Optimal Iterative Sketching with the Subsampled Randomized Hadamard Transform

Jonathan Lacotte[*]
Department of Electrical Engineering
Stanford University
lacotte@stanford.edu

Sifan Liu[*]
Department of Statistics
Stanford University
sfliu@stanford.edu

Edgar Dobriban
Department of Statistics
University of Pennsylvania
dobriban@wharton.upenn.edu

Mert Pilanci
Department of Electrical Engineering
Stanford University
pilanci@stanford.edu

Abstract

Random projections or sketching are widely used in many algorithmic and learning contexts. Here we study the performance of iterative Hessian sketch for least-squares problems. By leveraging and extending recent results from random matrix theory on the limiting spectrum of matrices randomly projected with the subsampled randomized Hadamard transform, and truncated Haar matrices, we can study and compare the resulting algorithms to a level of precision that has not been possible before. Our technical contributions include a novel formula for the second moment of the inverse of projected matrices. We also find simple closed-form expressions for asymptotically optimal step-sizes and convergence rates. These show that the convergence rate for Haar and randomized Hadamard matrices are identical, and asymptotically improve upon Gaussian random projections. These techniques may be applied to other algorithms that employ randomized dimension reduction.

1 Introduction

Random projections are a classical way of performing dimensionality reduction, and are widely used in many algorithmic and learning contexts, e.g., [32, 17, 35, 9] etc. In this work, we study the performance of the iterative Hessian sketch [24], in the context of overdetermined least-squares problems

$$x^* := \operatorname*{argmin}_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{2}\|Ax - b\|^2 \right\} . \tag{1}$$

Here $A \in \mathbb{R}^{n \times d}$ is a given data matrix with $n \geqslant d$ and $b \in \mathbb{R}^n$ is a vector of observations. For simplicity of notations, we assume throughout this work that $\text{rank}(A) = d$. We will leverage and extend recent results on the limiting spectral distributions of two classical subspace embeddings, random uniform projections and the subsampled randomized Hadamard transform (SRHT), to compare corresponding iterative Hessian sketch versions.

The iterative Hessian sketch (IHS) is an effective iterative method for solving least-squares [23, 24, 14, 28] (and more general convex optimal optimization problems [25]), and it aims to address the condition number dependency of standard iterative solvers as follows. Given step sizes $\{\mu_t\}$ and

[*]Equal contributions.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

momentum parameters $\{\beta_t\}$, it computes the update

$$x_{t+1} = x_t - \mu_t H_t^{-1} \nabla f(x_t) + \beta_t (x_t - x_{t-1}), \tag{2}$$

where the Hessian $H = A^\top A$ of f is approximated by $H_t = A^\top S_t^\top S_t A$, and $S_0, \dots, S_t, \dots$ are i.i.d. sketching (random) matrices with dimensions $m \times n$ and $m \ll n$. From now on, we refer to the i.i.d. property of the sketching matrices as *refreshed* matrices.

There are many possible choices for the sketching matrices S_t, and this is critical for the performance of the IHS. A classical sketch is a matrix $S \in \mathbb{R}^{m \times n}$ with independent and identically distributed (i.i.d.) Gaussian entries $\mathcal{N}(0, m^{-1})$, for which the matrix multiplication SA requires in general $\mathcal{O}(mnd)$ basic operations (using classical matrix multiplication). This is larger than the cost $\mathcal{O}(nd^2)$ of solving (1) with direct methods when $m \geqslant d$. Another well-studied embedding is the (truncated) $m \times n$ *Haar* matrix S, whose rows are orthonormal and with range uniformly distributed among the subspaces of $\mathbb{R}^n$ with dimension m. However, this requires time $\mathcal{O}(nm^2)$ to be formed, through a Gram-Schmidt procedure, which is also larger than $\mathcal{O}(nd^2)$.

The SRHT [1, 27] is another classical random orthogonal embedding. Due to the recursive structure of the Hadamard transform, the sketch SA can be formed in $\mathcal{O}(nd \log m)$ time, so that the SRHT is often viewed as a standard reference point for comparing sketching algorithms. Moreover, for many applications, random projections with i.i.d. entries perform worse compared to orthogonal projections [17, 18, 9]. More recently, this observation has also found some theoretical support in limited contexts [8, 36]. Works by [6] also showed the guaranteed improved performance in accuracy and/or speed. Consequently, along with computational considerations, these results favor the SRHT over Gaussian projections.

Our goal in this work is to design an optimal version of the IHS with SRHT and Haar embeddings. For this purpose, it is necessary to have a tight characterization of the spectral properties of the matrix $U^\top S^\top SU$ where U is an $n \times d$ partial orthogonal matrix (see, e.g., [13]). With Gaussian embeddings, the matrix $U^\top S^\top SU$ has the well-studied Wishart distribution, see e.g., [19, 3, 29, 5, 7, 38]. In fact, [13] provided an optimal IHS with Gaussian embeddings, and showed that the best achievable error $\|A(x_t - x^*)\|^2$ scales as $(d/m)^t$. However, a similar analysis does not work for SRHT and Haar sketches. To make progress on this problem, we aim to leverage powerful tools from asymptotic random matrix theory, and *we consider the asymptotic regime where we let the relevant dimensions go to infinity*.

Our technical analysis is based on asymptotic random matrix theory, see e.g., [3, 29, 5, 7, 38] etc. Classical results such as the Marchenko-Pastur law do not address well the case of the SRHT, and we leverage recent results on *asymptotically liberating sequences* established by [2] (see also [31] for prior work). Further, we are inspired by the work of [8], who, to our knowledge, first leveraged these results to study the SRHT. However, their results are limited to one-step "sketch-and-solve" methods, and do not address the iterative Hessian sketch. Moreover, while we build on their results, we also need to extend them significantly: for instance, we need to derive the second moment formula for $\theta_{2,h}$ in (11), which is novel and non-trivial to establish.

Beyond the IHS, there exist other randomized pre-conditioning methods [4, 10, 20, 26] for solving least-squares, which are based on the SRHT (or closely related sketches) which address effectively the condition number dependency of iterative solvers. Besides least-squares, SRHT sketches are widely used for a wide range of applications across numerical linear algebra, statistics and convex optimization, such as low-rank matrix factorization [11, 34], kernel regression [37], random subspace optimization [16], or sketch and solve linear regression [8], see the reviews above for applications. Hence, a refined analysis of the SRHT, including our specific technical contributions, may also lead to better algorithms in these fields.

Throughout the paper, we will consistently use the following assumptions and notations for the aspect ratios, $\gamma := \lim_{n,d \to \infty} \frac{d}{n} \in (0,1)$, $\xi := \lim_{n,m \to \infty} \frac{m}{n} \in (\gamma, 1)$ and $\rho_g := \frac{\gamma}{\xi} \in (0,1)$, and the subscript g (resp. h) will refer to Gaussian-related (resp. Haar and Hadamard-related) quantities. We use the notations $\|z\| \equiv \|z\|_2$ for the Euclidean norm of a real vector z, $\|M\|_2$ for the operator norm of a matrix M, and $\|M\|_F$ for its Frobenius norm. For a sequence of iterates $\{x_t\}$, we denote the error vector $\Delta_t := U^\top A(x_t - x^*)$, where U is the $n \times d$ matrix of left singular vectors of A. In particular, we have that $\|\Delta_t\|^2 = \|A(x_t - x^*)\|^2$.

1.1 Overview of our results, contributions and questions left open

All our contributions hold in the asymptotic limit $n, d, m \to \infty$, and under the aforementioned assumption that the aspect ratios (d/n) and (m/n) have finite limits.

We work with the matrix $U^\top S^\top S U$, where U is an $n \times d$ matrix with orthonormal columns and S is an $m \times n$ Haar or SRHT matrix. Our first results concern Haar projections (Section 3). By leveraging results about their limiting spectral distributions, and after some calculations with Stieljes transforms (defined below) we provide the following new trace formula (see Lemma 3.2):

$$\theta_{2,h} := \lim_{n\to\infty} \frac{1}{d} \operatorname{tr} \mathbb{E}\left[(U^\top S^\top S U)^{-2}\right] \;=\; \frac{(1-\gamma)(\gamma^2+\xi-2\gamma\xi)}{(\xi-\gamma)^3}\,.$$

As an application, we characterize explicitly the optimal step sizes μ_t and momentum parameters β_t of the IHS with Haar embeddings (Theorem 3.1). We emphasize that the optimal parameters have asymptotically closed form for any data matrix A, unlike for certain other propular methods such as gradient descent, which can be useful in practice. With these optimal parameters, we find that at *any* time step $t \geqslant 1$ (Theorem 3.1),

$$\lim_{n\to\infty} \frac{\mathbb{E}\|\Delta_t\|^2}{\|\Delta_0\|^2} = \rho_h^t\,, \tag{3}$$

where the convergence rate ρ_h is given by $\rho_h := \rho_g \cdot \frac{\xi(1-\xi)}{\gamma^2+\xi-2\xi\gamma}$, and always satisfies $\rho_h < \rho_g$. By comparing with the prior work [13], this implies that Haar embeddings have uniformly better performance than Gaussian ones. Further, as an immediate consequence of Theorem 2 in [13], we obtain that the optimal momentum parameters β_t are equal to 0, that is, Heavy-ball momentum does not accelerate the algorithm with refreshed Haar embeddings (Theorem 3.1 and following discussion). Thus, we are able to characterize explicitly the optimal version of the IHS with Haar embeddings.

Our next results concern SRHT sketches (Section 4). We prove that under the additional mild assumption on the initial error Δ_0 that $\mathbb{E}[\Delta_0\Delta_0^\top] = d^{-1} I_d$, the IHS with SRHT embeddings also has rate of convergence ρ_h (Theorem 4.1). This relies on novel formulas for the first two inverse moments of SRHT sketches (Lemma 4.3). Consequently, *SRHT matrices uniformly outperform Gaussian embeddings*. Then, we confirm numerically the above theoretical statements (Section 6).

We finally analyze the computational complexity of our method, in comparison to some standard randomized pre-conditioned solvers [26] for dense, ill-conditioned least-squares. We show that in our infinite-dimensional regime, we improve by a factor $\log d$ (Section 5).

Importantly, we specifically focus on the IHS with refreshed i.i.d. embeddings. An immediate variant of the IHS uses the same update (2), but with a fixed embedding S drawn only once at the first iteration, which is appealing in practice. In a concurrent paper [15] more recent to the initial version of the present work, it has been shown that, in the same asymptotic regime, the IHS with a fixed SRHT embedding achieves a better convergence rate. Thus, we emphasize that our core contributions are to develop novel techniques and results for analyzing the IHS with the SRHT, as this may be useful for future developments and extensions of this algorithm in different contexts (e.g., constrained least-squares, convex optimization).

