Optimal Randomized First-Order Methods for Least-Squares Problems

Jonathan Lacotte [1] **Mert Pilanci** [1]

Abstract

We provide an exact analysis of a class of randomized algorithms for solving overdetermined least-squares problems. We consider first-order methods, where the gradients are pre-conditioned by an approximation of the Hessian, based on a subspace embedding of the data matrix. This class of algorithms encompasses several randomized methods among the fastest solvers for least-squares problems. We focus on two classical embeddings, namely, Gaussian projections and subsampled randomized Hadamard transforms (SRHT). Our key technical innovation is the derivation of the limiting spectral density of SRHT embeddings. Leveraging this novel result, we derive the family of normalized orthogonal polynomials of the SRHT density and we find the optimal pre-conditioned first-order method along with its rate of convergence. Our analysis of Gaussian embeddings proceeds similarly, and leverages classical random matrix theory results. In particular, we show that for a given sketch size, SRHT embeddings exhibits a faster rate of convergence than Gaussian embeddings. Then, we propose a new algorithm by optimizing the computational complexity over the choice of the sketching dimension. To our knowledge, our resulting algorithm yields the best known complexity for solving least-squares problems with no condition number dependence.

1. Introduction

We study the performance of a randomized method, namely, the Hessian sketch (Pilanci & Wainwright, 2016), in the context of (overdetermined) least-squares problems,

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

where $A \in \mathbb{R}^{n\times d}$ 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 will assume throughout this work that $\operatorname{rank}(A) = d$.

Many works have developed randomized algorithms (Avron et al., 2010; Rokhlin & Tygert, 2008; Drineas et al., 2011; Pilanci & Wainwright, 2015) for solving (1), based on sketching methods. The latter involve using a random matrix $S \in \mathbb{R}^{m\times n}$ to project the data A and/or b to a lower dimensional space ($m \ll n$), and then approximately solving the least-squares problem using the sketch SA and/or Sb. The most 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 forming 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) through standard matrix factorization methods, provided that $m \geqslant d$. 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, it requires time $\mathcal{O}(nm^2)$ to be formed, through a Gram-Schmidt procedure, which is also larger than $\mathcal{O}(nd^2)$. An alternative embedding which verifies orthogonality properties is the SRHT (Ailon & Chazelle, 2006), which is based on the Walsh-Hadamard transform. Due to the recursive structure of the latter, 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.

Using the standard prediction (semi-)norm $\|A(\widetilde{x}-x^*)\|^2$ as the evaluation criterion for an approximate solution $\widetilde{x}$, iterative methods (e.g., gradient descent or the conjugate gradient algorithm) have time complexity which usually scales proportionally to the condition number κ of the matrix A – defined as the ratio between the largest and smallest singular values of A –, and this becomes prohibitively large when $\kappa \gg 1$. To address the latter issue, we introduce a pre-conditioning method, namely, the Hessian sketch (Pilanci & Wainwright, 2016), which approximates the Hessian $H = A^\top A$ of $f(x)$ by $H_S = A^\top S^\top SA$. This pre-conditioning technique has become widespread in the sketching literature for solving least-squares. For instance, it has been recently shown by (Ozaslan et al., 2019; Lacotte & Pilanci, 2019)

[1]Department of Electrical Engineering, Stanford University. Correspondence to: Jonathan Lacotte <lacotte@stanford.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

that the Heavy-ball update

$$x_{t+1} = x_t - \mu_t H_S^{-1} \nabla f(x_t) + \beta_t (x_t - x_{t-1}) \qquad (2)$$

yields a sequence of iterates whose convergence rate does not depend on the spectrum of A, but only on the concentration around the identity of the matrix

$$C_S := U^\top S^\top S U \,, \qquad (3)$$

where U is the matrix of left singular vectors of A. Further, they show that this convergence rate is equal to the ratio d/m both for Gaussian and SRHT embeddings. Notably, this rate does not depend on the sample size n. For a Gaussian embedding, this makes intuitively sense since the limiting spectral distribution of C_S is the Marchenko-Pastur law (Marchenko & Pastur, 1967) with scale parameter ρ, edge eigenvalues $a = (1-\sqrt{\rho})^2$ and $b = (1+\sqrt{\rho})^2$, and density

$$\mu_\rho(x) = \frac{\sqrt{(b-x)_+(x-a)_+}}{2\pi\rho x} \,, \qquad (4)$$

where $y_+ = \max\{y, 0\}$, and it does not depend on the sample size n but only on the limit ratio $\rho := \lim \frac{d}{m}$. However, for a SRHT embedding, it is unclear if the dimension n affects the best achievable convergence rate.

In a related vein, Lacotte et al. (2020) considered the Heavy-ball update (2) where at each iteration the sketching SRHT matrix $S = S_t$ is *refreshed*, i.e., re-sampled independently of $S_0, \ldots, S_{t-1}$, so that $H_S = H_{S_t}$ is also re-computed. They show that Haar and SRHT embeddings yield the same convergence rate $\rho_h^{\text{ref}} := \rho \cdot \frac{\xi(1-\xi)}{\gamma^2+\xi-2\gamma\xi}$, which indeed depends on the three relevant dimensions m, d, n through the aspect ratios

$$\rho := \lim \frac{d}{m} \,, \quad \gamma := \lim \frac{d}{n} \,, \quad \xi := \lim \frac{m}{n} \,. \qquad (5)$$

Importantly, this convergence rate ρ_h^{ref} is always strictly smaller than ρ, which is the convergence rate one would obtain with fixed or refreshed Gaussian embeddings (Ozaslan et al., 2019; Lacotte & Pilanci, 2019).

Although using refreshed SRHT embeddings yields a better convergence rate, it comes with two major shortcomings. First, refreshing the sketch at each iteration incurs additional computational costs compared to using the same sketch at each iteration. Second, the analysis of Lacotte et al. (2020) is specific to the Heavy-ball update (2), and it leaves an open problem we aim to address in this work, that is, whether *there exists a first-order method with a fixed sketch that provides better guarantees*. Formally, we will consider the following class of pre-conditioned first-order methods,

$$x_t \in x_0 + H_S^{-1} \cdot \text{span}\left\{\nabla f(x_0), \ldots, \nabla f(x_{t-1})\right\} \,, \qquad (6)$$

and this includes in particular the Heavy-ball update (2) with a fixed sketch.

