Abstract

We provide an exact analysis of a class of randomized methods 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 exhibit 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,

\[ x^* := \arg\min_{x \in \mathbb{R}^d} \left\{ f(x) := \frac{1}{2} \| Ax - b \|^2 \right\}, \]

where \( A \in \mathbb{R}^{n \times d} \) is a given data matrix with \( n \geq d \) and \( b \in \mathbb{R}^n \) is a vector of observations. For simplicity of notations, we will assume throughout this work that \( \text{rank}(A) = d \).

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 \( \mathcal{O}(nm^2) \) 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 \( m \geq d \). 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.

It has been observed in several contexts that 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). Consequently, along with computational considerations, this suggests to consider the SRHT over Gaussian or Haar projections. On the other hand, in order to pick optimal algorithm’s parameters, it is usually necessary to have a tight characterization of the spectrum of \( CS := U^\top S^\top SU \), where \( U \) is the matrix of left singular vectors of \( A \), which is the case for Gaussian embeddings.

Using the standard prediction (semi-)norm \( \| A(\tilde{x} - x^*) \|^2 \) as the evaluation criterion for an approximate solution \( \tilde{x} \), iterative methods (e.g., gradient descent or the conjugate

---

Optimal Randomized First-Order Methods for Least-Squares Problems

Jonathan Lacotte 1  Mert Pilanci 1

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

gradient algorithm) have time complexity which usually scales proportionally to the condition number \( \kappa \) 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 \( H_S = A^\top S^\top SA \). Then, we consider the following class of pre-conditioned first-order methods (also referred to as a quasi-Newton method in the optimization literature),

\[
x_t \in x_0 + H_S^{-1} \cdot \operatorname{span}\{\nabla f(x_0), \ldots, \nabla f(x_{t-1})\},
\]

Several versions of (2) have been studied. For instance, it has been recently shown by (Ozaslan et al., 2019) that the Heavy-ball update

\[
x_{t+1} = x_t - \mu_t H_S^{-1} \nabla f(x_t) + \beta_t (x_t - x_{t-1})
\]

(3)
yields a sequence of iterates whose convergence rate does not depend on the spectrum of \( A \), but only on the concentration of the matrix \( C_S \) around the identity, and 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 \( \rho \), edge eigenvalues \( a = (1 - \sqrt{\rho})^2 \) and \( b = (1 + \sqrt{\rho})^2 \), and density

\[
\mu_p(x) = \frac{\sqrt{(b - x)(x - a)}}{2\pi \rho x},
\]

(4)

where \( y_+ = \max\{y, 0\} \), and it does not depend on the sample size \( n \) but only on the limit ratio \( \rho = \lim d/m \).

For a SRHT embedding, it is unclear if the dimension \( n \) affects the best achievable convergence rate. In a related vein, the authors of Lacotte et al. (2020) considered the Heavy-ball update (3) 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 \( \rho_{\text{SRHT}} = \rho \cdot \frac{(1 - \xi)}{\sqrt{\rho - 2\xi+2}} \), which indeed depends on the three relevant dimensions \( m, d, n \) through the aspect ratios \( \rho \), \( \gamma := \lim d/m \) and \( \xi := \lim m/n \), which is always strictly smaller than \( \rho \), the convergence rate one would obtain with refreshed Gaussian embeddings (Lacotte & Pilanci, 2019).

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 \( p \geq 1 \), 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 \( O(n) \) compared to the cost \( O(n d \log m) \) 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 \( \pm 1 \) 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 \( \tilde{m} \times n \) orthogonal matrix with \( \tilde{m} \sim \operatorname{Binomial}(m/n, n) \), and the ratio \( \tilde{m}/n \) concentrates fast around \( \xi \) while \( n \to \infty \). Although the dimension \( \tilde{m} \) is random, we refer to \( S \) as an \( m \times n \) SRHT matrix.