Although we characterize the optimal step sizes and momentum parameters for the IHS with Haar embeddings, we only characterize the optimal step size in the absence of momentum for the IHS with the SRHT. It is thus left as an open question to know whether momentum can accelerate further our method.

2 Technical Background

We introduce a few needed definitions, and we refer the reader to [5, 3, 22, 38] for an extensive introduction to random matrix theory. Let $\{M_n\}_n$ be a sequence of Hermitian random matrices, where each M_n has size $n \times n$. For a fixed n, the empirical spectral distribution (e.s.d.) of M_n is the (cumulative) distribution function of its eigenvalues $\lambda_1, \ldots, \lambda_n$, i.e., $F_{M_n}(x) := \frac{1}{n}\sum_{j=1}^n \mathbf{1}\{\lambda_j \leqslant x\}$ for $x \in \mathbb{R}$, which has density $f_{M_n}(x) = \frac{1}{n}\sum_{j=1}^n \delta_{\lambda_j}(x)$ with δ_λ the Dirac measure at λ. Due to the randomness of the eigenvalues, F_{M_n} is random. The relevant aspect of some classes of large $n \times n$ symmetric random matrices M_n is that, almost surely, the e.s.d. F_{M_n} converges weakly towards a non-random distribution F, as $n \to \infty$. This function F, if it exists, will be called the *limiting spectral distribution* (l.s.d.) of M_n.

A powerful tool in the analysis of random matrices is the Stieltjes transform. For μ a probability measure supported on $[0, +\infty)$, its Stieltjes transform is defined over the complex space complementary to the support of μ as

$$m_\mu(z) := \int \frac{1}{x-z}\,\mathrm{d}\mu(x)\,. \tag{4}$$

It holds in particular that m_μ is analytic over $\mathbb{C} \setminus \mathbb{R}_+$, $m_\mu(z) \in \mathbb{C}^+$ for $z \in \mathbb{C}^+$, $m_\mu(z) \in \mathbb{C}^-$ for $z \in \mathbb{C}^-$ and $\mu_\mu(z) > 0$ for $z < 0$, where $\mathbb{R}_+$ is the set of positive reals and $\mathbb{C}^+$ is the set of complex numbers with positive imaginary part. Another useful transform for studying the product of random matrices is the S-transform, denoted S_μ. This is defined as the solution of the following equation, which is unique under certain conditions (see [33]),

$$m_\mu\left(\frac{z+1}{zS_\mu(z)}\right) + zS_\mu(z) = 0. \tag{5}$$

We introduce a few additional concepts from free probability that will be used in the proofs. We refer the reader to [33, 12, 21, 3] for an extensive introduction to this field. Consider the algebra $\mathcal{A}_n$ of $n \times n$ random matrices. For $X_n \in \mathcal{A}_n$, we define the linear functional $\tau_n(X_n) := \frac{1}{n}\mathbb{E}\left[\operatorname{trace} X_n\right]$. Then, we say that a family $\{X_{n,1}, \ldots, X_{n,I}\}$ of random matrices in $\mathcal{A}_n$ is *asymptotically free* if for every $i \in \{1, \ldots, I\}$, $X_{n,i}$ has a limiting spectral distribution, and if $\tau\left(\prod_{j=1}^m P_j\left(X_{n,i_j} - \tau\left(P_j(X_{n,i_j})\right)\right)\right) \to 0$ almost surely for any positive integer m, any polynomials $P_1, \ldots, P_m$ and any indices $i_1, \ldots, i_m \in \{1, \ldots, I\}$ with $i_1 \neq i_2, \ldots, i_{m-1} \neq i_m \neq i_1$. In particular, this definition implies that for two sequences of asymptotically free random matrices X_n, Y_n, we have the *trace decoupling* relation

$$\frac{1}{n}\mathbb{E}\left[\operatorname{trace} X_n Y_n\right] - \frac{1}{n}\mathbb{E}\left[\operatorname{trace} X_n\right]\frac{1}{n}\mathbb{E}\left[\operatorname{trace} Y_n\right] \to 0\,. \tag{6}$$

Essential to our analysis is the following result. If two $n \times n$ random matrices A_n and B_n are asymptotically free and have respective l.s.d. μ_A and μ_B with respective S-transforms S_A and S_B, then the matrix product $A_n B_n$ has l.s.d. μ_{AB} whose S-transform is $S_{AB}(z) = S_A(z)S_B(z)$. The distribution μ_{AB} is called the free multiplicative convolution of μ_A and μ_B, and we denote $\mu_{AB} = \mu_A \boxtimes \mu_B$.

We will also make use of an alternative form of the Stieltjes transform: the η-transform is defined for $z \in \mathbb{C} \setminus \mathbb{R}^-$ as

$$\eta_\mu(z) := \int \frac{1}{1+zx}\,\mathrm{d}\mu(x) = \frac{1}{z}m_\mu\left(-\frac{1}{z}\right)\,. \tag{7}$$

There are standard examples of classes of random matrices and their limiting spectral behavior. We recall a classical result [19]. If S is an $m \times d$ matrix with identically and independently distributed entries $\mathcal{N}(0, 1/m)$, then, as $m, d \to \infty$ with $m/d \to \rho \in (0,1)$, the Marchenko-Pastur theorem (see [19, 5]) states that the matrix $S^\top S$ has l.s.d. F_ρ, whose Stieltjes transform is the unique solution of a certain fixed point equation, and whose density is explicitly given by

$$\mu_\rho(x) = \frac{\sqrt{(b-x)_+(x-a)_+}}{2\pi\rho x}\,, \tag{8}$$

where $y_+ = \max\{0, y\}$, $a = (1-\sqrt{\rho})^2$ and $b = (1+\sqrt{\rho})^2$. In our analysis of Haar and SRHT matrices, we will encounter similar fixed-point equations satisfied by the Stieltjes (or η-) transform of their l.s.d.

3 Sketching with Haar matrices

Sketching matrices with i.i.d. entries are not ideal for sketching. Intuitively, i.i.d. projections distort the geometry of Euclidean space due to their non-orthogonality. In this section, we consider the IHS with refreshed Haar matrices $\{S_t\}$. The following result says that orthogonal projection has better performance than Gaussian projection.

Theorem 3.1 (Optimal IHS with Haar sketches)**.** *With refreshed Haar matrices $\{S_t\}$, step sizes $\mu_t = \theta_{1,h}/\theta_{2,h}$ (where $\theta_{i,h}$ are defined in Lemma 3.2) and momentum parameters $\beta_t = 0$, the sequence of error vectors $\{\Delta_t\}$ satisfies*

$$\rho_h := \left(\lim_{n\to\infty} \frac{\mathbb{E}\|\Delta_t\|^2}{\|\Delta_0\|^2} \right)^{1/t} = \rho_g \cdot \frac{\xi(1-\xi)}{\gamma^2 + \xi - 2\xi\gamma} \,. \tag{9}$$

Further, for any sequence of step sizes $\{\mu_t\}$ and momentum parameters $\{\beta_t\}$, we have that, for the resulting sequence of error vectors $\{\Delta_t\}$,

$$\rho_h \leqslant \liminf_{t\to\infty} \left(\lim_{n\to\infty} \frac{\mathbb{E}\|\Delta_t\|^2}{\|\Delta_0\|^2} \right)^{1/t} , \tag{10}$$

that is, ρ_h is the optimal rate one may achieve using Haar embeddings.

The proof of Theorem 3.1, whose details are deferred to Appendix A.2, is decomposed into two steps. First, we relate the asymptotic convergence rate ρ_h to the first and second moments of the inverse l.s.d. of the sketched matrix SU, and we adapt to the asymptotic setting the proof of Theorem 1 in [13]. Then, and this is our key technical contribution, we provide an explicit formula of this second moment, as given in the following technical lemma.

Lemma 3.2 (First two inverse moments of Haar sketches)**.** *Suppose that S is an $m \times n$ Haar matrix, and let U be an $n \times d$ deterministic matrix with orthonormal columns. It holds that*

$$\begin{aligned} \theta_{1,h} &:= \lim_{n\to\infty} \frac{1}{d} \operatorname{trace} \mathbb{E}\left[(U^\top S^\top S U)^{-1}\right] \;=\; \frac{1-\gamma}{\xi-\gamma} \\ \theta_{2,h} &:= \lim_{n\to\infty} \frac{1}{d} \operatorname{trace} \mathbb{E}\left[(U^\top S^\top S U)^{-2}\right] \;=\; \frac{(1-\gamma)(\gamma^2+\xi-2\gamma\xi)}{(\xi-\gamma)^3} \,. \end{aligned} \tag{11}$$

The formula of the second moment, to the best of our knowledge, is derived explicitly for the first time. We provide a proof sketch here. Note that $\theta_{i,h}$ $(i=1,2)$ is the average of the eigenvalues of $U^\top S^\top S U$ to the power of $-i$. Denoting F_h the limiting distribution of the eigenvalues of $U^\top S^\top S U$, we have $\theta_{i,h} = \int x^{-i} dF_h(x)$. This matrix has a specific structure whose l.s.d. has been studied in the random matrix literature. Specifically, given some diagonal non-negative matrices D, T and a squared Haar matrix W, Theorem 4.11 of [7] characterizes the l.s.d. of matrices of the form $D^{\frac{1}{2}} W T W^\top D^{\frac{1}{2}}$ through a system of functions involving its η-transform and the l.s.d. of D, T. Our setting is more intricate, as S, U are both partial orthogonal matrices, and we need to use an orthogonal complement trick. After getting the η-transform and thus the Stieltjes transform $m(z) = \int \frac{1}{x-z} dF_h(x)$, we can calculate $\theta_{1,h}, \theta_{2,h}$ by evaluating the first and second derivatives of $m(z)$ at 0. Fortunately in our case, the Stieltjes transform has a closed form, though the calculation is cumbersome. We defer the detailed proof to Appendix A.1.