Relatedly, the design of optimal first-order methods for quadratic optimization problems has been recently considered by Pedregosa & Scieur (2019). In contrast to our setting, they assume the data matrix A to be random. Then, by leveraging the limiting spectral properties of the matrix A, they are able to design a first-order method (without preconditioning) which is optimal in the average-case. Remarkably, their first-order method improves on the worst-case rates of convergence of standard first-order methods. However, their approach requires the limiting spectral distribution of the data matrix to be known beforehand, which might be impractical. By considering instead the *pre-conditioned* first-order methods (6), we will see that only the spectral distribution of the matrix C_S is required, and this is universal, i.e., independent of the spectrum of A. Therefore, by characterizing the l.s.d. of C_S for some classical embeddings, we will be able to optimize the *exact* error for any data matrix A, without the requirement of knowing its spectral properties beforehand.

We will focus exclusively on (pre-conditioned) *first-order methods* of the form (6) with a fixed embedding S, and our goal is to answer the following questions. What are the best achievable convergence rates for, respectively, Gaussian and SRHT embeddings? What are the corresponding optimal algorithms? How do these rates compare to each other and to that of state-of-the-art randomized iterative methods for solving (1)?

1.1. Technical background, notations and assumptions

We will assume that $\lim_{n\to\infty} \frac{d}{n} = \gamma \in (0,1)$, $\lim_{n\to\infty} \frac{m}{n} = \xi \in (\gamma, 1)$ and $\rho = \frac{\gamma}{\xi} \in (0,1)$. We denote $\|z\| \equiv \|z\|_2$ the Euclidean norm of a vector z, $\|M\|_2$ the operator norm of a matrix M, and $\|M\|_F$ its Frobenius norm. Given a sequence of iterates $\{x_t\}$, we denote the error at time t by $\Delta_t = U^\top A(x_t - x^*)$. Note that $\|\Delta_t\|^2 = \|A(x_t - x^*)\|^2$. Our evaluation criterion is the error $\lim_{n\to\infty} \mathbb{E}[\|\Delta_t\|^2]/\mathbb{E}[\|\Delta_0\|^2]$, and we call its (asymptotic) rate of convergence the quantity $\limsup_{t\to\infty} \left(\lim_{n\to\infty} \mathbb{E}[\|\Delta_t\|^2]/\mathbb{E}[\|\Delta_0\|^2]\right)^{1/t}$.

As we focus on infinite-dimensional regimes, our technical analysis is based on asymptotic random matrix theory, and we refer the reader to (Bai & Silverstein, 2010; Paul & Aue, 2014; Yao et al., 2015) for an extensive introduction to this field. For a random Hermitian matrix M_n of size $n \times n$, 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. Key to our analysis is the notion of orthogonal polynomials, which are fundamental both in optimization (Rutishauser, 1959) and in random matrix theory. We write $\mathbb{R}_t[X]$ the set of real polynomials with degree less than t, and $\mathbb{R}_t^0[X]$ the set of polynomials $P \in \mathbb{R}_t[X]$ such that $P(0) = 1$. For a complex number $z \in \mathbb{C}$, we denote respectively by $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$ its real and imaginary parts, and we use $\mathbb{C}_+$ for the complex numbers with positive imaginary parts, and $\mathbb{R}_+$ for the positive real numbers. For two sequences of real positive numbers $\{a_t\}$ and $\{b_t\}$, we write $a_t \asymp b_t$ if $\liminf \frac{a_t}{b_t} > 0$ and $\limsup \frac{a_t}{b_t} < \infty$.

We will assume that the first iterate x_0 is random such that $\mathbb{E}[x_0] = 0$, and, that the condition number of the matrix $U^\top A \mathbb{E}[x_0 x_0^\top] A^\top U + U^\top b b^\top U$ remains bounded as the dimensions grow. Essentially, this states that the condition number of A does not degenerate to $+\infty$ as the dimensions grow.

In this work, we consider a definition of the SRHT slightly different than its classical version (Ailon & Chazelle, 2006), which has been introduced in (Dobriban & Liu, 2019; Liu & Dobriban, 2019). For an integer $n = 2^p$ with $p \geqslant 1$, the Walsh-Hadamard transform is defined recursively as $H_n = \frac{1}{\sqrt{2}} \begin{bmatrix} H_{n/2} & H_{n/2} \\ H_{n/2} & -H_{n/2} \end{bmatrix}$ with $H_1 = 1$. 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 $S = BH_nDP$, 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 while $n \to \infty$. Although the dimension $\widetilde{m}$ is random, we refer to S as an $m \times n$ SRHT matrix.

1.2. Overview of our results and contributions

We have the following contributions.

1. For Gaussian embeddings, we characterize the algorithm (Algorithm 1) which attains the infimum of the error $\lim_{n\to\infty} \mathbb{E}[\|\Delta_t^2\|]/\mathbb{E}[\|\Delta_0\|^2]$, and we show that it corresponds to the Heavy-ball method with constant step size $\mu_t = (1-\rho)^2$ and momentum parameter $\beta_t = \rho$. Further, we show that the infimum of the error is equal to ρ^t.

2. For SRHT embeddings, we perform a similar analysis, and find the optimal first-order method (Algorithm 2). Notably, it is a Heavy-ball update with non-constant step sizes and momentum parameters. Further, we show that its rate of convergence is $\rho_h := \rho \cdot \frac{1-\xi}{1-\gamma}$, which is always strictly smaller than ρ and ρ_h^{ref}, i.e., Algorithm 2 has uniformly better convergence rate than that of Gaussian embeddings or the Heavy-ball method with refreshed SRHT embeddings. Even though our theoretical results hold asymptotically, we verify empirically that our theoretical predictions hold, even for sample sizes $n \gtrsim 1000$, and that Algorithm 2 is faster in practice than the other aforementioned algorithms.

3. We characterize explicitly the density $f_{h,r}$ of the l.s.d of the matrix $\frac{n}{m} C_S$, which is given by

$$f_{h,r}(x) = \frac{\sqrt{(\Lambda_{h,r} - x)_+ (x - \lambda_{h,r})_+}}{2\pi\rho x(1 - \xi x)}, \qquad (7)$$

   where the edge (i.e., extreme) eigenvalues are $\lambda_{h,r} = (\sqrt{1-\gamma} - \sqrt{(1-\xi)\rho})^2$ and $\Lambda_{h,r} = (\sqrt{1-\gamma} + \sqrt{(1-\xi)\rho})^2$. This characterization of the limiting density is of independent interest, as it might have several implications beyond least-squares optimization.

4. Finally, we show that Algorithm 2 has the best known complexity to solve (1) with no condition number dependence.

Except for the time complexity results, all our results regarding the SRHT hold exactly the same with Haar embeddings, since they both yield the same limiting spectral distributions.