We will focus exclusively on (pre-conditioned) first-order methods of the form (2) 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 \( \rho = \frac{d}{\tilde{m}} \in (0, 1) \). 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 \( ||\Delta_t||_2^2 = ||A(x_t - x^*)||_2^2 \). Our evaluation criterion is the error \( \lim_{n \to \infty} \mathbb{E}[||\Delta_t||_2^2]/\mathbb{E}[||\Delta_0||_2^2] \), and we call its (asymptotic) rate of convergence the quantity \( \limsup_{n \to \infty} \left( \lim_{n \to \infty} \mathbb{E}[||\Delta_t||_2^2]/\mathbb{E}[||\Delta_0||_2^2] \right)^{1/2} \).

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., \( F_{M_n}(x): = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\{\lambda_j \leq x\} \) for \( x \in \mathbb{R} \), which has density \( f_{M_n}(x): = \frac{1}{n} \sum_{j=1}^{n} \delta_{\lambda_j}(x) \) with \( \delta_{\lambda_j} \) the Dirac measure at \( \lambda_j \). 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 \( M_n \).
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 \( \mathbb{R}_t^d[X] \) 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 \( \text{Re}(z) \) and \( \text{Im}(z) \) 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 \( \lim \inf \frac{a_t}{b_t} > 0 \) and \( \lim \sup \frac{a_t}{b_t} < \infty \).

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.

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 \( \lim_{n \to \infty} \mathbb{E} \| \Delta t \| / \mathbb{E} \| \Delta 0 \| \), 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 rate of the error is equal to \( \rho^t \).

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 \( \rho_{\text{ref}} := \rho \cdot 1 - \frac{\gamma}{1 - \gamma} \), which is always strictly smaller than \( \rho \) and \( \rho_{\text{ref}} \), 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 \( \frac{1}{m} C_S \), which is given by

\[
 f_{h,r}(x) = \frac{\sqrt{(\lambda_{h,r} - x) + (\lambda_{h,r})}}{2\pi \rho \varepsilon (1 - \xi x)},
\]

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

The design of optimal first-order methods for quadratic optimization problems has been recently considered in Pedregosa & Scieur (2019). In contrast, they assume the data matrix to be random and they require its spectrum to be known beforehand, which is often impractical. On the other hand, our class of first-order methods applies a randomized pre-conditioning, so 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 are able to optimize the exact error for any data matrix \( A \). We note that existing methods do not directly minimize the error, but a worst-case upper bound as in Chebyshev iteration and Conjugate Gradient methods (Saad, 2003).

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 \( S \) – such that the condition number of \( AP^{-1} \) is \( O(1) \), 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; Wit- ten & 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.

2. Optimal first-order method for classical embeddings

Let \( S \) be an \( m \times n \) Gaussian or SRHT embedding. Denote by \( \mu \) the l.s.d. of \( C_S \). We say that a family of polynomials \( \{ R_k \} \) is orthogonal with respect to \( \mu \) if \( \int R_k R_\ell d\mu = 0 \) for any \( k \neq \ell \). The next result establishes the link between polynomials and the pre-conditioned first-order methods (2) we consider, and its proof is deferred to Appendix B.1.

Lemma 2.1. Let \( \{ x_t \} \) be generated by some first-order method (2). Then, for any iteration \( t \geq 0 \), there exists a
polynomial \( p_t \in \mathbb{R}^t[X] \) such that \( \Delta_t = p_t(C_S^{-1}) \cdot \Delta_0 \).

Further, it holds that
\[
\lim_{n \to \infty} \frac{\mathbb{E}[\|\Delta_t\|^2]}{\mathbb{E}[\|\Delta_0\|^2]} = \int_{\mathbb{R}} p_t^2(\lambda^{-1}) \, d\mu(\lambda) .
\]

(6)

Thus, the best achievable error is lower bounded by the infimum of the following variational problem,
\[
\mathcal{L}_{\mu,t}^* := \min_{p \in \mathbb{R}^t[X]} F_{\mu}(p),
\]

(7)

where \( F_{\mu}(p) := \int p^2(\lambda^{-1}) \, d\mu(\lambda) \). Using the change of variable \( x = 1/\lambda \) and setting \( d\nu(x) = x^{-1}d\mu(x^{-1}) \), we have that \( F_{\mu}(p) = G_{\nu}(p) \) where \( G_{\nu}(p) := \int p^2(x) \frac{1}{2} d\nu(x) \).