One might wonder how the l.s.d. of Haar matrices and that of Gaussian embeddings – the Marchenko-Pastur law μ_{ρ_g} – differ. Consider the re-scaled matrix $\frac{n}{m} S_{1,n}^\top S_{1,n}$, whose expectation is equal to the identity. Crucially, the l.s.d. μ_{ρ_g} does not depend on the sample size n but only on the limit ratio between d and m, whereas the distribution F_h involves the ratios γ and ξ. Numerically, we observe in Figure 1 that, for fixed $\gamma = 0.2$, as ξ increases, the empirical Haar density departs from the Marchenko-Pastur density μ_{ρ_g}, and concentrates more and more relatively to μ_{ρ_g}. Importantly, we see that the support of F_h is included within the support of μ_{ρ_g}, and thus, more concentrated around 1. According to Theorem 3.1 orthogonal projections are uniformly better than Gaussian i.i.d. projections. Indeed, the ratio between the convergence rates ρ_h and ρ_g is equal to $\xi(1-\xi)/(\gamma^2+\xi-2\gamma\xi)$, and is *always strictly smaller* than 1. To see this, note that $\xi(1-\xi)/(\gamma^2+\xi-2\gamma\xi) < 1$ if and only if $\xi(1-\xi) < \gamma^2+\xi-2\gamma\xi$, and after simplification, we obtain the condition $(\xi-\gamma)^2 > 0$. In the small sketch size regime $d \leqslant m \ll n$, we have $\rho_h/\rho_g \approx 1$. As the sketch size m increases relatively to n, the convergence rates' ratio scales as $\rho_h/\rho_g \approx (1-\xi)$, and one can improve on the number of iterations – and thus, data passes – with Haar embeddings by making $1-\xi$ bounded away from 1. Further, observe that if we do not reduce the size of the original matrix, so that $m = n$ and $\xi = 1$, then the algorithm converges in one iteration. This means that we do not lose any information in the linear model. In contrast, Gaussian projections introduce more distortions than rotation, even though the rows of a Gaussian matrix are almost orthogonal to each other in the high-dimensional setting. The reason is that the eigenvalues are not close to unity.

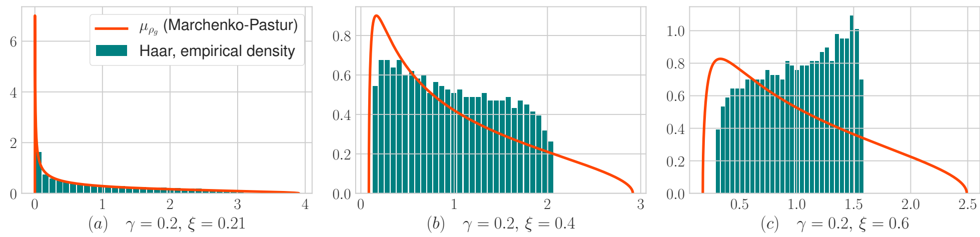

Figure 1: Empirical density of the matrix $\frac{n}{m} U^\top S^\top S U$ for S an $m \times n$ Haar matrix, versus Marchenko-Pastur density with shape parameter d/m. We use $n = 4096$, $d = 820$ and $m \in \{860, 1640, 2450\}$, so that $\gamma \approx 0.2$ and $\xi \in \{0.2, 0.4, 0.6\}$.

Interestingly, momentum does not accelerate the refreshed sketch with Haar embeddings. Leveraging past information through the Heavy-ball update (2) does not provide any benefit, possibly due to the independence between the sketching matrices $\{S_t\}$. Our proof of this fact is actually an immediate consequence of Theorem 2 in [13]. On the other hand, it remains an open question whether there exists a first-order method which uses past iterates along with refreshed matrices, and provide acceleration over gradient descent updates.

We also emphasize that *the optimal parameters have asymptotically closed forms, for any data matrix* A! This is quite unexpected and can be useful in practice. The reason is that random projections introduce a great deal of regularity, leading to a "universal" behavior of certain quantities, including those we need. For methods such as gradient descent with momentum, the optimal parameters (e.g, stepsize, momentum), can depend on quantities that can be nontrivial to estimate (e.g, the Lipschitz constant), and require extra computational work.

However, the time complexity of generating an $m \times n$ Haar matrix using the Gram-Schmidt procedure is $O(nm^2)$, which is, for instance, larger than the classical cost $\mathcal{O}(nd^2)$ for solving the least-squares problem (1), and we now turn to the analysis of another orthogonal matrix, the SRHT, which contains less randomness, but is more structured and faster to generate.

4 Sketching with SRHT matrices

We have seen in the previous section that Haar random projections have a better performance than Gaussian i.i.d. random projections. However, they are still slow to generate and apply. Can we get the same good statistical performance as Haar projections with faster methods? Here we consider the SRHT. This is faster as it relies on the well-structured Walsh-Hadamard transform, which is defined as follows. For an integer $n = 2^p$ with $p \geqslant 1$, the Walsh-Hadamard transform is defined recursively as $H_n = \begin{bmatrix} H_{n/2} & H_{n/2} \\ H_{n/2} & -H_{n/2} \end{bmatrix}$ with $H_1 = 1$. We consider a version of the SRHT which is slightly different from the classical SRHT [1]. Our transform $A \mapsto SA$ first *randomly permutes* the rows of A, before applying the classical transform. This has negligible cost $\mathcal{O}(n)$ compared to the cost $\mathcal{O}(nd \log m)$ of the matrix multiplication $A \mapsto SA$, and *breaks the non-uniformity* in the data. That is, we define the $n \times n$ subsampled randomized Hadamard matrix as $S = BH_n DP/\sqrt{n}$, where B is an $n \times n$ diagonal sampling matrix of i.i.d. Bernoulli random variables with success probability m/n, H_n is the $n \times n$ Walsh-Hadamard matrix, D is an $n \times n$ diagonal matrix of i.i.d. sign random variables, equal to ± 1 with equal probability, and $P \in \mathbb{R}^{n \times n}$ is a uniformly distributed permutation matrix. At the last step, we discard the zero rows of S, so that it becomes an $\widetilde{m} \times n$ orthogonal matrix with $\widetilde{m} \sim \mathrm{Binomial}(m/n, n)$, and the ratio $\widetilde{m}/n$ concentrates fast around ξ as $n \to \infty$. Although the dimension $\widetilde{m}$ is random, we refer to S as an $m \times n$ SRHT matrix.

The following theorem characterizes the exact convergence rate of the IHS with refreshed SRHT embeddings.

Theorem 4.1 (IHS with SRHT sketches)**.** *Suppose that the initial point x_0 is random and that the error vector Δ_0 satisfies the condition $\mathbb{E}\left[\Delta_0 \Delta_0^\top\right] = d^{-1} I_d$. Then, with refreshed SRHT matrices $\{S_t\}$, step sizes $\mu_t = \theta_1^h/\theta_2^h$ and momentum parameters $\beta_t = 0$, the sequence of error vectors $\{\Delta_t\}$*

satisfies

$$\rho_s := \left(\lim_{n\to\infty} \frac{\mathbb{E}\|\Delta_t\|^2}{\mathbb{E}\|\Delta_0\|^2}\right)^{1/t} = \rho_g \cdot \frac{\xi(1-\xi)}{\gamma^2+\xi-2\xi\gamma} = \rho_h\,. \tag{12}$$

Here we impose an additional mild assumption on the initialization of the least-squares problem (1). We note that the initialization condition $\mathbb{E}\left[\Delta_0\Delta_0^\top\right] = d^{-1}I_d$ can be achieved by picking x_0 uniformly on the unit d-sphere $\mathbb{S}^{d-1}$, followed by a uniformly random signed permutation and scaling to the columns of A. The key challenge to avoid this is that we need to evaluate $\mathbb{E}\left[\|\Delta_t\|^2\right] = \operatorname{trace}\mathbb{E}\left[Q_0\dots Q_{t-1}Q_{t-1}\dots Q_0\Delta_0\Delta_0^\top\right]$, where $Q_t = I_d - \mu_t\,(U^\top S_t^\top S_t U)^{-1}$ and U are the left singular vectors of A. Understanding this for general Δ_0 requires properties that are not currently known in random matrix theory (see Appendix A.4 and Remark A.5 for more details). Further we can only analyze the case $\beta_t = 0$, and we do not have a proof for optimality, but *we conjecture that it is true* based on numerical simulations.

We also present an upper-bound on the error, which holds for any deterministic or random initialization x_0 and exhibits an identical convergence rate. This is weaker by a factor of d, but this is negligible for large t.

Theorem 4.2. *For any initialization x_0, with refreshed SRHT matrices $\{S_t\}$, step sizes $\mu_t = \theta_1^h/\theta_2^h$ and momentum parameters $\beta_t = 0$, the sequence of error vectors $\{\Delta_t\}$ satisfies*

$$\limsup_{n\to\infty} \left(\frac{\mathbb{E}\|\Delta_t\|^2}{d\cdot\mathbb{E}\|\Delta_0\|^2}\right)^{1/t} \leq \rho_h\,. \tag{13}$$

The proofs of Theorem 4.1 and 4.2 are deferred to Appendix A.4. While providing significant computational benefits for forming the sketch SA, SRHT embeddings are still able to match the convergence rate of orthogonal projections, and thus, also improves on Gaussian sketches. This result follows from the observation that, althouth SRHT has much less randomness than Haar projection, their first two inverse moments behave the same asymptotically. This is formally stated in the following lemma.

Lemma 4.3 (First two inverse moments of SRHT sketches)**.** *Let S be an $m\times n$ SRHT matrix, S_h be an $m\times n$ Haar matrix, and U an $n\times d$ deterministic matrix with orthonormal columns. Then, the matrices $U^\top S^\top SU$ and $U^\top S_h^\top S_h U$ have the same limiting spectral distribution. Consequently, with $\theta_{1,h}, \theta_{2,h}$ as defined in Lemma 3.2, it holds that*

$$\lim_{n\to\infty} \frac{1}{d}\operatorname{trace}\mathbb{E}\left[(U^\top S^\top SU)^{-1}\right] = \theta_{1,h}\,, \tag{14}$$

$$\lim_{n\to\infty} \frac{1}{d}\operatorname{trace}\mathbb{E}\left[(U^\top S^\top SU)^{-2}\right] = \theta_{2,h}\,. \tag{15}$$

The proof is based on recent results about *asymptotically liberating sequences* from the free probability literature [2], which proves the asymptotic freeness for Hadamard matrices. This technique is also used in [8] to study SRHT. Specifically, they defined the bi-signed-permutation Hadamard matrix $W = P^\top DHDP$, where H is a Hadamard matrix, D is a sign-flipping diagonal matrix, and P is a permutation. Corollary 3.5, 3.7 of [2] showed that the Bernoulli-sampling diagonal matrix B and $WUU^\top W$ are asymptotically free in the non-commutative probability space of random matrices. Another observation is that, by changing the definition of S to $S = BP^\top DHDP = BW$, the l.s.d. of $U^\top S^\top SU$ remain the same as when $S = BHDP$. The asymptotic freeness shows that the l.s.d. of $U^\top S^\top SU$ for S an SRHT is the same as when S is a Haar matrix. So we get the same results as in Lemma 3.2. The detailed proof is defered to Appendix A.3.

In Figure 2, we verify that the empirical densities with Haar and SRHT matrices are indeed very close.

5 Complexity Analysis

Let us now turn to a complexity analysis of the IHS with SRHT embeddings, and compare it, in an asymptotic sense, to the complexity of the standard pre-conditioned conjugate gradient method [26].