1.3. Other related work

Besides the Hessian sketch, there are many other efficient pre-conditioned iterative methods which aim to address the aforementioned conditioning issue, based on an SRHT sketch of the data (or closely related sketches based on the Fourier transform). Randomized *right* pre-conditioning methods (Avron et al., 2010; Rokhlin & Tygert, 2008) compute first a matrix P – which itself depends on SA – such that the condition number of AP^{-1} is $\mathcal{O}(1)$, and then apply any standard iterative algorithm to the pre-conditioned least-squares objective $\|AP^{-1}y - b\|^2$. SRHT sketches are also used for a wide range of applications across numerical linear algebra, statistics and convex optimization, such as low-rank matrix factorization (Halko et al., 2011; Witten & Candes, 2015), kernel regression (Yang et al., 2017),

random subspace optimization (Lacotte et al., 2019), or, sketch and solve linear regression (Dobriban & Liu, 2019). Hence, a refined analysis of the SRHT may also lead to better algorithms in these fields.

Our work also substantiates the observation that, in a growing number of contexts, random projections with i.i.d. entries degrade the performance of the approximate solution compared to orthogonal projections (Mahoney, 2011; Mahoney & Drineas, 2016; Drineas & Mahoney, 2016; Dobriban & Liu, 2019).

2. Optimal first-order method for classical embeddings

Let S be an $m \times n$ Gaussian or SRHT embedding. Denote by μ the l.s.d. of C_S. We say that a family of polynomials $\{R_k\}$ is orthogonal with respect to μ if $\int R_k R_\ell \, \mathrm{d}\mu = 0$ for any $k \neq \ell$. The next result establishes the link between polynomials and the pre-conditioned first-order methods (6) we consider, and its proof is deferred to Appendix B.1.

Lemma 2.1. *Let $\{x_t\}$ be generated by some first-order method (6). Then, for any iteration $t \geqslant 0$, there exists a polynomial $p_t \in \mathbb{R}^0_t[X]$ such that $\Delta_t = p_t\big(C_S^{-1}\big) \cdot \Delta_0$. Further, it holds that*

$$\lim_{n\to\infty} \frac{\mathbb{E}[\|\Delta_t\|^2]}{\mathbb{E}[\|\Delta_0\|^2]} = \int_{\mathbb{R}} p_t^2\big(\lambda^{-1}\big) \, \mathrm{d}\mu(\lambda) \,. \tag{8}$$

Thus, the best achievable error is lower bounded by the infimum of the following variational problem,

$$\mathcal{L}^*_{\mu,t} := \min_{p \in \mathbb{R}^0_t[X]} F_\mu(p) \,, \tag{9}$$

where $F_\mu(p) := \int p^2\big(\lambda^{-1}\big) \, \mathrm{d}\mu(\lambda)$. Using the change of variable $x = 1/\lambda$ and setting $\mathrm{d}\nu(x) = x^{-1}\mathrm{d}\mu\big(x^{-1}\big)$, we have that $F_\mu(p) = G_\nu(p)$ where $G_\nu(p) := \int p^2(x) \, \frac{1}{x} \mathrm{d}\nu(x)$. The optimal polynomial can be constructed by leveraging the following result.

Lemma 2.2. *Let ν be some measure with bounded support in $(0, +\infty)$, and suppose that $\{\Pi_t\}$ is a family of orthogonal polynomials with respect to ν such that $\deg(\Pi_t) = t$ and $\Pi_t(0) = 1$. Then, the polynomial Π_t is the unique solution of the optimization problem $\min G_\nu(p)$ over $p \in \mathbb{R}^0_t[X]$.*

Proof. Let $p \in \mathbb{R}^0_t[X]$. Since $\Pi_t(0) = 1$, the polynomial $(p - \Pi_t)$ has a root at 0. Hence, $(p - \Pi_t)(x) = xQ(x)$ with $Q \in \mathbb{R}_{t-1}[X]$. Then,

$$\begin{aligned} G_\nu(p) &= \int p^2(x) x^{-1} \mathrm{d}\nu(x) \\ &= \int \Pi_t^2(x) x^{-1} \mathrm{d}\nu(x) + 2 \int \Pi_t Q(x) \mathrm{d}\nu(x) \\ &\quad + \int x Q^2(x) \mathrm{d}\nu(x) \,. \end{aligned}$$

The cross-term is equal to 0 since Q in the span of $\Pi_0, \ldots, \Pi_{t-1}$, which are orthogonal to Π_t. The third term is non-negative, and equal to 0 if and only if that $Q = 0$. Therefore, the unique solution to (9) is Π_t. □

Based on such an orthogonal family $\{\Pi_t\}$, we aim to derive a first-order method which achieves the lower bound $\mathcal{L}^*_{\mu,t}$. We recall a standard result, that is, for such a family of polynomials $\{\Pi_t\}$, there exist sequences $\{a_t\}$ and $\{b_t\}$ such that $\Pi_0(x) = 1$, $\Pi_1(x) = 1 + b_1 x$ and for any $t \geqslant 2$,

$$\Pi_t(x) = (a_t + b_t x)\Pi_{t-1}(x) + (1 - a_t)\Pi_{t-2}(x) \,. \tag{10}$$

Then we can construct an optimal first-order method according to the following result, which is inspired by the work of Pedregosa & Scieur (2019) and whose proof is deferred to Appendix A.1.

Theorem 1. *Given $x_0 \in \mathbb{R}^d$, set $x_1 = x_0 + b_1 H_S^{-1} \nabla f(x_0)$, and for $t \geqslant 2$,*

$$x_t = x_{t-1} + b_t H_S^{-1} \nabla f(x_{t-1}) + (1 - a_t)(x_{t-2} - x_{t-1}) \,. \tag{11}$$

Then, the sequences of iterates $\{x_t\}$ is asymptotically optimal, i.e.,

$$\lim_{n\to\infty} \frac{\mathbb{E}\|\Delta_t\|^2}{\mathbb{E}\|\Delta_0\|^2} = \mathcal{L}^*_{\mu,t} \,. \tag{12}$$

Consequently, a strategy to find the optimal first-order method proceeds as follows. First, we characterize the l.s.d. μ of the matrix C_S, and we find the polynomial $\Pi_t \in \mathbb{R}^0_t[X]$ which achieves the lower bound $\mathcal{L}^*_{\mu,t}$. Then, according to Theorem 1, we build from the three-terms recursion (10) of the orthogonal polynomials $\{\Pi_t\}$ a first-order method which yields an asymptotically optimal sequence of iterates $\{x_t\}$. Our analysis of the Gaussian case is based on standard random matrix theory results, that we recall in details as we leverage them for the analysis of the SRHT case. For the latter, most technicalities actually lie in characterizing the l.s.d. μ of C_S, and in constructing an orthogonal basis of polynomials for the distribution $\mathrm{d}\nu(x) = x^{-1}\mathrm{d}\mu\big(x^{-1}\big)$.

