1 as n tends to infinity.

Abstract:Continued

- difference between the trace of the inverse of the local Gram matrix and the trace of the inverse of the scaled global Gram matrix converges to zero at a specific rate.
- Experimental results on the sublinear growth of p with respect to n and the linear growth of p with respect to n.

Keywords: Distributed Sketching, Ordinary Least Squares (OLS), Randomized Hadamard Transform (RHT), Partitioned Machines, Relative Efficiency, Gaussian Mixture Model (GMM)

Introduction: Distributed Sketching for Linear Regression

Background:

- Distributed sketching is a common tool for efficiently training linear regression models on large datasets by partitioning data into K blocks, performing local regression, and averaging parameters.
- Prior research often focused on:
  - Optimized absolute sketching errors or bias proportions (Wang, 2018[1]).
  - Accurate expectation of OLS approximation error (Derezinski, 2023[2]).
- Less focus on comparing relative efficiency of different sampling methods (e.g., SRHT, LBS).
- Dobriban and Liu[3] discussed various efficiencies in distributed OLS sketching, noted as relevant for further study.

Introduction: Continued

Our Main Research Contributions:

- To identify specific data matrix $\mathbf{X}$ distributions where equally partitioned uniform sampling drastically decreases OLS relative efficiency (referencing Section 3.2 of Dobriban and Sheng[4]).
- To demonstrate that Randomized Hadamard Transform (RHT) improves OLS relative efficiency by:
  - Flattening the variance of local Gram matrices.
  - Enabling equal weight averaging, saving time on weight adjustments (from 2.6).
- To rigorously prove that RHT with equal-weighted partitions achieves a relative OLS efficiency of 1 when $n \to \infty$ and $p \to \infty$ (with p having sublinear growth of n larger than $\log n$) (see 1).
  - Interpretation: The difference $\text{tr}[(\tilde{\mathbf{X}}_i^\top \tilde{\mathbf{X}}_i)^{-1}] - \text{tr}[(\frac{1}{K}\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1}]$ converges to 0 at rate $\mathcal{O}(\sqrt{p^3 \frac{\log n}{n^3}})$ (4.5), stabilizing the local Gram matrix.

Introduction: Continued

In Short:

- By using RHT to flatten local Gram matrix variance, we can use equal weight $\frac{1}{K}$ averaging for K machines, avoiding complex weight adjustments for local OLS estimators.

Theoretical Foundations & Methodology:

- RHT intuition from Tropp[5] and Cherapanamjeri[6].
- Proofs utilize probability inequalities (Bernstein, Matrix Chernoff) from Vershynin[7], Tropp[5], and [8] for Corollary 1.

Normalized Hadamard Matrix

- Definition: A square matrix $\mathbf{H}$ of dimension $n \times n$ is a Hadamard matrix if:
  - $\mathbf{H}/\sqrt{n}$ is orthogonal
  - $|\mathbf{H}_{ij}| = 1$ for all $i, j = 1, \ldots, n$
- Example: Walsh-Hadamard matrix, defined recursively:

$$\mathbf{H}_n = \begin{pmatrix} \mathbf{H}_{n/2} & \mathbf{H}_{n/2} \\ \mathbf{H}_{n/2} & -\mathbf{H}_{n/2} \end{pmatrix}, \quad \mathbf{H}_1 = 1$$

- Property: Orthogonal, i.e., $\mathbf{H}_n \mathbf{H}_n^T = n\mathbf{I}_n$
- Role: Used in Randomized Hadamard Transform (RHT) to preprocess data

OLS Problems: Global and Distributed Estimators

- Model: $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \mathbf{I}_n)$
- Global OLS estimator:

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

  - MSE: $M(\hat{\beta}) = \mathbb{E}\|\beta - \hat{\beta}\|^2$
  - Challenge: High computational cost as $n \to \infty$
- Distributed estimator:

$$\hat{\beta}_{\text{dist}} = \sum_{i=1}^{K} w_i \hat{\beta}_i$$

  - $\hat{\beta}_i$: Local OLS estimate on partition i
  - MSE: $M(\hat{\beta}_{\text{dist}}) = \mathbb{E}\|\beta - \hat{\beta}_{\text{dist}}\|^2$
  - Solution: Partition data into K blocks, reducing computational burden

Relative Efficiency Interpretation

- Definition (Lemma 1):

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

- Context: Uniform sampling with GMM distribution (Definition 1) reduces efficiency drastically
  - See finite sample results by Dobriban and Sheng [4]
- Main Result: RHT flattens variance, achieving:

$$\mathbb{E}(\mathbf{I}_p, \tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_K) = 1 \text{ as } n, p \to \infty$$