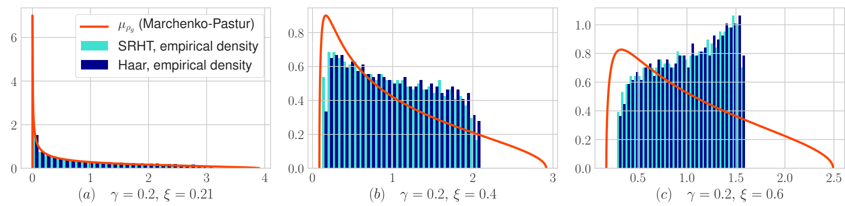

Figure 2: Empirical densities of the matrices $\frac{n}{m} U^\top S^\top S U$ for S an $m \times n$ Haar matrix and SRHT matrix, versus Marchenko-Pastur density with shape parameter d/m. We use $n = 4096$, $d = 820$ and $m \in \{860, 1640, 2450\}$, so that $\gamma \approx 0.2$ and $\xi \in \{0.21, 0.4, 0.6\}$.

The latter uses a sketch SA to compute a pre-conditioning matrix P, such that AP^{-1} has a small condition number, and then it solves the least-squares problem $\min_y \|AP^{-1}y - b\|^2$, using the conjugate-gradient method. As for the IHS, it can be decomposed into three parts: sketching, factoring (computing P and AP^{-1} versus computing H_t), and iterating. The pre-conditioned conjugate gradient prescribes the sketch size $m \approx d \log d$ to guarantee convergence with high-probability. This lower bound is based on the finite-sample bounds on the extremal eigenvalues of the matrix $U^\top S^\top S U$ derived by [30]. Then, given $\varepsilon > 0$ and with $m \approx d \log d$, the resulting complexity to achieve $\|\Delta_t\|^2 \leqslant \varepsilon$ scales as $\mathcal{C}_c \asymp nd \log d + d^3 \log d + nd \log(1/\varepsilon)$, where $nd \log d$ is the cost of forming SA, the term $d^3 \log d$ is the factoring cost, and $nd \log(1/\varepsilon)$ is the per-iteration cost times the number of iterations. In contrast, we obtain that the IHS with the SRHT can use $m \approx d$, with resulting complexity $\mathcal{C}_n \asymp (nd \log d + d^3 + nd) \log(1/\varepsilon)$. Note that the number of iterations multiplies the sum of the sketching, factoring and per-iteration costs, and this is due to refreshing the sketches. Then, treating the term $\log(1/\varepsilon)$ as a constant independent of the dimensions, we find that, as n, d, m grow to infinity, we have that $C_n / C_c \asymp 1/\log d$.

6 Numerical Simulations

6.1 Comparison of the different variants of the iterative Hessian sketch

We evaluate the performance of the IHS with refreshed Haar/SRHT sketches against refreshed Gaussian sketches.

First, we generate a synthetic data matrix $A \in \mathbb{R}^{n \times d}$ with exponential spectral decay (its j-th singular value of A is $\sigma_j = 0.98^j$) and where $n = 8192$ and $d = 800$. We consider the sketch sizes $m \in \{980, 2450, 4100\}$. For the SRHT, we use the step size $\mu_t = \theta_{1,h}/\theta_{2,h}$ prescribed in Theorem 4.1, where we replace ξ and γ by their finite sample approximations $\xi \approx \frac{m}{n}$ and $\gamma \approx \frac{d}{n}$. For refreshed Gaussian embeddings, we use the optimal parameters μ_t and β_t derived in [13]. Results are reported in Figure 3. As m increases, Haar/SRHT embeddings are increasingly better compared to Gaussian projections. Further, the empirical curves match closely our theoretical predictions: the algorithmic parameters derived from our asymptotic analysis are useful in practice when they are replaced by their finite-sample approximations. Second, we carry out a similar experiment with

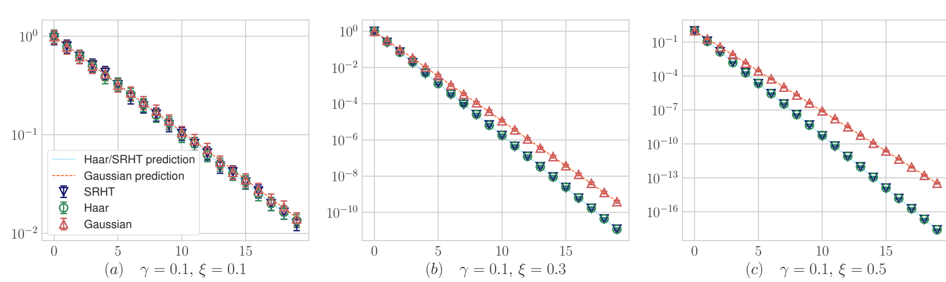

Figure 3: Synthetic dataset: Error $\|\Delta_t\|^2 / \|\Delta_0\|^2$ versus number of iterations for the iterative Hessian sketch: (a) $m = 980$, (b) $m = 2450$ and (c) $m = 4100$. We average over 50 independent trials and empirical standard deviations are shown in the form of error bars.

the CIFAR10 dataset, for which we consider one-vs-all classification. Here, we have $n = 60000$, $d = 3072$ and we use the sketch sizes $m \in \{6000, 18000, 30000\}$. Results are reported in 4, and we observe similar quantitative results as for the aforementioned synthetic dataset.

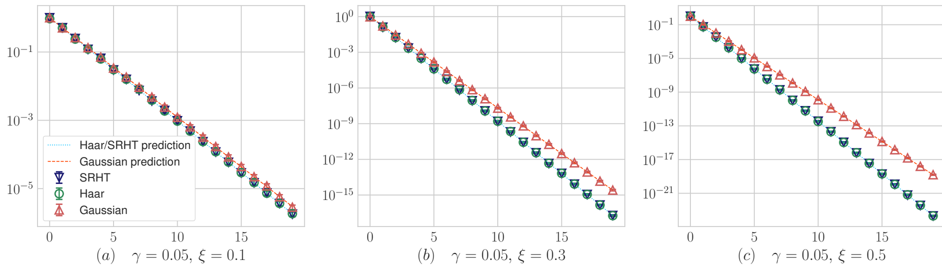

Figure 4: CIFAR10 dataset: Error $\|\Delta_t\|^2/\|\Delta_0\|^2$ versus number of iterations for the iterative Hessian sketch: (a) $m = 6000$, (b) $m = 18000$ and (c) $m = 30000$. We average over 50 independent trials and empirical standard deviations are shown in the form of error bars.

6.2 Comparison of the iterative Hessian sketch to standard iterative solvers

We compare the IHS with the SRHT against the conjugate gradient (CG) method and its preconditioned (pCG) version [26]. We also consider a variant of the IHS, for which we do not refresh the embedding at every iteration. We generate a synthetic data matrix $A \in \mathbb{R}^{n \times d}$ with exponential spectral decay ($\sigma_j = 0.98^j$), $n = 4096$ and $d = 200$. We consider the sketch sizes $m \in \{1000, 1500, 2000\}$. We observe that the IHS which refreshes embeddings at every iteration has the best convergence rate. More generally, the higher this update frequency, the better the performance. In comparison, CG has the worst convergence rate, which is expected since the data matrix is ill-conditioned, and pCG performs slightly worse than the IHS with update frequency equal to 1.

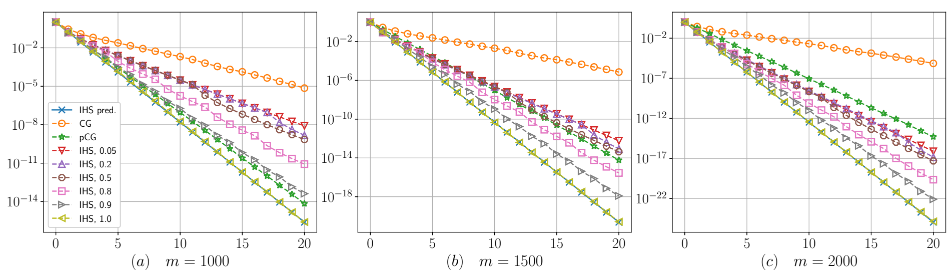

Figure 5: Error $\|\Delta_t\|^2/\|\Delta_0\|^2$ versus number of iterations for the iterative Hessian sketch with the SRHT and different sketch sizes. We average over 50 independent trials. For instance, 'IHS, 0.2' refers to the IHS with update frequency equal to 0.2. For clarity, we do not show error bars for the mean empirical standard deviation which are barely visible.

Acknowledgments and Disclosure of Funding

This work was partially supported by the National Science Foundation under grants IIS-1838179 and ECCS-2037304, Facebook Research, Adobe Research and Stanford SystemX Alliance.

References

[1] N. Ailon and B. Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563. ACM, 2006.

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

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

[4] H. Avron, P. Maymounkov, and S. Toledo. Blendenpik: Supercharging lapack's least-squares solver. *SIAM Journal on Scientific Computing*, 32(3):1217–1236, 2010.

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

[6] K. M. Choromanski, M. Rowland, and A. Weller. The unreasonable effectiveness of structured random orthogonal embeddings. In *Advances in Neural Information Processing Systems*, pages 219–228, 2017.

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

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

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

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

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

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

[13] J. Lacotte and M. Pilanci. Faster least squares optimization. *arXiv preprint arXiv:1911.02675*, 2019.

[14] J. Lacotte and M. Pilanci. Effective dimension adaptive sketching methods for faster regularized least-squares optimization. *arXiv preprint arXiv:2006.05874*, 2020.

[15] J. Lacotte and M. Pilanci. Optimal randomized first-order methods for least-squares problems. *arXiv preprint arXiv:2002.09488*, 2020.

[16] J. Lacotte, M. Pilanci, and M. Pavone. High-dimensional optimization in adaptive random subspaces. In *Advances in Neural Information Processing Systems*, pages 10846–10856, 2019.

[17] M. W. Mahoney. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

[18] M. W. Mahoney and P. Drineas. Structural properties underlying high-quality randomized numerical linear algebra algorithms., 2016.

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

[20] X. Meng, M. A. Saunders, and M. W. Mahoney. Lsrn: A parallel iterative solver for strongly over-or underdetermined systems. *SIAM Journal on Scientific Computing*, 36(2):C95–C118, 2014.

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

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

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

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

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

[26] V. Rokhlin and M. Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212–13217, 2008.

[27] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 143–152. IEEE, 2006.

[28] S. Sridhar, M. Pilanci, and A. Özgür. Lower bounds and a near-optimal shrinkage estimator for least squares using random projections. *arXiv preprint arXiv:2006.08160*, 2020.

[29] T. Tao. *Topics in Random Matrix Theory*, volume 132. American Mathematical Soc., 2012.

[30] J. A. Tropp. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(01n02):115–126, 2011.

[31] A. M. Tulino, G. Caire, S. Shamai, and S. Verdú. Capacity of channels with frequency-selective and time-selective fading. *IEEE Transactions on Information Theory*, 56(3):1187–1215, 2010.