2.1. The Gaussian case

Consider an $m \times n$ matrix S with i.i.d. entries $\mathcal{N}\big(0, m^{-1}\big)$. The l.s.d. of C_S is the Marchenko-Pastur law with density μ_ρ given in (4). Denote by $a = (1 - \sqrt{\rho})^2$ and $b = (1 + \sqrt{\rho})^2$ the edge eigenvalues. Let $\{\Delta_t\}$ be the sequence of error vectors generated by a first-order method as in (6). According to Lemma 2.1, there exists a sequence of polynomials $p_t \in \mathbb{R}^0_t[X]$ such that $\Delta_t = p_t\big(C_S^{-1}\big) \, \Delta_0$, and

$$\lim_{n\to\infty} \frac{\mathbb{E}\|\Delta_t\|^2}{\mathbb{E}\|\Delta_0\|^2} = \int_a^b p_t^2(\lambda^{-1}) \mu_\rho(\lambda) \mathrm{d}\lambda \,, \tag{13}$$

Lemma 2.3. *Under the above assumptions and notations, and setting* $P_t(x) = p_t\left(\frac{x}{(1-\rho)^2}\right)$, *we have*

$$\lim_{n\to\infty} \frac{\mathbb{E}\|\Delta_t\|^2}{\mathbb{E}\|\Delta_0\|^2} = (1-\rho)\int_a^b P_t^2(x)\frac{1}{x}\mu_\rho(x)\,\mathrm{d}x\,. \quad (14)$$

Consequently, if $\{\Pi_t\}$ *is an orthogonal basis of polynomials with respect to* μ_ρ *such that* $\deg(\Pi_t) = t$ *and* $\Pi_t(0) = 1$ *then* $\overline{\Pi}_t(x) := \Pi_t\big((1-\rho)^2 x\big)$ *achieves the lower bound* $\mathcal{L}^*_{\mu_\rho,t}$.

Proof. Using the change of variable $x = (1-\rho)^2/\lambda$, a simple calculation yields that $p_t^2(\lambda^{-1})\mu_\rho(\lambda)\mathrm{d}\lambda = (1-\rho)P_t^2(x)\frac{1}{x}\mu_\rho(x)\mathrm{d}x$. Applying Lemma 2.2 with $\nu = \mu_\rho$, we get that the optimal polynomial P_t is equal to Π_t, and thus, p_t is exactly $\overline{\Pi}_t(x)$. □

The Marchenko-Pastur law μ_ρ is well-studied, and such a construction of polynomials is classical. In this section, we provide a definition by recursion, which is enough to state the optimal algorithm. However, for the proof of the next results, we will consider an alternative construction, from which we establish several intermediate properties useful to the analysis. Define $\Pi_0(x) = 1$, $\Pi_1(x) = 1 - x$, and for $t \geqslant 2$,

$$\Pi_t(x) = (1+\rho-x)\Pi_{t-1}(x) - \rho\,\Pi_{t-2}(x)\,. \quad (15)$$

Lemma 2.4. *The family of polynomials* $\{\Pi_t\}$ *is orthogonal with respect to* μ_ρ. *Further, we have* $\Pi_t(0) = 1$ *and* $\deg(\Pi_t) = t$ *for all* $t \geqslant 0$.

Proof. We defer the proof to Section B.4. □

Now, set $\overline{\Pi}_t(x) = \Pi_t\big((1-\rho)^2 x\big)$. From (15), we obtain that $\overline{\Pi}_0(x) = 1$, $\overline{\Pi}_1(x) = 1 - (1-\rho)^2 x$, and for $t \geqslant 2$,

$$\overline{\Pi}_t(x) = (1+\rho-(1-\rho)^2 x)\overline{\Pi}_{t-1}(x) - \rho\,\overline{\Pi}_{t-2}(x)\,. \quad (16)$$

According to Lemma 2.3, the polynomial $\overline{\Pi}_t$ achieves the lower bound $\mathcal{L}^*_{\mu_\rho,t}$. Further, we identify the recursion formula (16) with the three-terms recursion (10) by setting $b_t = -(1-\rho)^2$ for $t \geqslant 1$, and $a_t = 1+\rho$ for $t \geqslant 2$. Using Theorem 1, we immediately have the asymptotically optimal first-order method, which we present in Algorithm 1 in its finite-sample approximation.

Algorithm 1 Optimal First-Order Method for Gaussian embeddings.

Input: Data matrix $A \in \mathbb{R}^{n\times d}$, sketch size $m \geqslant d+1$, initial point $x_0 \in \mathbb{R}^d$ and (finite-sample) ratio $\rho := d/m$.
Sample $S \in \mathbb{R}^{m\times n}$ with i.i.d. entries $\mathcal{N}(0, 1/m)$.
Compute the sketched matrix $S_A = S \cdot A$.
Compute and cache a factorization of $H_S = S_A^\top S_A$.
Set $x_1 = x_0 - (1-\rho)^2 H_S^{-1} \cdot A^\top(Ax_0 - b)$.
for $t = 2$ **to** T **do**
  Compute the gradient $g_{t-1} = A^\top(Ax_{t-1} - b)$.
  Perform the update

$$x_t = x_{t-1} + \rho(x_{t-1} - x_{t-2}) - (1-\rho)^2 \cdot H_S^{-1} g_t\,. \quad (17)$$

end for
Return the last iterate $\mathbf{x_T}$.

Surprisingly, up to the initialization of the first iterate x_1, Algorithm 1 corresponds exactly to the Heavy-ball method (2) using the fixed step size $\mu = (1-\rho)^2$ and the fixed momentum parameter $\beta = \rho$, which was obtained in (Ozaslan et al., 2019; Lacotte & Pilanci, 2019) based on edge eigenvalues analysis. Hence, in the Gaussian case, leveraging the whole shape of the limiting distribution, as opposed to using only the edge eigenvalues, yields the same algorithm. We complete the analysis of the Gaussian case by providing the exact asymptotic error $\mathcal{L}^*_{\mu_\rho,t}$.

Theorem 2. *The sequence of iterates* $\{x_t\}$ *given by Algorithm 1 is asymptotically optimal within the class of first order algorithms as in* (6)*, and the optimal error is given by* $\mathcal{L}^*_{\mu_\rho,t} = \rho^t$.

Proof. We have already argued that $\{x_t\}$ is asymptotically optimal. It remains to show that $\mathcal{L}^*_{\mu_\rho,t} = \rho^t$, whose proof is deferred to Appendix A.2. □

2.2. The SRHT case

Haar random projections have been shown to have a better performance than Gaussian embeddings in several contexts. However, they are slow to generate and apply, and we consider instead the SRHT. We recall the definition of the Stieltjes transform m_μ of a distribution μ supported on $[0, +\infty)$, which, for $z \in \mathbb{C}\backslash\mathbb{R}_+$, is given by $m_\mu(z) := \int_{\mathbb{R}} \frac{1}{x-z}\,\mathrm{d}\mu(x)$. It has been recently shown that the SRHT behaves asymptotically as Haar embeddings, as formally stated by the next result.