(Corollary 1)

- Implication: Equal weights ($w_i = \frac{1}{K}$) suffice with RHT (Lemma 2)

Remark (Growth Conditions for Parameter $p(n)$)

In our asymptotic analysis, we require the parameter $p(n)$, a function $p : \mathbb{N} \to \mathbb{R}^+$, to adhere to a specific growth regime ensuring desirable convergence properties for probability bounds. We term this controlled intermediate growth, defined by the following conditions as $n \to \infty$:

(C1) *$p(n)$ is superlogarithmic with respect to n:* $\ln n = o(p(n))$, *meaning* $\lim_{n\to\infty} \frac{\ln n}{p(n)} = 0$.

(C2) *The product $p(n) \ln p(n)$ is sublinear with respect to n:* $p(n) \ln p(n) = o(n)$, *meaning* $\lim_{n\to\infty} \frac{p(n) \ln p(n)}{n} = 0$.

Condition (C1) ensures $p(n)$ dominates logarithmic factors of n. Condition (C2) ensures $p(n)$ does not grow too rapidly, specifically, slower than $n/\ln p(n)$. Standard examples of $p(n)$ satisfying (C1) and (C2) include $p(n) = n^a$ for any $a \in (0, 1)$, $p(n) = (\ln n)^k$ for $k > 1$, and $p(n) = n/(\ln n)^k$ for $k > 1$.

Corollary

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 *satisfies Remark1.*
$\mathbf{\breve{X}} = \mathbf{H}_n\mathbf{D}\mathbf{X}$, *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 $\mathbf{\breve{X}}_1, \mathbf{\breve{X}}_2, \ldots, \mathbf{\breve{X}}_K$ *are the partitioned machine submatrices after being Randomzied Hadamard Transformed (RHT) from Remark 2.*
Then, based on the 2.3 result of Lemma 1, we have the following corollary:

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

as long as the condition of 2.8 is satisfied.

finite sample results of Dobriban and Sheng [4] 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 2.*

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

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

Our intuition for equal partitions

1. 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 $\mathbf{M}_i = \mathbf{X}_i^\top \mathbf{X}_i$.
2. 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 $\mathbf{M}$.
3. 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.

Lemma

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

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

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 4.

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:

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

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

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

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

Our proposed RHT and the common uniform sampling

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 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* $\tilde{\mathbf{X}}$*.*

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

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

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

Definition:Continued

The mixture parameters satisfy

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

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

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.

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:*

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

Proof of Corollary 1 (MAIN RESULT)

To prove this corollary rigorously, we need to show that as $n \to \infty$ and p satisfies Remark1, we have:

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

where $\tilde{\mathbf{X}} \in \mathbb{R}^{n\times p}$ is a matrix with growing number of rows n and p satisfies Remark1. 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 p satisfies Remark1.

Proof Continued

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., $|\text{tr}(M)| \leq \sqrt{p}\|M\|_F$, hence we have:

$$\begin{aligned} |\text{tr}(A^{-1}) - \text{tr}(B^{-1})| &= |\text{tr}(A^{-1}(A - B)B^{-1})| \\ &\leq \sqrt{p}\,\|\,A^{-1}(A - B)B^{-1}\|_F \\ &\leq \sqrt{p}\|A^{-1}\|_2\|A - B\|_F\|B^{-1}\|_2 \end{aligned} \tag{4.0}$$

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

Proof Continued

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.

Proof Continued: 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

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

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:

$$\mathbf{A}_{jk} - \mathbf{B}_{jk} = \sum_{r=1}^{n} \left(\mathbf{I}_r - \frac{1}{K}\right) (\mathbf{Y}_r)_j \, (\mathbf{Y}_r)_k .$$

Proof Continued

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.:

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

Follow the book[7], we navigate to the theorem of Bernstein's inequality.

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.

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

Proof Continued

Now we rewrite:

$$\cdot_{jk}^{(i)} = \sum_{r=1}^{n} \left(\mathbf{I}_r - \frac{1}{K} \right) (\mathbf{Y}_r)_j \, (\mathbf{Y}_r)_k = \sum_{r=1}^{n} \left(\mathbf{I}_r - \frac{1}{K} \right) \mathbf{z}_r$$

where 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.