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

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

[34] R. Witten and E. Candes. Randomized algorithms for low-rank matrix factorizations: sharp performance bounds. *Algorithmica*, 72(1):264–281, 2015.

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

[36] F. Yang, S. Liu, E. Dobriban, and D. P. Woodruff. How to reduce dimension with pca and random projections? *arXiv preprint arXiv:2005.00511*, 2020.

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

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

A Proofs of main theorems

A.1 Calculations of $\theta_{1,h}$ and $\theta_{2,h}$ for Haar sketch

We first prove some lemmas and provide the proof of 3.2 in Section A.1.1.

This lemma characterizes the Stieltjes transform of the l.s.d. of S_nU_n.

Lemma A.1 (Stieltjes transform of l.s.d. of S_nU_n)**.** *We set $S_{1,n} = S_nU_n$. Then the matrix $S_{1,n}^\top S_{1,n}$ admits a l.s.d. whose Stieltjes transform m_h is given by*

$$m_h(z) = \frac{z(2\gamma-1)+\xi-\gamma-\sqrt{(\gamma+\xi-2+z)^2+4(z-1)(1-\gamma)(1-\xi)}}{2\gamma z(1-z)}\,, \tag{16}$$

for any $z \in \mathbb{C} \setminus \mathbb{R}_+$.

Proof. First, observe that since both S_n and U_n are rectangular orthogonal matrices, we can embed them into full orthogonal matrices as $\mathbb{S}_n = \begin{pmatrix} S_n \\ S_n^\perp \end{pmatrix}$ and $\mathbb{U}_n = \begin{pmatrix} U_n & U_n^\perp \end{pmatrix}$. Then, we can write

$$S_{1,n} = \begin{pmatrix} I_m & 0 \end{pmatrix} \mathbb{S}_n\mathbb{U}_n \begin{pmatrix} I_d \\ 0 \end{pmatrix}. \tag{17}$$

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

$$C_n := \begin{pmatrix} S_{1,n}S_{1,n}^\top & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix} \mathbb{W}_n \begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix} \mathbb{W}_n^\top \begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix}. \tag{18}$$

The matrix C_n is related to our matrix of interest $S_{1,n}^\top S_{1,n}$, as they have exactly the same non-zero eigenvalues. Thus, as a first step to establish Lemma A.1, we characterize the l.s.d. of C_n.

The matrix C_n admits a l.s.d. F_C, whose Stieltjes transform m_C is given by

$$m_C(z) = \frac{z+\gamma+\xi-2-\sqrt{(\gamma+\xi-2+z)^2+4(z-1)(1-\gamma)(1-\xi)}}{2z(1-z)}\,, \tag{19}$$

for any $z \in \mathbb{C} \setminus \mathbb{R}_+$. The above expression (18) of the matrix C_n has the required form to apply Theorem 4.11 by [7], and hence characterize the e.s.d. of C_n through its η-transform which has to satisfy a fixed-point equation. We defer details of the proof to Section B.2. Now, we use the fact that the matrices $S_{1,n}^\top S_{1,n}$ and C_n have the same non-zero eigenvalues. Almost surely, there are exactly d of them, which we denote $\lambda_1, \ldots, \lambda_d$. Then, the e.s.d. F_{C_n} of C_n can be decomposed as

$$F_{C_n}(x) = \left(1-\frac{d}{n}\right)\mathbf{1}_{\{x\geqslant 0\}} + \frac{1}{n}\sum_{i=1}^{d}\mathbf{1}_{\{x\geqslant\lambda_i\}} = \left(1-\frac{d}{n}\right)\mathbf{1}_{\{x\geqslant 0\}} + \frac{d}{n}F_{h,n}(x)\,, \tag{20}$$

where $F_{h,n}$ is the e.s.d. of $S_{1,n}^\top S_{1,n}$. Taking the limit $n \to \infty$, we find that $F_{1,n}$ converges weakly almost surely to

$$F_h(x) = \frac{1}{\gamma}\left(F_C(x) - (1-\gamma)\mathbf{1}_{\{x\geqslant 0\}}\right)\,. \tag{21}$$

By definition of m_h and using (21), it follows that for $z \in \mathbb{C} \setminus \mathbb{R}_+$

$$m_h(z) = \int \frac{1}{x-z}\,\mathrm{d}F_h(x) = \frac{1}{\gamma}\int \frac{1}{x-z}\,\mathrm{d}F_C(x) - \frac{1-\gamma}{\gamma}\int \frac{1}{x-z}\,\delta_0(x)\mathrm{d}x \tag{22}$$

$$= \frac{1}{\gamma}m_C(z) + \frac{1-\gamma}{\gamma z}\,. \tag{23}$$

Plugging-in the expression of m_C, we obtain the claimed formula (16) for m_h.

□

We will need the following result regarding the support of F_h, which is proved in Appendix B.1.

Lemma A.2. *The support of F_h satisfies*

$$\inf \operatorname{supp}(F_h) \geqslant \frac{(1-\sqrt{\rho_g})^2}{\left(1+\frac{1}{\sqrt{\xi}}\right)^2}\,. \tag{24}$$

Thus, the support of F_h is bounded away from 0, so is the intersection of the support of F_C and $\mathbb{R}^*$. Further, the distribution F_C has a point mass at 0 equal to $1-\gamma$. We now turn to the trace calculations in Lemma 3.2.

A.1.1 Proof of Lemma 3.2

1. **Computing** $\theta_{1,h}$

   Using the facts that F_C has support within $[0, +\infty)$ and a point mass equal to $(1-\gamma)$ at 0, its η-transform η_C is well-defined on $\{z \in \mathbb{R} \mid z > 0\}$, and, for $z > 0$, it can be decomposed as

   $$\eta_C(z) = 1-\gamma + \int_{x\neq 0} \frac{1}{1+zx}\mathrm{d}F_C(x)\,. \tag{25}$$

   The function $\frac{1}{x}$ is integrable on the set $\{x > 0\}$ with respect to F_C, since the support of F_C on $\mathbb{R}^*$ is bounded away from 0. Since $|\frac{z}{1+xz}| < \frac{1}{x}$ when $z > 0, x > 0$, it follows by the dominated convergence theorem that

   $$\lim_{z\to\infty}\int_{x\neq 0}\frac{z}{1+xz}\mathrm{d}F_C(x) = \int_{x\neq 0}\lim_{z\to\infty}\frac{z}{1+xz}\mathrm{d}F_C(x) = \int_{x\neq 0}\frac{1}{x}\mathrm{d}F_C(x)\,. \tag{26}$$

   Using (25), it follows that

   $$\lim_{z\to\infty} z\left(\eta_C(z) - (1-\gamma)\right) = \int_{x\neq 0}\frac{1}{x}\,\mathrm{d}F_C(x)\,, \tag{27}$$

On the other hand, we have that

$$\lim_{z\to\infty} \eta_C(z) = (1-\gamma) + \lim_{z\to\infty} \int_{x\neq 0} \frac{1}{1+zx}\,\mathrm{d}F_C(t) \tag{28}$$

$$= (1-\gamma) + \int_{x\neq 0} \lim_{z\to\infty} \frac{1}{1+zx}\,\mathrm{d}F_C(x) \tag{29}$$

$$= 1-\gamma\,. \tag{30}$$

where the second equality is again justified by the dominated convergence theorem. Subtracting $1-\gamma$ from both sides of (48), multiplying by $z\left(1+\frac{\xi-1}{\eta_C(z)}\right)$ and letting $z\to\infty$, we obtain

$$\lim_{z\to\infty} z\left(1+\frac{\xi-1}{\eta_C(z)}\right)(\eta_C(z)-(1-\gamma)) = \lim_{z\to\infty} z\left(1+\frac{\xi-1}{\eta_C(z)}\right)\left(\frac{\gamma}{1+z(1+\frac{\xi-1}{\eta_C(z)})}\right)\,.$$

Note that the right-hand side of the above equation is equal to γ, and the left-hand side satisfies

$$\begin{aligned}\lim_{z\to\infty} z\left(1+\frac{\xi-1}{\eta_C(z)}\right)(\eta_C(z)-(1-\gamma)) &= \lim_{z\to\infty} z\,(\eta_C(z)-(1-\gamma))\left(1+\frac{\xi-1}{1-\gamma}\right)\\ &= \frac{\xi-\gamma}{1-\gamma}\cdot\int_{x\neq 0}\frac{1}{x}\,\mathrm{d}F_C(x),\end{aligned}$$

where we used (27) and (30). This shows that $\gamma = \frac{\xi-\gamma}{1-\gamma}\int_{x\neq 0}\frac{1}{x}\,\mathrm{d}F_C(x)$. We conclude by observing that

$$\theta_{1,h} = \lim_{n\to\infty}\frac{1}{d}\operatorname{trace}\mathbb{E}\left[(S_{1,n}^\top S_{1,n})^{-1}\right] = \frac{1}{\gamma}\cdot\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{d}\frac{1}{\lambda_i}\right] = \frac{1}{\gamma}\int_{x\neq 0}\frac{1}{x}\,\mathrm{d}F_C(x)\,,$$

and consequently, $\theta_{1,h} = \frac{1-\gamma}{\xi-\gamma}$, which is the claimed result.