The optimal polynomial can be constructed by leveraging the following result.

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

Proof. Let \( p \in \mathbb{R}^0_0[X] \). 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,
\[
G_{\nu}(p) = \int p^2(x) x^{-1} d\nu(x)
= \int \Pi_t^2(x) x^{-1} d\nu(x) + 2 \int \Pi_t Q(x) d\nu(x)
+ \int xQ^2(x) d\nu(x).
\]

The cross-term is equal to 0 since \( Q \) in the span of \( \{\Pi_0, \ldots, \Pi_{t-1}\} \), which are orthogonal to \( \Pi_t \). The third term is non-negative, and equal to 0 if and only if that \( Q = 0 \). Therefore, the unique solution to (7) is \( \Pi_t \). \( \square \)

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 \( t \geq 2 \),
\[
\Pi_t(x) = (a_t + b_t x) \Pi_{t-1}(x) + (1 - a_t) \Pi_{t-2}(x) .
\]

(8)

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 \( t \geq 2 \),
\[
x_t = x_{t-1} + b_t H_S^{-1} \nabla f(x_{t-1}) + (1 - a_t) (x_{t-2} - x_{t-1}) .
\]

(9)

Then, the sequences of iterates \( \{x_t\} \) is asymptotically optimal, i.e.,
\[
\lim_{n \to \infty} \frac{\mathbb{E}[\|\Delta_t\|^2]}{\mathbb{E}[\|\Delta_0\|^2]} = \mathcal{L}_{\mu,t}^* .
\]

(10)

Consequently, a strategy to find the optimal first-order method proceeds as follows. First, we characterize the l.s.d. \( \mu \) of the matrix \( C_S \), and we find the polynomial \( \Pi_t(0) = 1 \) which achieves the lower bound \( \mathcal{L}_{\mu,t}^* \). Then, according to Theorem 1, we build from the three-terms recursion (8) 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. \( \mu \) of \( C_S \), and in constructing an orthogonal basis of polynomials for the distribution \( d\nu(x) = x^{-1}d\mu(x^{-1}) \).

### 2.1. The Gaussian case

Consider an \( m \times n \) matrix \( S \) with i.i.d. entries \( \mathcal{N}(0, m^{-1}) \). The l.s.d. \( \mu \) of \( C_S \) is the Marchenko-Pastur law with density \( \mu_{\rho} \) 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 (2). According to Lemma 2.1, there exists a sequence of polynomials \( p_t \in \mathbb{R}^t[X] \) such that \( \Delta_t = p_t(C_S^{-1}) \Delta_0 \), and
\[
\lim_{n \to \infty} \frac{\mathbb{E}[\|\Delta_t\|^2]}{\mathbb{E}[\|\Delta_0\|^2]} = \int_a^b p_t^2(\lambda^{-1}) \mu_{\rho}(\lambda) d\lambda ,
\]

(11)

**Lemma 2.3.** Under the above assumptions and notations, and setting \( P_t(x) = p_t \left( \frac{1}{1 - \rho} \right)^{1/2} \), we have
\[
\lim_{n \to \infty} \frac{\mathbb{E}[\|\Delta_t\|^2]}{\mathbb{E}[\|\Delta_0\|^2]} = (1 - \rho) \int_a^b P_t^2(x) \frac{1}{x} \mu_{\rho}(x) dx .
\]

(12)