Proof Continued

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 $\mathcal{N}_p(\mathbf{0}, \Sigma)$, and there is a probability of a_2 that this row is drawn from $\mathcal{N}_p(\mu_2, c\Sigma)$. Thus we have the following for the j-th element $\mathbf{X}_{rj}$ of the row $(\mathbf{X})_{r,*}$:

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

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 [7]:

Lemma (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 $\|Z_1\|_{\psi_2} \leq \gamma\sqrt{\Sigma_{jj}}$ for some absolute constant $\gamma > 0$, and we have $\|Z_2 - \mu_2\|_{\psi_2} \leq \omega\sqrt{c\Sigma_{jj}}$ for some absolute constant $\omega > 0$, this implies that we have $\|Z_2\|_{\psi_2} \leq \omega\sqrt{c\Sigma_{jj}} + \|\mu_2\|_{\psi_2}$.
Then we define $\kappa = \max\{\gamma\sqrt{\Sigma_{jj}}, \omega\sqrt{c\Sigma_{jj}} + \|\mu_2\|_{\psi_2}\}$ and have:

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

Proof Continued

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

Theorem (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[7] which is Lemma 2.7.7, we have:

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

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 ν^2:

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

Thus we could plug this ν^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:

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

To let the second term dominate, we choose:

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

Thus we have:

$$\Pr\{|\Delta_{jk}^{(i)}| \geq \chi\sqrt{n\log n}\} \;\leq\; 2\,\exp\left[-c\,\frac{\chi^2}{b^2}(\log n)\right] \;=\; 2\,n^{-c\,\chi^2/b^2}$$

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

$$\Pr\left\{\max_{j,k}|\Delta_{jk}^{(i)}| \;\geq\; \chi\sqrt{n\,\log n}\right\} \;\leq\; 2p^2\,n^{-\frac{c\,\chi^2}{b^2}}. \qquad \text{(Remark.I.1)}$$

Thus we have :

$$\|\mathbf{A}-\mathbf{B}\|_F^2 \;=\; \sum_{j,k}(|\Delta_{jk}^{(i)}|)^2 \;\leq\; \sum_{j,k}\left(\max_{j,k}|\Delta_{jk}^{(i)}|\right)^2 \;=\; p^2\,\chi^2\,n\,\log n. \qquad \text{(Remark.I.2)}$$

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

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

Here it should be noted that in order for $2p^2\,n^{-\frac{c\,\chi^2}{b^2}}$ to converge to 0 as n and p tends to infinity as well as p satisfies Remark 1, we could utilize the inequality here that $2p^2\,n^{-\frac{c\,\chi^2}{b^2}} < 2n^2\,n^{-\frac{c\,\chi^2}{b^2}}$ and only have to make sure of the following:

$$\chi > \sqrt{\frac{2b^2}{c}} \qquad \text{(Remark.I.3)}$$

L2 norm of inverse of gram matrix

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 (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

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

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

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

and for any $\epsilon \geq 0$,

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

Corollary (Simplified lower-tail bound)

Under the same hypotheses as Theorem 3, for any $\epsilon \in [0, 1]$ *one has*

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

Proof.

Starting from the bound:

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

We use the standard inequality here that is: $\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}} \le e^{-\epsilon^2/2}$, which is valid for $0 \le \epsilon \le 1$.
Substituting this inequality gives the right hand side exponential result:

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

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:

$$\Pr\left\{\lambda_{\min}\left(\sum_{l}\mathbf{X}_{l}\right) \leq (1-\epsilon)\,\mu_{\min}\right\} \;\leq\; p\,\exp\left(-\frac{\epsilon^{2}\,\mu_{\min}}{2R}\right)$$

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 $\mathbf{X}$ 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[7]:

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

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:

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

Again, we use the Theorem 1 here to derive the bound of $\|\mathbf{Y}_r\|_2^2$.
We take $t = p$ here, and we will reason on this choice in the following steps by steps. Firstly, when $t = p$, the left object in the maximum function of Theorem 1 is $\frac{p}{b^2}$ and the right object is $\frac{p}{b}$, thus we take $\alpha = \max\{b^2, b\}$ and the probability becomes $2\exp\{-\frac{c}{\alpha}p\}$
Again by The Law of Union of Probability and plugging with $t = p$, we have:

$$\Pr\left\{\max_{1\le r\le n}\mathbf{R}_r > p\right\} \le \frac{2n}{\exp\{\frac{c}{\alpha}p\}} \tag{4.2}$$

By the condition $(C1)$ of Remark1, it is easy to see that

$$\lim_{p,n\to\infty}\frac{\log n}{p} = 0 \Longrightarrow \lim_{p,n\to\infty}\frac{2n}{\exp\left(\frac{c}{\alpha}p\right)} = 0$$

And inversely, the condition $(C1)$ is derived just because we need to ganrantee that the probability on the right hand side of 4.2 should converge to 0.