2. **Computing** $\theta_{2,h}$

Unrolling its definition, we have that

$$\theta_{2,h} = \lim_{n\to\infty}\frac{1}{d}\operatorname{trace}\mathbb{E}\left[(S_{1,n}^\top S_{1,n})^{-2}\right] = \frac{1}{\gamma}\cdot\lim_{n\to\infty}\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{d}\frac{1}{\lambda_i^2}\right] = \frac{1}{\gamma}\int_{\{x\neq 0\}}\frac{1}{x^2}\,\mathrm{d}F_C(x)\,,$$

where the limit in the third equation holds and is finite since F_C has support bounded away from 0 on $\mathbb{R}^*$. By definition of m_C and using the fact that F_C has point mass $1-\gamma$ at 0, we get that

$$\frac{\mathrm{d}m_C(z)}{\mathrm{d}z} = \int\frac{1}{(x-z)^2}\,\mathrm{d}F_C(x) = \frac{1-\gamma}{z^2} + \int_{\{x\neq 0\}}\frac{1}{(x-z)^2}\,\mathrm{d}F_C(x)\,.$$

Using again the fact that F_C has support bounded away from 0 on $\mathbb{R}^*$ and the dominated convergence theorem, we have that $\gamma\theta_{2,h} = \lim_{z\to 0}\int_{x\neq 0}\frac{1}{(x-z)^2}\,\mathrm{d}F_C(x)$, and thus,

$$\gamma\theta_{2,h} = \lim_{z\to 0}\left\{\frac{\mathrm{d}m_C(z)}{\mathrm{d}z} - \frac{1-\gamma}{z^2}\right\}\,.$$

We denote

$$\begin{aligned}\triangle &:= (\gamma+\xi-2+z)^2 + 4(z-1)(1-\gamma)(1-\xi)\,,\\ \triangle' &:= \frac{\mathrm{d}\triangle}{\mathrm{d}z} = 2(z+\gamma+\xi-2) + 4(1-\gamma)(1-\xi)\,.\end{aligned}$$

Then, using the expression (19) of m_C and taking the derivative, it follows that

$$\frac{\mathrm{d}m_C(z)}{\mathrm{d}z} - \frac{1-\gamma}{z^2} = \frac{1-\frac{1}{2\sqrt{\triangle}}(2(z+\gamma+\xi-2)+4(1-\gamma)(1-\xi))}{2z(1-z)} \tag{31}$$

$$+ \frac{(z+\gamma+\xi-2-\sqrt{\triangle})(2z-1)}{2z^2(z-1)^2} + \frac{\gamma-1}{z^2} \tag{32}$$

$$= \frac{1}{2z^2(z-1)^2}[\triangle_1 + (2\gamma\xi-\gamma-\xi)\triangle_2 - \triangle_3 + \triangle_4], \tag{33}$$

where

$$\begin{cases} \triangle_1 = \frac{z^2(z-1)}{\sqrt{\triangle}} \\ \triangle_2 = \frac{z(z-1)}{\sqrt{\triangle}} \\ \triangle_3 = (2z-1)\sqrt{\triangle} \\ \triangle_4 = z(1-z) + (z+\gamma+\xi-2)(2z-1) + 2(\gamma-1)(z-1)^2. \end{cases}$$

According to L'Hospital rule,

$$\gamma\theta_{2,h} = \lim_{z\to 0} \frac{\triangle_1'' + (2\gamma\xi-\gamma-\xi)\triangle_2'' - \triangle_3'' + \triangle_4''}{2(12z^2-12z+2)} = \lim_{z\to 0} \frac{\triangle_1'' + (2\gamma\xi-\gamma-\xi)\triangle_2'' - \triangle_3'' + \triangle_4''}{4}, \tag{34}$$

where $\triangle_i''$ denotes the second derivative of $\triangle_i$ with respect to z. After some calculations, we find that

$$\begin{aligned} \triangle_1''|_{z=0} &= -\frac{2}{\xi-\gamma}, \\ \triangle_2''|_{z=0} &= \frac{2}{\xi-\gamma} + \frac{4\gamma\xi-2\gamma-2\xi}{(\xi-\gamma)^3}, \\ \triangle_3''|_{z=0} &= \frac{4(2\gamma\xi-\gamma-\xi)-1}{\xi-\gamma} + \frac{(2\gamma\xi-\gamma-\xi)^2}{(\xi-\gamma)^3}, \\ \triangle_4''|_{z=0} &= 2(2\gamma-1). \end{aligned}$$

Using (34), it follows that

$$\gamma\theta_{2,h} = \frac{1}{4}\left(\frac{-(2\gamma-1)^2}{\xi-\gamma} + \frac{(2\gamma\xi-\gamma-\xi)^2}{(\xi-\gamma)^3}\right) = \frac{\gamma(1-\gamma)(\gamma^2+\xi-2\gamma\xi)}{(\xi-\gamma)^3},$$

and finally, we obtain the claimed expression, that is, $\theta_{2,h} = \frac{(1-\gamma)(\gamma^2+\xi-2\gamma\xi)}{(\xi-\gamma)^3}$.

A.2 Proof of Theorem 3.1

Proof. Let $\{S_t\}$ be a sequence of independent $m \times n$ Haar matrices, and let $\{x_t\}$ be the sequence of iterates generated by the update (2) with $\mu_t = \theta_{1,h}/\theta_{2,h}$ and $\beta_t = 0$. Recall that we denote $\Delta_t = U^\top A(x_t - x^*)$, where $A = U\Sigma V^\top$ is a thin singular value decomposition of A. For $t \geqslant 0$, we have that

$$\begin{aligned} A\left(A^\top S^\top S A\right)^{-1} A^\top &= U\Sigma V^\top \left(V\Sigma U^\top S^\top S U \Sigma V^\top\right)^{-1} V\Sigma U^\top \\ &= U\Sigma V^\top V \Sigma^{-1}(U^\top S^\top S U)^{-1}\Sigma^{-1} V V^\top \Sigma U^\top \\ &= U(U^\top S^\top S U)^{-1} U^\top \end{aligned}$$

Multiplying both sides of the update formula (2) by A, subtracting Ax^* and using the normal equation $A^\top A x^* = A^\top b$, we find that

$$A(x_{t+1} - x^*) = \left(I_n - \mu_t U(U^\top S_t^\top S_t U)^{-1} U^\top\right) A(x_t - x^*). \tag{35}$$

Multiplying both sides of (35) by $U^\top$, using the definition of Δ_t and the fact that $U^\top U = I_d$, it follows that

$$\begin{aligned} \Delta_{t+1} &= U^\top \left(I_n - \mu_t U(U^\top S_t^\top S_t U)^{-1} U^\top\right) A(x_t - x^*) \\ &= \left(U^\top - \mu_t U^\top U(U^\top S_t^\top S_t U)^{-1} U^\top\right)(Ax_t - x^*) \\ &= \left(I_d - \mu_t (U^\top S_t^\top S_t U)^{-1}\right)\Delta_t, \end{aligned}$$

and then, taking the squared norm,

$$\|\Delta_{t+1}\|^2 = \Delta_t^\top \left(I_d - \mu_t (U^\top S_t^\top S_t U)^{-1}\right)^2 \Delta_t.$$

Taking the expectation with respect to S_t and using the independence of S_t with respect to $S_0, \dots, S_{t-1}$, we obtain that

$$\mathbb{E}_{S_t}\left[\|\Delta_{t+1}\|^2\right] = \Delta_t^\top \mathbb{E}\left[\left(I_d - \mu_t (U^\top S_t^\top S_t U)^{-1}\right)^2\right] \Delta_t \tag{36}$$

$$= \Delta_t^\top \Big(I_d - 2\mu_t \,\mathbb{E}\left[(U^\top S_t^\top S_t U)^{-1}\right] + \mu_t^2 \,\mathbb{E}\left[(U^\top S_t^\top S_t U)^{-2}\right]\Big)\Delta_t\,. \tag{37}$$

We write the spectral decomposition $U^\top S_t^\top S_t U = V \Sigma V^\top$ where Σ is diagonal with positive entries $\lambda_1, \dots, \lambda_d$ and $V_t = [v_1, \dots, v_d]$ is a $d \times d$ orthogonal matrix. The matrix $S_t U$ is distributed as the $m \times d$ upper-left block of an $n \times n$ Haar matrix. Therefore, $S_t U$ is right rotationally invariant, and so is the matrix V. It follows that $\lambda_i v_{ik} v_{i\ell} \overset{\mathrm{d}}{=} -\lambda_i v_{ik} v_{i\ell}$ for any index i and any indices $k \neq \ell$. Then, for any $p \in \{1, 2\}$ and any $k \neq \ell$, we have

$$\mathbb{E}\left[\left((U^\top S^\top S U)^{-p}\right)_{k\ell}\right] = \sum_{i=1}^d \mathbb{E}\left[\lambda_i^{-p} v_{ik} v_{i\ell}\right] = -\sum_{i=1}^d \mathbb{E}\left[\lambda_i^{-p} v_{ik} v_{i\ell}\right]\,,$$

which implies that the off-diagonal term $\mathbb{E}\left[\left((U^\top S^\top S U)^{-p}\right)_{k\ell}\right]$ is equal to 0. Further, by permutation invariance of the matrix V, we get that for any k,

$$\mathbb{E}\left[\left((U^\top S^\top S U)^{-p}\right)_{kk}\right] = \frac{1}{d}\,\mathrm{trace}\,\mathbb{E}\left[(U^\top S^\top S U)^{-p}\right]\,,$$

or equivalently, $\mathbb{E}\left[(U^\top S^\top S U)^{-p}\right] = \theta_{p,n} I_d$ where $\theta_{p,n} := d^{-1}\,\mathrm{trace}\,\mathbb{E}\left[(U^\top S^\top S U)^{-p}\right]$. Then, using (37), it follows that

$$\begin{aligned}
\mathbb{E}_{S_t}\left[\|\Delta_{t+1}\|^2\right] &= \Delta_t^\top \Big(I_d - 2\mu_t\, \theta_{1,n} I_d + \mu_t^2\, \theta_{2,n} I_d\Big)\Delta_t \\
&= (1 - 2\mu_t \theta_{1,n} + \mu_t^2 \theta_{2,n}) \cdot \|\Delta_t\|^2 \\
&= \left(1 - \frac{{\theta_{1,n}}^2}{\theta_{2,n}} + \left(\frac{\theta_{1,n}}{\sqrt{\theta_{2,n}}} - \mu_t \sqrt{\theta_{2,n}}\right)^2\right) \cdot \|\Delta_t\|^2\,.
\end{aligned}$$

By induction, we further obtain

$$\frac{\mathbb{E}\left[\|\Delta_t\|^2\right]}{\|\Delta_0\|^2} = \prod_{j=0}^{t-1}\left(1 - \frac{{\theta_{1,n}}^2}{\theta_{2,n}} + \left(\frac{\theta_{1,n}}{\sqrt{\theta_{2,n}}} - \mu_j \sqrt{\theta_{2,n}}\right)^2\right)\,.$$

Taking the limit $n \to \infty$ and using the definition $\theta_{h,p} = \lim_{n\to\infty} \theta_{p,n}$ for $p \in \{1, 2\}$, we find that

$$\lim_{n\to\infty} \frac{\mathbb{E}\left[\|\Delta_t\|^2\right]}{\|\Delta_0\|^2} = \prod_{j=0}^{t-1}\left(1 - \frac{{\theta_{1,h}}^2}{\theta_{2,h}} + \left(\frac{\theta_{1,h}}{\sqrt{\theta_{2,h}}} - \mu_j \sqrt{\theta_{2,h}}\right)^2\right)\,.$$

The above right-hand side is minimized at $\mu_j = \theta_{1,h}/\theta_{2,h}$ for all times steps $j \geqslant 0$, which yields the error formula

$$\lim_{n\to\infty} \frac{\mathbb{E}\left[\|\Delta_t\|^2\right]}{\|\Delta_0\|^2} = \left(1 - \frac{{\theta_{1,h}}^2}{\theta_{2,h}}\right)^t\,.$$

Plugging-in the expressions of $\theta_{1,h}$ and $\theta_{2,h}$, we obtain the claimed convergence rate ρ_h.