Consequently, if \( \{\Pi_t\} \) is an orthogonal basis of polynomials with respect to \( \mu_{\rho} \) such that \( \deg(\Pi_t) = t \) and \( \Pi_t(0) = 1 \) then \( \Pi_t(x) := \Pi_t((1 - \rho)^2 x) \) 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 \( p_t^2(\lambda^{-1}) \mu_{\rho}(\lambda) d\lambda = (1 - \rho) P_t^2(x) \frac{1}{x} \mu_{\rho}(x) dx \). Applying Lemma 2.2 with \( \nu = \mu_{\rho} \), we get that the optimal polynomial \( p_t \) is equal to \( \Pi_t \), and thus, \( p_t \) is exactly \( \Pi_t(x) \). \( \square \)

The Marchenko-Pastur law \( \mu_{\rho} \) 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 $t \geq 2$,

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

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

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

Now, set $\Pi_0(x) = \Pi_1((1 - \rho)^2 x)$. From (13), we obtain that $\Pi_0(x) = 1$, $\Pi_1(x) = 1 - (1 - \rho)^2 x$, and for $t \geq 2$,

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

(14)

According to Lemma 2.3, the polynomial $\Pi_t$ achieves the lower bound $L_{\mu_\rho,t}$.

Further, we identify the recursion formula (14) with the three-terms recursion (8) by setting $b_t = -(1 - \rho)^2$ for $t \geq 1$, and $a_t = 1 + \rho$ for $t \geq 2$.

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 $m \geq d + 1$, 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

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

end for

Return the last iterate $x_T$.

Surprisingly, up to the initialization of the first iterate $x_1$, Algorithm 1 corresponds exactly to the Heavy-ball method (3) 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 $L_{\mu_\rho,t}$.

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

Proof. We have already argued that $\{x_t\}$ is asymptotically optimal. It remains to show that $L_{\mu_\rho,t} = \rho^t$, 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 $m_\mu$ of a distribution $\mu$ supported on $[0, +\infty)$, which, for $z \in \mathbb{C} \setminus \mathbb{R}_+$, is given by $m_\mu(z) := \int \frac{1}{z - x} d\mu(x)$. 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

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

(16)

where

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

**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 $\mathbb{R}$,

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

(17)

where

$$
\lambda_h := \left( \sqrt{(1 - \gamma)\xi - (1 - \xi)\gamma} \right)^2,
$$

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

Proof. The proof is essentially based on the expression (16) of the Stieltjes transform $m_h$, and on the inversion formula,

$$
f_h(x) = \lim_{y \to 0^+} \frac{1}{\pi} \text{Im} \left( m_h(x + iy) \right),
$$

where $y \in \mathbb{R}_+$.

(18)
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

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

for

$$\left\{ \begin{array}{l}
\lambda_{h,r} = \lambda_{h}/\xi = \left(\sqrt{1 - \gamma - \sqrt{(1 - \xi)\rho}}\right)^2 ,
\lambda_{h,r} = \Lambda_{h}/\xi = \left(\sqrt{1 - \gamma + \sqrt{(1 - \xi)\rho}}\right)^2 .
\end{array} \right.$$  \hfill (19)

The density $f_{h,r}$ resembles the Marchenko-Pastur density $\mu_{\rho}$ 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 $f_{h,r}(x)$ is close to $\mu_{\rho}(x)$. 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 $\rho$ (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 $\mu_{\rho}$, and then concentrates more and more around 0. Note in particular that the support of $f_{h,r}$ is always within that of $\mu_{\rho}$. This can be formally verified by comparing their respective edge eigenvalues.