Lemma 2.5 (Theorem 4.1 in Lacotte et al. (2020))**.** *Let* S *be an* $m\times n$ *SRHT embedding and* S_h *be an* $m\times n$ *Haar embedding. Then, the matrices* C_S *and* C_{S_h} *have the same limiting spectral distribution* F_h, *with support included within the*

interval $(0,1)$ *and whose Stieltjes transform* m_h *is given by*

$$m_h(z) = \frac{1}{2\gamma}\left(\frac{2\gamma-1}{1-z} + \frac{\xi-\gamma}{z(1-z)} - \frac{R(z)}{z(1-z)}\right), \quad (18)$$

where

$$R(z) = \sqrt{(\gamma+\xi-2+z)^2 + 4(z-1)(1-\gamma)(1-\xi)},$$

Remark 1. *Due to the computational benefits of the SRHT over Haar projections, we state all our next results for the former, although all statements also apply to the latter (except for the time complexity results).*

In order to characterize the optimal first-method with SRHT embeddings, we first derive the density of F_h.

Theorem 3. *The distribution* F_h *admits the following density on* $\mathbb{R}$*,*

$$f_h(x) = \frac{1}{2\gamma\pi}\frac{\sqrt{(\Lambda_h - x)_+(x-\lambda_h)_+}}{x(1-x)}, \quad (19)$$

where

$$\begin{cases} \lambda_h := \left(\sqrt{(1-\gamma)\xi} - \sqrt{(1-\xi)\gamma}\right)^2, \\ \Lambda_h := \left(\sqrt{(1-\gamma)\xi} + \sqrt{(1-\xi)\gamma}\right)^2. \end{cases}$$

Proof. The proof is essentially based on the expression (18) of the Stieltjes transform m_h, and on the inversion formula,

$$f_h(x) = \lim_{y\to 0^+}\frac{1}{\pi}\mathrm{Im}\left(m_h(x+iy)\right), \text{ where } y \in \mathbb{R}_+. \quad (20)$$

which holds for any $x \in \mathbb{R}$ provided that the above limit exists (Silverstein & Choi, 1995). We defer the calculations to Appendix A.3. □

Using the change of variable $y = x/\xi$, we can also derive the limiting density of the *rescaled* matrix $\frac{n}{m}C_S$ – whose expectation is equal to the identity – which is given by

$$f_{h,r}(y) = \xi f_h(\xi y) = \frac{\sqrt{(\Lambda_{h,r}-y)_+(y-\lambda_{h,r})_+}}{2\rho\pi y(1-\xi y)}, \quad (21)$$

where

$$\begin{cases} \lambda_{h,r} = \lambda_h/\xi = \left(\sqrt{1-\gamma} - \sqrt{(1-\xi)\rho}\right)^2, \\ \Lambda_{h,r} = \Lambda_h/\xi = \left(\sqrt{1-\gamma} + \sqrt{(1-\xi)\rho}\right)^2. \end{cases}$$

The density $f_{h,r}$ resembles the Marchenko-Pastur density μ_ρ, up to the factor $(1-\xi y)$ and corrections in the edge eigenvalues $\lambda_{h,r}$ and $\Lambda_{h,r}$. When $\xi, \gamma \approx 0$, then $\lambda_{h,r} \approx (1-\sqrt{\rho})^2$, $\Lambda_{h,r} \approx (1+\sqrt{\rho})^2$, and $f_{h,r}(x) \approx \mu_\rho(x)$. This is consistent with the fact that provided $m, d = o(n)$ so that $\xi, \gamma = 0$, then the l.s.d. of $\frac{n}{m}C_S$ is the Marchenko-Pastur law with parameter ρ (see (Jiang, 2009) for a formal statement). In Figure 1, we compare the empirical spectral density of the matrix $\frac{n}{m}C_S$ with S an $m \times n$ SRHT to $f_{h,r}$, for fixed d and n, and several values of m. We observe that these two densities match very closely, and so does the empirical spectral density using a Haar projection with $f_{h,r}$. Further, as m increases, the limiting density $f_{h,r}$ departs from μ_ρ, and then concentrates more and more around 1. Note in particular that the support of $f_{h,r}$ is always within that of μ_ρ. This can be formally verified by comparing their respective edge eigenvalues.

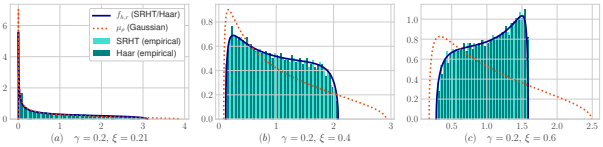

Figure 1. We use $n = 8192$, $\gamma \approx \frac{d}{n} = 0.2$ and $\xi \approx \frac{m}{n} \in \{0.21, 0.4, 0.6\}$.

2.2.1. Orthogonal polynomials and optimal first-order method

Given a first-order method as in (6), we know from Lemma 2.1 that for a given iteration t, there exists a polynomial $p \in \mathbb{R}_t^0[X]$ such that $\Delta_t = p\left(C_S^{-1}\right)\Delta_0$, and

$$\lim_{n\to\infty}\frac{\mathbb{E}\|\Delta_t\|^2}{\mathbb{E}\|\Delta_0\|^2} = \int_{\lambda_h}^{\Lambda_h} p^2\left(\lambda^{-1}\right) f_h(\lambda)\,\mathrm{d}\lambda. \quad (22)$$

Introducing the scaling parameters $\tau = \left(\frac{\sqrt{\Lambda_h}-\sqrt{\lambda_h}}{\sqrt{\Lambda_h}+\sqrt{\lambda_h}}\right)^2$, $c = \frac{4}{\left(\sqrt{1/\Lambda_h}+\sqrt{1/\lambda_h}\right)^2}$, $\alpha = (1-\sqrt{\tau})^2$, $\beta = (1+\sqrt{\tau})^2$, the rescaled polynomial $P(x) = p(x/c)$, and using the change of variable $x = c/\lambda$, we find that

$$\lim_{n\to\infty}\frac{\mathbb{E}\|\Delta_t\|^2}{\mathbb{E}\|\Delta_0\|^2} \quad (23)$$

$$= \frac{c\tau}{(1-\tau)\gamma}\int_\alpha^\beta P^2(x)\frac{\sqrt{(x-\alpha)(\beta-x)}}{2\pi\tau x(x-c)}\,\mathrm{d}x \quad (24)$$

$$= \frac{c\tau}{(1-\tau)\gamma}\int_\alpha^\beta P^2(x)\frac{\mu_\tau(x)}{x-c}\,\mathrm{d}x \quad (25)$$

Thus, according to Lemma 2.2, it suffices to find a family of polynomials $\{R_t\}$ orthogonal with respect to the density $\frac{x\mu_\tau(x)}{x-c}$ such that $\deg(R_t) = t$ and $R_t(0) = 1$, in which case the minimizer over $P \in \mathbb{R}_t^0[X]$ of the integral in (25) is equal to R_t, and the minimizer of (22) is then $\overline{R}_t(x) = R_t(cx)$.