It remains to prove that ρ_h is the best rate one may achieve with the update (2) along with Haar embeddings. It is actually an immediate consequence of Theorem 2 in [13] whose assumptions (precisely, Assumption 1 in [13]) are trivially satisfied by Haar embeddings.

□

A.3 Calculations of $\theta_{1,h}$ and $\theta_{2,h}$ for SRHT

Our analysis proceeds in a way similar to the analysis of the Haar case, and we describe in this paragraph the main steps. Denote by F_S the l.s.d. of $U^\top S^\top S U$ and by $F_{S,n}$ its e.s.d. As we did for the Haar case with the matrix C_n, we introduce here an auxiliary matrix G_n whose e.s.d. is related to $F_{S,n}$. Then, we characterize the η-transform η_G of its l.s.d. F_G. Our analysis for η_G uses recent results on *asymptotically liberating sequences* from free probability [2]. This technique has also been used in the prior work [8]. Finally, we show that η_G is equal to the η-transform η_C of F_C, and we conclude that $F_S = F_h$.

Let $S = BH_nDP$ be the $n \times n$ SRHT matrix (before discarding the rows) as defined in Section 4 in the paper, and U be an $n \times d$ deterministic matrix with orthonormal columns. Note that whether we consider the zero rows or not in the matrix S, the matrix $U^\top S^\top SU$ remains the same, and so does its l.s.d. The matrices B, H_n and D are all symmetric matrices, and they respectively satisfy $B^2 = B$, $H_n^2 = I_n$ and $D^2 = I_n$, and P is also an orthogonal matrix. Then, we have that $S^\top S = P^\top DH_nBH_nDP$, and further,

$$(S^\top S)^2 = P^\top DH_nBH_nDPP^\top DH_nBH_nDP = P^\top DH_nBH_nDP = S^\top S\,.$$

We first have the following observation, whose proof is deferred to Appendix B.3.

Lemma A.3. *For P, B, D, H_n and U defined as above, we have the following equality in distribution*

$$U^\top(P^\top DH_n)B(HDP)U \overset{\mathrm{d}}{=} U^\top(P^\top DH_nDP)B(P^\top DH_nDP)U\,. \tag{38}$$

We now proceed with asymptotic statements, and we introduce the subscript n to all matrices. We set $W_n := P_n^\top D_nH_nD_nP_n$. It holds that the matrix $U_n^\top W_nB_nW_nU_n$ has the same nonzero eigenvalues as $G_n := B_nW_nU_nU_n^\top W_nB_n$, so that we first find the l.s.d. of the matrix G_n. The reader may notice that G_n plays a similar role in the analysis of the SRHT case, to that of the matrix C_n in the analysis of the Haar case.

The following result states the asymptotic freeness of the matrices B_n and $W_nU_nU_n^\top W_n$. Its proof follows directly from Corollaries 3.5 and 3.7 by [2].

Lemma A.4. *Let B_n, W_n, U_n be defined as above. Then, the matrices $\{B_n, W_nU_nU_n^\top W_n\}$ are asymptotically free in the limit of the non-commutative probability spaces of random matrices. Consequently, the e.s.d. of the matrix $G_n = B_nW_nU_nU_n^\top W_nB_n$ converges to the freely multiplicative convolution of the l.s.d. F_B of B_n and the l.s.d. F_U of $U_nU_n^\top$, that is, G_n has l.s.d. given by $F_G = F_B \boxtimes F_U$.*

Since the density of the l.s.d. F_B is $f_B = \xi\delta_1 + (1-\xi)\delta_0$ and and the density of F_U is $f_U = \gamma\delta_1 + (1-\gamma)\delta_0$, we have that the S-transforms S_B of F_B and S_U of F_U are respectively equal to $S_B(y) = \frac{y+1}{y+\xi}$ and $S_U(y) = \frac{y+1}{y+\gamma}$. From Lemma A.4, it follows that the S-transform S_G of F_G is the product of S_B and S_U, i.e.,

$$S_G(y) = S_U(y)S_B(y) = \frac{(y+1)^2}{(y+\xi)(y+\gamma)}\,. \tag{39}$$

First, note that using their respective definitions, the S-transform of F_G and its η-transform η_G are related by the equation $\eta_G\left(-\frac{y}{y+1}S_G(y)\right) = y+1$. Plugging-in the expression (39) of $S_G(y)$ into the latter equation, we obtain that

$$\eta_G\left(-\frac{y(y+1)}{(y+\gamma)(y+\xi)}\right) = y+1\,.$$

Letting $z = -\frac{(y+\gamma)(y+\xi)}{y(y+1)}$ and using the relationship (7) between the Stieltjes and η-transforms, we find that the Stieltjes transform m_G of G is equal to

$$m_G(z) = \frac{z+\gamma+\xi-2-\sqrt{g(z)}}{2z(1-z)}\,,$$

where $g(z) = (\gamma+\xi-2+z)^2 + 4(z-1)(1-\gamma)(1-\xi)$. Hence, we get that $m_G(z) = m_C(z)$, that is, $F_G = F_C$.

Further, the matrix G_n has the same non-zero eigenvalues as the matrix $U_n^\top W_nB_nW_nU_n$ which, according to Lemma A.3, is equal in distribution to $U_n^\top S_n^\top S_nU_n$. Denote by $\lambda_1, \ldots, \lambda_{\widetilde{d}}$ the non-zero eigenvalues of $U_n^\top S_n^\top S_n U_n$, where $\widetilde{d}$ is itself a random number due to the randomness of non-zero rows $\widetilde{m}$. Hence, the e.s.d $F_{G,n}$ of G_n and the e.s.d. $F_{S,n}$ of $U_n^\top S_n^\top S_n U_n$ satisfy (see Appendix B.4)

$$F_{G_n}(x) \overset{\mathrm{d}}{=} \left(1 - \frac{d}{n}\right) \mathbf{1}_{\{x \geqslant 0\}} + \frac{d}{n} F_{S,n}(x)\,. \tag{40}$$

Thus, we obtain that $F_{S,n}$ converges weakly almost surely to the distribution

$$F_S(x) := \frac{1}{\gamma}\left(F_G(x) - (1-\gamma)\mathbf{1}_{\{x \geqslant 0\}}\right) = \frac{1}{\gamma}\left(F_C(x) - (1-\gamma)\mathbf{1}_{\{x \geqslant 0\}}\right)\,. \tag{41}$$

The latter expression is equal to $F_h(x)$ according to (21), so that $F_S(x) = F_h(x)$. The analysis of the traces of the expected first and second inverse moments only involves the limiting distribution (we refer the reader to the proof of the expressions of $\theta_{1,h}$ and $\theta_{2,h}$, in Section A.1). Due to the equality $F_h = F_S$, they remain the same with SRHT matrices, which concludes the proof of Lemma 4.3.

A.4 Proof of Theorem 4.1 and 4.2

Let $\{S_t\}$ be a sequence of independent $m \times n$ SRHT matrices, and let $\{x_t\}$ be the sequence of iterates generated by the update (2) with $\mu_t = \theta_{1,h}/\theta_{2,h}$ and $\beta_t = 0$. Denote $\Delta_t = U^\top A(x_t - x^*)$ the sequence of error vectors. The proof follows exactly the same lines as for Theorem 4.1 up to the relationship (37), which we recall here,

$$\mathbb{E}_{S_t}\left[\|\Delta_{t+1}\|^2\right] = \mathbb{E}_{S_t}\left[\Delta_t^\top \left(I_d - \mu_t\,(U^\top S_t^\top S_t U)^{-1}\right)^2 \Delta_t\right]\,. \tag{42}$$

Denote $Q_t = I_d - \mu_t\,(U^\top S_t^\top S_t U)^{-1}$. It holds that $\Delta_{t+1} = Q_t \Delta_t$ as previously shown. Hence, by induction, we obtain that

$$\mathbb{E}\left[\|\Delta_t\|^2\right] = \operatorname{trace} \mathbb{E}\left[Q_0 \dots Q_{t-1} Q_{t-1} \dots Q_0 \Delta_0 \Delta_0^\top\right]\,. \tag{43}$$

Using the independence of Δ_0 and the Q_i, and the assumption $\mathbb{E}\left[\Delta_0\Delta_0^\top\right] = I_d/d$, it follows that

$$\mathbb{E}\left[\|\Delta_t\|^2\right] = \frac{1}{d}\operatorname{trace} \mathbb{E}\left[Q_1 \dots Q_{t-1} Q_{t-1} \dots Q_0^2\right]\,. \tag{44}$$

It holds that the matrix Q_0^2 is asymptotically free from $Q_{t-1} \dots Q_1$. Therefore, using the trace decoupling relation (6), we have that

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[\|\Delta_t\|^2\right] &= \lim_{n\to\infty} \frac{1}{d}\operatorname{trace} \mathbb{E}\left[Q_1 \dots Q_{t-1} Q_{t-1} \dots Q_0^2\right] \\
&= \lim_{n\to\infty} \frac{1}{d}\operatorname{trace} \mathbb{E}\left[Q_0^2\right] \cdot \lim_{n\to\infty} \frac{1}{d}\operatorname{trace} \mathbb{E}\left[Q_2 \dots Q_{t-1} Q_{t-1} \cdots Q_1^2\right]\,.
\end{aligned}$$

Note that $\lim_{n\to\infty} \frac{1}{d}\operatorname{trace} \mathbb{E}\left[Q_0^2\right] = (1 - 2\mu_0\theta_{1,h} + \mu_0^2\theta_{2,h})$. Repeating the same asymptotic freeness argument between Q_1^2 and $Q_{t-1}\dots Q_2$ and plugging-in $\mu_j = \theta_{1,h}/\theta_{2,h}$, we finally obtain the claimed result,

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[\|\Delta_{t+1}\|^2\right] &= \prod_{j=0}^{t-1}\left(1 - \mu_j\theta_{1,h} + \mu_j^2\theta_{2,h}\right) \\
&= \left(1 - \frac{\theta_{1,h}^2}{\theta_{2,h}}\right)^t\,.
\end{aligned}$$

The proof of Theorem 4.2 immediately follows from an alternative upper-bound on the expression (43) for the norm of the error. In particular, we note that

$$\begin{aligned}
\mathbb{E}\left[\|\Delta_t\|^2\right] &= \operatorname{trace} \mathbb{E}\left[Q_0 \dots Q_{t-1} Q_{t-1} \dots Q_0 \Delta_0 \Delta_0^\top\right] \\
&\leq \|\Delta_0\Delta_0^\top\|_2 \operatorname{trace} \mathbb{E}\left[Q_0 \dots Q_{t-1} Q_{t-1} \dots Q_0\right] \\
&= d\|\Delta_0\|_2^2 \frac{1}{d}\operatorname{trace} \mathbb{E}\left[Q_0 \dots Q_{t-1} Q_{t-1} \dots Q_0\right].
\end{aligned}$$

We then combine the earlier expression (44) with the above upper-bound and complete the proof.