![Figure 1](image.png)

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 (2), we know from Lemma 2.1 that for a given iteration $t$, there exists a polynomial $p \in \mathbb{R}^{0}[X]$ such that $\Delta_{t} = p(C_{S}^{-1}) \Delta_{0}$, and

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}[\|\Delta_{t}\|^{2}]}{\mathbb{E}[\|\Delta_{0}\|^{2}]} = \int_{\lambda_{h}}^{\lambda_{h}} p^{2}(\lambda^{-1}) f_{h}(\lambda) d\lambda .$$

(20)

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

$$\lim_{n \rightarrow \infty} \frac{\mathbb{E}[\|\Delta_{t}\|^{2}]}{\mathbb{E}[\|\Delta_{0}\|^{2}]} = \frac{c \tau}{(1 - \tau)^{1/2}} \int_{C} P^{2}(x) \sqrt{(x - \alpha)/(\beta - x)} \frac{d\alpha}{2 \pi \tau x}$$

$$\int_{C} P^{2}(x) \frac{d\alpha}{x - c} d\alpha$$

Thus, according to Lemma 2.2, it suffices to find a family of polynomials $\{R_{t}\}$ orthogonal with respect to the density $x^{2m+2}/2\pi \rho$ such that $\deg(\Pi_{t}) = t$ and $R_{t}(0) = 1$, in which case the minimizer over $P \in \mathbb{R}^{0}[X]$ of the integral in (23) is equal to $\Pi_{t}$, and the minimizer of (20) is then $\Pi_{t}(x) = R_{t}(cx)$.

**Theorem 4.** Define the parameters $\omega = (\sqrt{\beta} - \sqrt{\alpha})^2$ and $\kappa = \left(\frac{\sqrt{\beta} - \sqrt{\alpha}}{\sqrt{\beta} + \sqrt{\alpha}}\right)^2$. Let $\{\Pi_{t}\}$ be the orthogonal family of polynomials with respect to $\mu_{\kappa}$, that is, $\Pi_{0}(x) = 1, \Pi_{1}(x) = 1 - x, \text{ and for } t \geq 2$, $\Pi_{t}(x) = (1 + \kappa - x)\Pi_{t-1}(x) - \kappa \Pi_{t-2}(x) .$$

(24)

Define the polynomials $R_{t}(x) = \Pi_{t}(\omega(x - c))/\Pi_{t}(\omega - c)$. Then, it holds that $R_{t}(0) = 1$, $\deg(\Pi_{t}) = t$, and the family $\{R_{t}\}$ is orthogonal with respect to the density $x^{2m+2}/2\pi \rho$.

**Proof.** For $k \neq t$, we have that $\int R_{k}(x) R_{t}(x) \frac{dx}{x - c} \propto \int_{C} \Pi_{k} p_{k}(\omega(x - c)) \Pi_{t} p_{t}(\omega(x - c)) \frac{dx}{2\pi \rho(x - c)}$. Using the change of variable $y = \omega(x - c)$, we find that the latter integral is (up to a constant) equal to $\int_{C} \Pi_{k}(y) \Pi_{t}(y) \mu_{\kappa}(y) dy$, which is itself equal to 0 due to the orthogonality of the $\Pi_{t}$ with respect to $\mu_{\kappa}$.

In order to derive the optimal first-order method, we need to find the three-terms recursion relationship satisfied by the polynomials $\{R_{t}\}$. First, let us compute the normalization factor $u_{t} := \Pi_{t}(\omega)$. Evaluating (24) at $x = \omega - c$ and denoting $\eta := 1 + \kappa - \omega$, we find that $u_{t+1} = \eta u_{t} - \kappa u_{t-1}$, with the initial conditions $u_{0} = 1$ and $u_{1} = \Pi_{1}(\omega - c) = 1 + \omega - \kappa = \eta - \kappa$. Thus, after solving this second-order linear system, we obtain that

$$u_{t} = \frac{x_{1} - \kappa}{x_{1} - x_{2}} x_{1} + \frac{\kappa - x_{2}}{x_{1} - x_{2}} x_{2},$$

(25)