Theorem 4. *Define the parameters* $\omega = \frac{4}{\left(\sqrt{\beta-c}+\sqrt{\alpha-c}\right)^2}$ *and* $\kappa = \left(\frac{\sqrt{\beta-c}-\sqrt{\alpha-c}}{\sqrt{\beta-c}+\sqrt{\alpha-c}}\right)^2$. *Let* $\{\Pi_t\}$ *be the orthogonal*

family of polynomials with respect to μ_κ, that is, $\Pi_0(x) = 1$, $\Pi_1(x) = 1 - x$, and for $t \geqslant 2$,

$$\Pi_t(x) = (1 + \kappa - x)\Pi_{t-1}(x) - \kappa\Pi_{t-2}(x)\,. \tag{26}$$

Define the polynomials $R_t(x) = \Pi_t(\omega(x - c))/\Pi_t(-\omega c)$. Then, it holds that $R_t(0) = 1$, $\deg(R_t) = t$, and the family $\{R_t\}$ is orthogonal with respect to the density $\frac{x\mu_\tau(x)}{x-c}$.

Proof. For $k \neq \ell$, we have that $\int_{\mathbb{R}} R_k(x)R_\ell(x)\frac{x\mu_\tau(x)}{x-c}\,\mathrm{d}x \propto \int_\alpha^\beta \Pi_k\left(\omega(x-c)\right)\Pi_\ell\left(\left(\omega(x-c)\right)\frac{\sqrt{(\beta-x)(x-\alpha)}}{2\pi\rho(x-c)}\,\mathrm{d}x$. Using the change of variable $y = \omega(x-c)$, we find that the latter integral is (up to a constant) equal to $\int_{\mathbb{R}} \Pi_k(y)\Pi_\ell(y)\mu_\kappa(y)\,\mathrm{d}y$, which is itself equal to 0 due to the orthogonality of the Π_t with respect to μ_κ. □

In order to derive the optimal first-order method, we need to find the three-terms recursion relationship satisfied by the polynomials $\{\overline{R}_t\}$. First, let us compute the normalization factor $u_t := \Pi_t(-\omega c)$. Evaluating (26) at $x = -\omega c$ and denoting $\eta := 1+\kappa+\omega c$, we find that $u_{t+1} = \eta u_t - \kappa u_{t-1}$, with the initial conditions $u_0 = 1$ and $u_1 = \Pi_1(-\omega c) = 1 + \omega c = \eta - \kappa$. Thus, after solving this second-order linear system, we obtain that

$$u_t = \frac{x_1 - \kappa}{x_1 - x_2}x_1^t + \frac{\kappa - x_2}{x_1 - x_2}x_2^t\,, \tag{27}$$

where $x_1 = \frac{\eta}{2} + \sqrt{\frac{\eta^2}{4} - \kappa}$ and $x_2 = \frac{\eta}{2} - \sqrt{\frac{\eta^2}{4} - \kappa}$. It is easy to check that $\eta^2/4 > \kappa$, so that x_1 and x_2 are indeed distinct and real. Then, using the change of variable $y = \omega(x-c)$ in (26), we get the following three-terms recurrence relationship, that is, $\overline{R}_0(x) = 1$, $\overline{R}_1(x) = 1 + b_{h,1}x$ and for $k \geqslant 2$,

$$\overline{R}_t(x) = (a_{h,t} + xb_{h,t})\overline{R}_{t-1}(x) + (1 - a_{h,t})\overline{R}_{t-2}(x)\,, \tag{28}$$

where $a_{h,t} = \frac{\eta\, u_{t-1}}{u_t}$ for $t \geqslant 1$, and $b_{h,t} = -\frac{\omega c\, u_{t-1}}{u_t}$ for $t \geqslant 2$. Using Theorem 1, we obtain the optimal first-order method, which we present in Algorithm 2 in its finite-sample approximation.

Algorithm 2 Optimal First-Order Method for SRHT (or Haar) embeddings.

Input: Data matrix $A \in \mathbb{R}^{n\times d}$, sketch size $m \geqslant d + 1$, initial point $x_0 \in \mathbb{R}^d$.
Sample an $m \times n$ SRHT S.
Compute the sketched matrix $S_A = S \cdot A$.
Compute and cache a factorization of $H_S = S_A^\top S_A$.
Set $x_1 = x_0 + b_{h,1}H_S^{-1}A^\top(Ax_0 - b)$.
for $t = 2$ **to** T **do**
  Compute the gradient $g_{t-1} = A^\top(Ax_{t-1} - b)$.
  Perform the update

$$x_t = x_{t-1} + b_{h,t}H_S^{-1}g_t + (1 - a_{h,t})(x_{t-2} - x_{t-1})\,. \tag{29}$$

  where $a_{h,t}$ and $b_{h,t}$ are as described in Section 2.2.1.
end for
Return the last iterate $\mathbf{x_T}$.

Differently from the Gaussian case, Algorithm 2 does not correspond to the Heavy-ball method (2) using the fixed step size $\mu = (1 - \rho)^2$ and the fixed momentum parameter $\beta = \rho$, which was obtained by (Lacotte & Pilanci, 2019) based on edge eigenvalues analysis and standard finite-sample concentration bounds on the spectrum of SRHT matrices (Tropp, 2011).

Using the new asymptotically exact extreme eigenvalues we derived in Theorem 3 – which are different from the bounds obtained by (Tropp, 2011) – and following the same extreme eigenvalues analysis proposed by (Lacotte & Pilanci, 2019), we can derive an optimal Heavy-ball method for which the step size μ_h and momentum parameter β_h are given by $\mu_h = \frac{4}{\left(\frac{1}{\sqrt{\Lambda_h}} + \frac{1}{\sqrt{\lambda_h}}\right)^2}$ and $\beta_h = \left(\frac{\sqrt{\Lambda_h} - \sqrt{\lambda_h}}{\sqrt{\Lambda_h} + \sqrt{\lambda_h}}\right)^2$.