By the illustration above, the $\lambda_{\max}$ of $\mathbf{W}_r$ is bounded by p with probability at least $1 - \frac{2n}{\exp\{\frac{c}{\alpha}p\}}$. That is:

$$\Pr\{\lambda_{\max}(\mathbf{W}_r) \leq p\} \geq 1 - \frac{2n}{\exp\{\frac{c}{\alpha}p\}}$$

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 $\Sigma^\star$ first.

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

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

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

Then we have by the Corollary 2:

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

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

And the probability on the right hand side could converge to 0 as n and p tends to infinity, which is ganranteed by ($C2$) of Remark1.
That is:

$$\lim_{p,n\to\infty} \frac{p\log p}{n} = 0 \implies \lim_{p,n\to\infty} p\exp\left(-\frac{\iota n}{p}\right) = 0$$

And inversely, the condition ($C2$) is derived just because we need to ganrantee that the probability on the right hand side here should converge to 0.
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:

$$1 - p\exp\Big(-\frac{\iota n}{p}\Big) - \frac{2n}{\exp\{\frac{c}{\alpha}p\}}$$

We have

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

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:

$$\Pr\left\{\|\mathbf{B}^{-1}\|_2 \leq \frac{K}{(1-\sigma)\lambda^\star}\cdot\frac{1}{n}\right\} \geq 1 - p\exp\Big(-\frac{\iota n}{p}\Big) - \frac{2n}{\exp\{\frac{c}{\alpha}p\}} \tag{4.3}$$

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 $\|\mathbf{A}^{-1}\|_2$:

$$\Pr\left\{\|\mathbf{A}^{-1}\|_2 \leq \frac{K}{(1-\sigma)\lambda^\star} \cdot \frac{1}{n}\right\} \geq 1 - p\exp\left(-\frac{\iota n}{p}\right) - \frac{2n}{\exp\left\{\frac{c}{\alpha}p\right\}} \tag{4.4}$$

Again, only make sure the dimension n and p satisfies Remark1.

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:

$$1 - 2p\exp\left(-\frac{\iota n}{p}\right) - \frac{4n}{\exp\{\frac{c}{\alpha}p\}} - 2p^2\, n^{-\frac{c\,\chi^2}{b^2}}$$

We have:

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

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

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

With the probability that has been discussed above and only to make sure to choose that χ is larger than $\sqrt{\frac{2b^2}{c}}$.

The right hand side of the absolute difference converge to 0 as n and p tends to infinity, which is ganranteed by Remark1 for its sublinear condition of p. Thus the final result here is that:

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

Result 4.5 implies that $\operatorname{tr}(\mathbf{A}^{-1})$ converges to $\operatorname{tr}(\mathbf{B}^{-1})$ with probability 1 as $n\to\infty$ and $p\to\infty$ and p satisfies Remark1. This is the result we desire to show in the very beginning:

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

Remark 1 of n and p

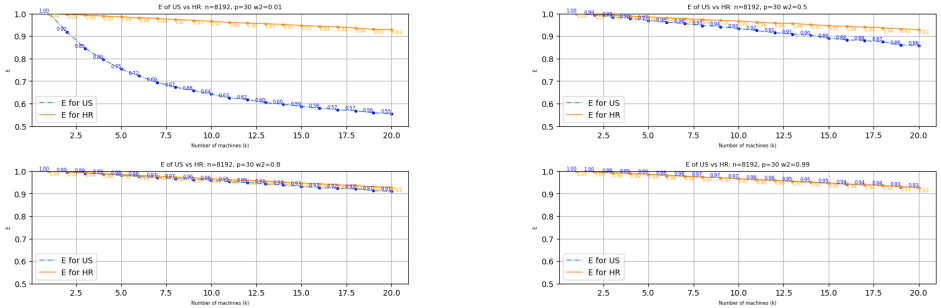

Figure: Comparison of Relative Efficiency between RHT Sampling and Uniform Sampling with varying a_2 proportion. Settings: $c = 100$, $\mu_2 = (5, \ldots, 5)^\top$, $p = 30$.