Remark A.5. *In view of equations (4-6) in [2], one can show that asymptotic freeness between* $U^\top S^\top SU$ *and a rank-one matrix* $vv^\top$ *holds provided that* $\|v\|_2 < \infty$ *as the dimensions grow to infinity. One could then wonder whether such a result can be applied to our setting, in order to remove the assumption* $\mathbb{E}\Delta_0\Delta_0^\top = \frac{1}{d}\cdot I_d$*. Using* (43)*, dividing by* $\mathbb{E}\|\Delta_0\|^2$ *and denoting* $\widetilde{\Delta}_0 = \frac{\Delta_0}{\sqrt{\mathbb{E}\|\Delta_0\|^2/d}}$*, we get*

$$\frac{\mathbb{E}\|\Delta_t\|^2}{\mathbb{E}\|\Delta_0\|^2} = \frac{1}{d}\,\mathrm{trace}\,\mathbb{E}\left[Q_0\dots Q_{t-1}Q_{t-1}\dots Q_0\widetilde{\Delta}_0\widetilde{\Delta}_0^\top\right].$$

Provided we have asymptotic freeness between $\widetilde{\Delta}_0\widetilde{\Delta}_0^\top$ *and* $Q_0\dots Q_{t-1}Q_{t-1}\dots Q_0$*, then we have*

$$\lim_{n\to\infty}\frac{\mathbb{E}\|\Delta_t\|^2}{\mathbb{E}\|\Delta_0\|^2} = \lim_{n\infty}\frac{1}{d}\,\mathrm{trace}\,\mathbb{E}\left[Q_0\dots Q_{t-1}Q_{t-1}\dots Q_0\right]\cdot\lim_{n\infty}\frac{1}{d}\,\mathrm{trace}\,\mathbb{E}\left[\widetilde{\Delta}_0\widetilde{\Delta}_0^\top\right]$$

According to our previous analysis, the term $\lim_{n\infty}\frac{1}{d}\,\mathrm{trace}\,\mathbb{E}\left[Q_0\dots Q_{t-1}Q_{t-1}\dots Q_0\right]$ *is equal to* $(1-\frac{\theta_{1,h}^2}{\theta_{2,h}})^t$*. On the other hand, the term* $\lim_{n\infty}\frac{1}{d}\,\mathrm{trace}\,\mathbb{E}\left[\widetilde{\Delta}_0\widetilde{\Delta}_0^\top\right]$ *is equal to* 1*, so that we would get the claimed result. But, for asymptotic freeness to hold between* $\widetilde{\Delta}_0\widetilde{\Delta}_0^\top$ *and* $Q_0\dots Q_{t-1}Q_{t-1}\dots Q_0$*, we need* $\|\widetilde{\Delta}_0\| < \infty$*, and this assumption seems too strong: for instance, if* Δ_0 *is deterministic, then* $\|\widetilde{\Delta}_0\| = \sqrt{d}$ *which is unbounded as the dimensions grow to infinity.*

B Proofs of the auxiliary results

B.1 Proof of the bounds on the support of F_h (Lemma A.2)

Proof. We show that the support of F_h satisfies

$$\inf\ \mathrm{supp}(F_h) \geqslant \frac{\left(1-\sqrt{\rho_g}\right)^2}{(1+\frac{1}{\sqrt{\xi}})^2}.$$

Let S be an $m\times n$ Haar matrix, U an $n\times d$ deterministic matrix with orthonormal columns, and S_g be an $m\times n$ matrix independent of S, with i.i.d. entries $\mathcal{N}(0,1/m)$. Write $S_g = \Omega_\ell\Sigma\Omega_r$ a singular value decomposition of S_g. It holds that Ω_ℓ is an $m\times m$ Haar matrix, independent of the $m\times m$ diagonal matrix of singular values Σ, and $\Omega_r \overset{\mathrm{d}}{=} S$, so that $\Omega_\ell\Sigma S \overset{\mathrm{d}}{=} S_g$. Further, the operator norm of Σ satisfies $\lim_{n\to\infty}\|\Sigma\|_2 = \left(1+\frac{1}{\sqrt{\xi}}\right)$ almost surely. Then,

$$\begin{aligned}\sigma_{\min}(SU) = \min_{\|x\|=1}\|SUx\| &\geqslant \min_{\|x\|=1}\frac{\|\Sigma SUx\|}{\|\Sigma\|_2}\\ &= \frac{1}{\|\Sigma\|_2}\cdot\min_{\|x\|=1}\|\Omega_\ell\Sigma SUx\|.\end{aligned}$$

Almost surely, $\min_{\|x\|=1}\|\Omega_\ell\Sigma Sx\| \to (1-\sqrt{\rho_g})$ as $n\to\infty$. Thus, almost surely, $\liminf_{n\to\infty}\sigma_{\min}(SU) \geqslant \frac{\left(1-\sqrt{\rho_g}\right)}{(1+\frac{1}{\sqrt{\xi}})}$, which yields the claimed lower bound on the support of F_h. □

B.2 Characterization of the e.s.d. of C_n

Recall the definition (18) of the matrix C_n,

$$C_n = \begin{pmatrix} I_m & 0\\ 0 & 0\end{pmatrix}\mathbb{W}_n\begin{pmatrix} I_d & 0\\ 0 & 0\end{pmatrix}\mathbb{W}_n^\top\begin{pmatrix} I_m & 0\\ 0 & 0\end{pmatrix}.$$

We leverage Theorem 4.11 from [7], which we recall for the sake of completeness.

Theorem B.1 (Theorem 4.11, [7])**.** *Let* $D_n\in\mathbb{R}^{n\times n}$ *and* $T_n\in\mathbb{R}^{n\times n}$ *be diagonal non-negative matrices, and* $\mathbb{W}_n\in\mathbb{R}^{n\times n}$ *be a Haar matrix. Denote* F_D *and* F_T *the respective l.s.d. of* D_n *and*

T_n. Denote C_n the matrix $C_n := D_n^{\frac{1}{2}} \mathbb{W}_n T_n \mathbb{W}_n^\top D_n^{\frac{1}{2}}$. Then, as n tends to infinity, the e.s.d. of C_n converges to F whose η-transform η_F satisfies

$$\begin{aligned} \eta_F(z) &= \int \frac{1}{z\gamma(z)x+1} \, \mathrm{d}F_D(x) \,, \\ \gamma(z) &= \int \frac{x}{\eta_F(z) + z\delta(z)x} \, \mathrm{d}F_T(x) \,, \\ \delta(z) &= \int \frac{x}{z\gamma(z)x+1} \mathrm{d}F_D(x) \,. \end{aligned}$$

The e.s.d. of $\begin{pmatrix} I_d & 0 \\ 0 & 0 \end{pmatrix}$ converges to the distribution F_γ with density $\gamma\delta_1 + (1-\gamma)\delta_0$, and the e.s.d. of $\begin{pmatrix} I_m & 0 \\ 0 & 0 \end{pmatrix}$ converges to the distribution F_ξ with density $\xi\delta_1 + (1-\xi)\delta_0$. Then, according to Theorem B.1, the e.s.d. of C_n converges to a distribution F_C, whose η-transform η_C is solution of the following system of equations,

$$\eta_C(z) = \int \frac{1}{z\gamma(z)x+1} \, \mathrm{d}F_\xi(x) \,, \tag{45}$$

$$\gamma(z) = \int \frac{x}{\eta_C(z) + z\delta(z)x} \, \mathrm{d}F_\gamma(x) \,, \tag{46}$$

$$\delta(z) = \int \frac{x}{z\gamma(z)x+1} \, \mathrm{d}F_\xi(x) \,. \tag{47}$$

Plugging the above expressions of F_ξ and F_γ into the above equations, and after simplification, we obtain that η_C is solution of the following second-order equation

$$\eta_C(z) = (1-\gamma) + \frac{\gamma}{1 + z\left(1 + \frac{\xi - 1}{\eta_C(z)}\right)} \,, \tag{48}$$

Plugging the relationship (7) between the Stieltjes and η-transforms into (48), we find that

$$m_C(z) = \frac{z + \gamma + \xi - 2 - \sqrt{g(z)}}{2z(1-z)} \,, \tag{49}$$

where $g(z) = (\gamma+\xi-2+z)^2 + 4(z-1)(1-\gamma)(1-\xi)$, and we choose the branch of the square-root such that $m_C(z) \in \mathbb{C}^+$ for $z \in \mathbb{C}^+$, $m_C(z) \in \mathbb{C}^-$ for $z \in \mathbb{C}^-$ and $m_C(z) > 0$ for $z < 0$.

B.3 Proof of Lemma A.3

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

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

and

$$DP = P\tilde{D}, \quad P^\top D = \tilde{D}P^\top \,. \tag{50}$$

It follows that

$$\begin{aligned} U^\top P^\top DH_n DPBP^\top DH_n DPU &= U^\top P^\top DH_n P\tilde{D}B\tilde{D}P^\top H_n DPU \\ &= U^\top P^\top DH_n PB\tilde{D}^2 P^\top H_n DPU \\ &= U^\top P^\top DH_n PBP^\top H_n DPU \\ &= U^\top P^\top DH_n \tilde{B} H_n DPU \\ &\overset{d}{=} U^\top P^\top DH_n BH_n DPU, \end{aligned}$$

where the first equation follows from (50), the second equation holds because $\tilde{D}$ and B are diagonal so they commute, while the third equation holds because $\tilde{D}^2 = I_n$. $\square$

B.4 Proof of the identity (40)

We note that

$$
\begin{aligned}
F_{G_n}(x) &\overset{\mathrm{d}}{=} \left(1 - \frac{\widetilde{d}}{n}\right) \mathbf{1}_{\{x \geqslant 0\}} + \frac{1}{n} \sum_{j=1}^{\widetilde{d}} \mathbf{1}_{\{x \geqslant \lambda_j\}} \\
&= \left(1 - \frac{\widetilde{d}}{n}\right) \mathbf{1}_{\{x \geqslant 0\}} + \frac{d}{n} \cdot \frac{1}{d} \sum_{j=1}^{\widetilde{d}} \mathbf{1}_{\{x \geqslant \lambda_j\}} \\
&= \left(1 - \frac{\widetilde{d}}{n}\right) \mathbf{1}_{\{x \geqslant 0\}} + \frac{d}{n} \left(F_{S,n}(x) - \left(\frac{d - \widetilde{d}}{d}\right) \mathbf{1}_{\{x \geqslant 0\}}\right) \\
&= \left(1 - \frac{d}{n}\right) \mathbf{1}_{\{x \geqslant 0\}} + \frac{d}{n} F_{S,n}(x)\,,
\end{aligned}
$$

which proves (40).