Hence, leveraging the whole shape of the limiting distribution, as opposed to using only the edge eigenvalues, yields an optimal first-order method which is different, and has non-constant step sizes and momentum parameters. But interestingly, it holds that as the iteration number t grows to $+\infty$, then the update coefficients $a_{h,t}$ and $b_{h,t}$ have respective limits $1 + \beta_h$ and $-\mu_h$, which yields exactly this Heavy-ball method. Thus, we expect the latter and Algorithm 2 to have a similar performance as t grows large.

We complete our analysis of the SRHT case by characterizing the asymptotic error $\mathcal{L}^*_{f_h,t}$.

Theorem 5. *The sequence of iterates $\{x_t\}$ given by Algorithm 2 is asymptotically optimal, and the optimal error satisfies $\mathcal{L}^*_{f_h,t} \asymp \frac{(1-\xi)^t}{(1-\gamma)^t}\,\rho^t$.*

Proof. We have already argued that $\{x_t\}$ is asymptotically optimal. It remains to show that $\mathcal{L}^*_{f_h,t} \asymp \frac{(1-\xi)^t}{(1-\gamma)^t}\,\rho^t$, whose

proof is deferred to Appendix A.4. □

Of natural interest is to compare the rate of convergence $\rho_h := \frac{(1-\xi)}{(1-\gamma)}\rho$ of Algorithm 2 to the rate ρ of Algorithm 1. We have $\frac{\rho_h}{\rho} = \frac{(1-\xi)}{(1-\gamma)}$, which is always smaller than 1 since $\xi > \gamma$. Hence, these rotation matrices yield an optimal first-order method which is uniformly better than that with Gaussian embeddings, by a factor which can be made arbitrarily large by increasing the sketch size m relatively to the other dimensions. Further, 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 by sketching. In contrast, Gaussian projections introduce more distortions than rotations, even though the rows of a Gaussian matrix are almost orthogonal to each other in the high-dimensional setting.

Further, we compare the rate of Algorithm 2 to the rate of the best Heavy-ball method with refreshed SRHT embeddings which is equal to $\rho_h^{\text{ref}} = \rho \cdot \frac{\xi(1-\xi)}{\gamma^2+\xi-2\xi\gamma}$. We have $\rho_h < \rho_h^{\text{ref}}$ if and only if $\frac{1-\xi}{1-\gamma} < \frac{\xi(1-\xi)}{\gamma^2+\xi-2\xi\gamma}$, which is equivalent to $\gamma^2 + \xi - 2\gamma\xi < \xi - \gamma\xi$, again equivalent to $\gamma^2 < \gamma\xi$, i.e., $\gamma < \xi$, which holds by assumption. Thus, a fixed embedding yields a first-order method which is uniformly faster than the best Heavy-ball method with refreshed sketches. However, it remains an open problem whether one can find a first-order method with refreshed sketches which yields a rate better than ρ_h^{ref}. We recapitulate the different convergence rates in Table 1.

Table 1. Asymptotic rates of convergence for the best first-order method (6) and the best Heavy-ball method (2), with fixed or refreshed Gaussian or SRHT embeddings. For the best Heavy-ball method rates, we use previously derived results from (Ozaslan et al., 2019; Lacotte & Pilanci, 2019; Lacotte et al., 2020).

Algorithm	Fixed Gaussian	Refreshed Gaussian	Fixed SRHT	Refreshed SRHT
Best first-order method (6)	ρ	unknown	$\frac{1-\xi}{1-\gamma}\rho$	unknown
Best Heavy-ball method (2)	ρ	ρ	ρ	$\frac{\xi(1-\xi)}{\gamma^2+\xi-2\xi\gamma}$

3. Numerical simulations

First, we generate a synthetic dataset with $n = 8192$, $d = 1600$ and $m \in \{1700, 3500, 5700\}$. Although our results are universal in the sense that it does not depend on the spectral decay of the matrix A, we still consider a challenging setting for first-order methods, that is, we generate an $n \times d$ matrix A with an exponential spectral decay. In Figure 2, we verify numerically that Algorithm 2 is faster than the best Heavy-ball method with refreshed SRHT sketches ("SRHT (refreshed)"), and than Algorithm 1. Further, we compare Algorithm 2 to the Heavy-ball method with fixed SRHT embedding whose parameters are found based on edge eigenvalues analysis, using either our new density f_h ("SRHT (edge eig.)") – as described previously in Section 2.2.1 –, or, the previous bounds derived by (Tropp, 2011) ("SRHT (baseline)"). As predicted, Algorithm 2 performs very similarly to the former, and better than the latter. We mention that we use small perturbations of the algorithmic parameters derived from our asymptotic analysis. Following the notations introduced in Theorem 1, instead of a_t and b_t, we use $a_t^\delta = (1+\delta)a_t$ and $b_t^\delta = (1-\delta)b_t$ with $\delta = 0.01$. These conservative perturbations are necessary in practice due to the finite-sample approximations. We defer a detailed description of the experimental setup to Appendix C. Second, we verify that our predicted convergence

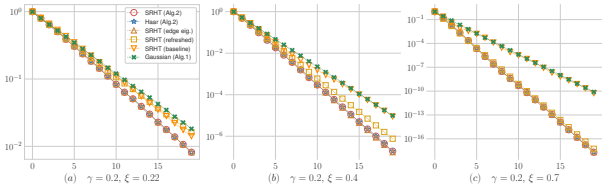

Figure 2. Error $\mathbb{E}\|\Delta_t\|^2/\mathbb{E}\|\Delta_0\|^2$ versus number of iterations. We use $n = 8192$, $d/n \approx \gamma = 0.2$ and $m/n \approx \xi \in \{0.22, 0.4, 0.7\}$.

rates for Algorithms 1 and 2 are matched empirically, on Figure 3. For this purpose, we generate a synthetic dataset with $n = 8192$, $d \in \{500, 1250, 2000\}$ and varying sketching size, with a data matrix having an exponential spectral decay. Lastly, we verify our results on standard machine learning

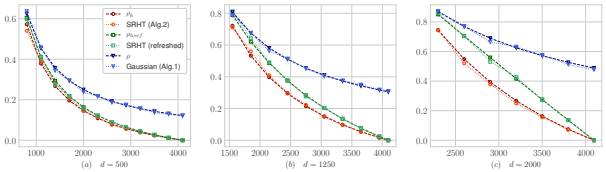

Figure 3. Empirical and theoretical convergence rates versus sketch size m. We use $n = 8192$ and $d \in \{500, 1250, 2000\}$.

benchmarks, that is, we consider the MNIST and CIFAR10 datasets, for which results are respectively reported on Figures 4 and 5. Our observations for these datasets are qualitatively similar to those made on the aforementioned synthetic dataset. This confirms the universality of our methods, i.e., they do not depend on the data considered.

4. Complexity Analysis

We turn to a complexity analysis of Algorithm 2 and compare it to the currently best known algorithmic complexities for solving (1).