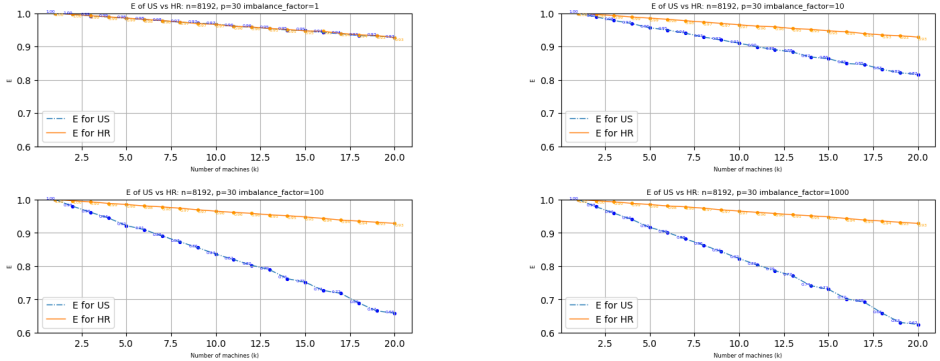

Figure: Comparison of Relative Efficiency between RHT Sampling and Uniform Sampling with varying c inflated ratio. Settings: $a_2 = 0.2$, $\mu_2 = (5, \ldots, 5)^\top$, $p = 30$.

Continuing

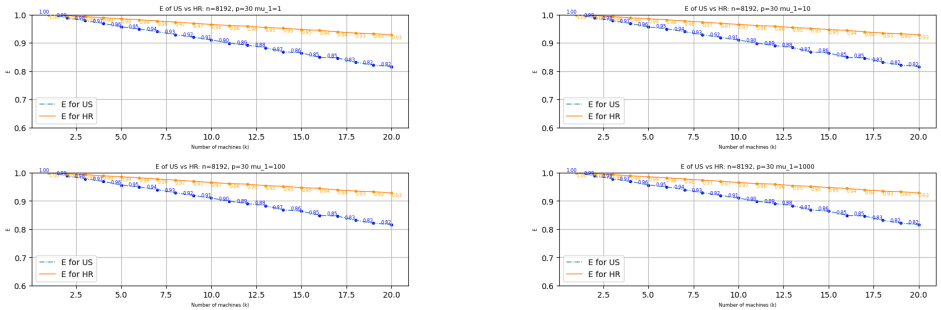

Figure: Comparison of Relative Efficiency between RHT Sampling and Uniform Sampling with varying μ_2 mean vector. Settings: $a_2 = 0.2$, $c = 10$, $p = 30$.

p is linear growth of n

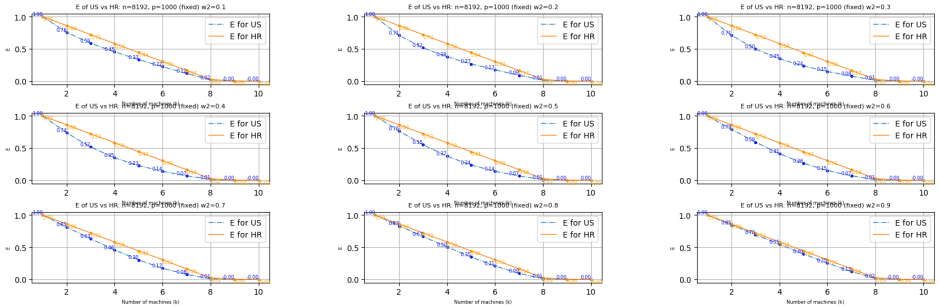

Figure: Comparison of Relative Efficiency between RHT Sampling and Uniform Sampling with varying a_2 proportion. Settings: $c = 100$, $\mu_2 = (5, \ldots, 5)^\top$, $p = 1000$.

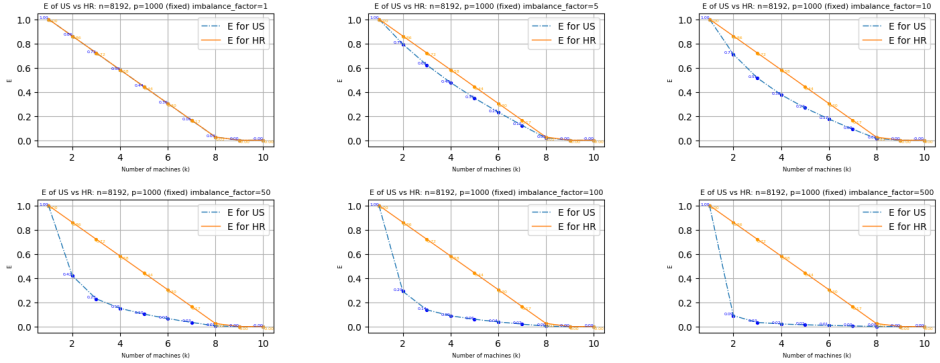

Figure: Comparison of Relative Efficiency between RHT Sampling and Uniform Sampling with varying c inflated ratio. Settings: $a_2 = 0.2$, $\mu_2 = (5, \ldots, 5)^\top$, $p = 1000$.

References I

[1] 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.

[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.

References II

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

[6] 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.

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

[8] 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.