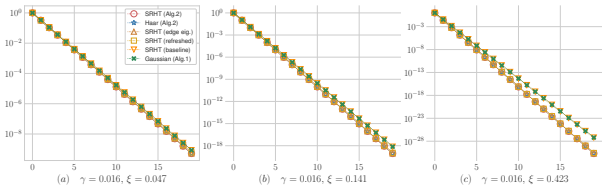

Figure 4. Error $\mathbb{E}\|\Delta_t\|^2/\mathbb{E}\|\Delta_0\|^2$ versus number of iterations, for the MNIST dataset. We have $n = 50000$, $d/n \approx \gamma = 0.015$ and we use $m/n \approx \xi \in \{0.047, 0.141, 0.423\}$.

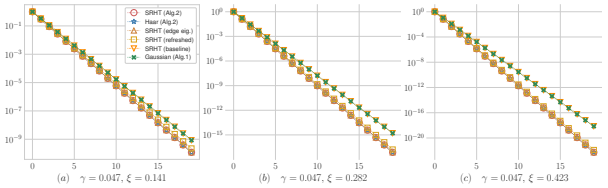

Figure 5. Error $\mathbb{E}\|\Delta_t\|^2/\mathbb{E}\|\Delta_0\|^2$ versus number of iterations, for the CIFAR10 dataset. We have $n = 50000$, $d/n \approx \gamma = 0.047$ and we use $m/n \approx \xi \in \{0.141, 0.282, 0.423\}$.

Given a fixed (and independent of the dimensions) error $\varepsilon > 0$, we aim to find $\widetilde{x}$ such that $\|A(\widetilde{x} - x^*)\|^2 \leqslant \varepsilon$. Among the best complexity algorithms is the pre-conditioned conjugate gradient algorithm (Rokhlin & Tygert, 2008). As described in Section 1, it is decomposed into three parts: sketching the data matrix, factoring the pre-conditioned matrix, and then the iterations of the conjugate gradient method. This algorithm prescribes at least the sketch size $m \asymp d \log d$ in order to converge with high-probability guarantees. This theoretical prescription is based on the finite-sample bounds on the extremal eigenvalues of the matrix C_S derived by (Tropp, 2011). Then, the resulting complexity scales as

$$\mathcal{C}_{\text{cg}} \asymp nd \log d + d^3 \log d + nd \log(1/\varepsilon)\,, \qquad (30)$$

where $nd \log d$ is the sketching cost, $d^3 \log d$ the pre-conditioning cost, and $nd \log(1/\varepsilon)$ is the per-iteration cost nd times the number of iterations $\log 1/\varepsilon$.

Our analysis shows that for $m \asymp d$, Algorithm 2 yields a complexity no larger than

$$\mathcal{C}_{\text{fhs}} \asymp nd \log d + d^3 + nd \log(1/\varepsilon)\,, \qquad (31)$$

Note that in the above complexity, we omit the rate of convergence – which would yield an even smaller complexity – to simplify the comparison. Since ε is independent of the dimensions, it follows that

$$\frac{C_{\text{fhs}}}{C_{\text{cg}}} \asymp \frac{1}{\log d}\,, \quad d \to \infty\,. \qquad (32)$$

Hence, with a smaller sketch size, the resulting complexity improves by a factor $\log d$ over the current state-of-the-art in randomized preconditioning for dense problems (e.g., see (Boutsidis & Gittens, 2013; Nelson & Nguyên, 2013)). We also note that the $O(d^3)$ term can be improved to $O(d^\omega)$, where ω is the exponent of matrix multiplication.

It has also been shown by Lacotte et al. (2020) that the Heavy-ball update (2) with refreshed SRHT embeddings yields a complexity $\mathcal{C}_{\text{ihs}}$ such that $C_{\text{ihs}}/C_{\text{cg}} \asymp \frac{1}{\log d}$, provided that $m \asymp d$. In order to compare more finely Algorithm 2 with this algorithm, we consider an arbitrary sketch size m. Then, the complexity of Algorithm 2 is

$$\mathcal{C}_{\text{fhs}} \asymp nd \log m + md^2 + nd \frac{\log(1/\varepsilon)}{\log \rho_h}\,, \qquad (33)$$

whereas the former algorithm yields

$$\mathcal{C}_{\text{ihs}} \asymp \left(nd \log m + md^2 + nd\right) \frac{\log(1/\varepsilon)}{\log \rho_h^{\text{ref}}}\,. \qquad (34)$$

Since, in particular, ρ_h is uniformly smaller than ρ_h^{ref}, it always holds that

$$\mathcal{C}_{\text{fhs}} \leqslant \mathcal{C}_{\text{ihs}}\,. \qquad (35)$$

It should be noted that we translate our asymptotic results to finite-sample versions. Although it is beyond our scope, we believe that our results could be extended to finite-sample versions with high-probability guarantees and with similar rates of convergence.

Acknowledgements

The authors thank Sifan Liu and Edgar Dobriban for helpful discussions, and the reviewers for their careful feedback and suggestions. This work was partially supported by the National Science Foundation under grant IIS-1838179.

References

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

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

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

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

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

Boutsidis, C. and Gittens, A. Improved matrix algorithms via the subsampled randomized hadamard transform. *SIAM Journal on Matrix Analysis and Applications*, 34 (3):1301–1340, 2013.

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

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

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

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

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

Jiang, T. Approximation of haar distributed matrices and limiting distributions of eigenvalues of jacobi ensembles. *Probability theory and related fields*, 144(1-2):221–246, 2009.

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

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

Lacotte, J., Liu, S., Dobriban, E., and Pilanci, M. Limiting spectrum of randomized hadamard transform and optimal iterative sketching methods. *arXiv:2002.00864*, 2020.

Liu, S. and Dobriban, E. Ridge regression: Structure, cross-validation, and sketching. 2019.

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

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

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

Nelson, J. and Nguyên, H. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 IEEE 54th annual symposium on foundations of computer science*, pp. 117–126. IEEE, 2013.

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

Ozaslan, I., Pilanci, M., and Arikan, O. Iterative hessian sketch with momentum. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7470–7474. IEEE, 2019.

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

Pedregosa, F. and Scieur, D. Acceleration through spectral modeling. *NeurIPS workshop "Beyond First Order Methods in ML"*, 2019.

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

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

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

Rutishauser, H. Theory of gradient methods. In *Refined iterative methods for computation of the solution and the eigenvalues of self-adjoint boundary value problems*, pp. 24–49. Springer, 1959.

Silverstein, J. W. and Choi, S.-I. Analysis of the limiting spectral distribution of large dimensional random matrices. *J. Multivariate Anal.*, 54(2):295–309, 1995.

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

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

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

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

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