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 (24), we get the following three-terms recurrence
Optimal Randomized First-Order Methods for Least-Squares Problems

relationship, that is, \( \mathcal{R}_0(x) = 1 \), \( \mathcal{R}_1(x) = 1 + b_{h,1}x \) and for \( k \geq 2 \),
\[
\mathcal{R}_k(x) = (a_{h,t} + x b_{h,t}) \mathcal{R}_{k-1}(x) + (1 - a_{h,t}) \mathcal{R}_{k-2}(x),
\]
where \( a_{h,t} = \frac{\mu_{h,t-1}}{\lambda_{h,t}} \) for \( t \geq 1 \), and \( b_{h,t} = -\frac{\omega_{h,t-1}}{\gamma_{h,t}} \) for \( t \geq 2 \). 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 \( m \geq d + 1 \), 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
\[
x_t = x_{t-1} + b_{h,t} H_S^{-1} g_t + (1 - a_{h,t}) (x_{t-2} - x_{t-1}).
\]

end for

Return the last iterate \( x_T \).

Differently from the Gaussian case, Algorithm 2 does not correspond to the Heavy-ball method (3) 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 \( \mu_h \) and momentum parameter \( \beta_h \) are given by
\[
\mu_h = \left( \frac{1}{\sqrt{\Lambda_h}} + \frac{1}{\sqrt{\Lambda_{h+\gamma}}} \right)^2 \quad \text{and} \quad \beta_h = \left( \frac{\sqrt{\Lambda_{h+\gamma}} - \sqrt{\Lambda_h}}{\sqrt{\Lambda_{h+\gamma}} + \sqrt{\Lambda_h}} \right)^2.
\]

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 \( \mathcal{L}_{f_h,t}^* \lesssim (1 - \xi)^t \rho^t \).

**Proof.** We have already argued that \( \{x_t\} \) is asymptotically optimal. It remains to show that \( \mathcal{L}_{f_h,t}^* \lesssim (1 - \xi)^t \rho^t \), whose proof is deferred to Appendix A.4. \( \square \)

Of natural interest is to compare the rate of convergence \( \rho_h := \frac{(1 - \xi_t)}{(1 - \xi)} \rho \) of Algorithm 2 to the rate \( \rho \) of Algorithm 1.

We have \( \rho_h = \frac{(1 - \xi)}{(1 - \gamma)} \rho \), 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 \( \rho_h^\text{ref} = \rho \cdot \frac{(1 - \xi)}{\gamma} \cdot \frac{1 - \xi}{\gamma + \xi - 2\xi^2} \). We have \( \rho_h < \rho_h^\text{ref} \) if and only if \( \frac{1 - \xi}{\gamma} < \gamma \), which is equivalent to \( \gamma^2 + \xi - 2\xi^2 < \xi - \gamma \xi \), 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 \( \rho_h^\text{ref} \). We recapitulate the different convergence rates in Table 1.

**Table 1.** Asymptotic rates of convergence for the best first-order method (2) and the best Heavy-ball method (3), 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).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Fixed Gaussian</th>
<th>Refreshed Gaussian</th>
<th>Fixed SRHT</th>
<th>Refreshed SRHT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best first-order method (2)</td>
<td>( \rho )</td>
<td>( \rho )</td>
<td>( \frac{1 - \xi}{\gamma} \rho )</td>
<td>( \frac{1 - \xi}{\gamma} \rho )</td>
</tr>
<tr>
<td>Best Heavy-ball method (3)</td>
<td>( \rho )</td>
<td>( \rho )</td>
<td>( \frac{(1 - \xi)}{\gamma} \rho )</td>
<td>( \frac{(1 - \xi)}{\gamma} \rho )</td>
</tr>
</tbody>
</table>
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. Finally, we verify that our predicted convergence rates for Algorithms 1 and 2 are matched empirically, on Figure 3.

We turn to a complexity analysis of Algorithm 2 and compare it to the currently best known algorithmic complexities for solving (1). Given a fixed (and independent of the dimensions) error \( \varepsilon > 0 \), we aim to find \( \tilde{x} \) such that \( \| A (\tilde{x} - x^*) \|^2 \leq \varepsilon \). 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 \( m \approx d \log d \) 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

\[
C_{\text{cg}} \approx nd \log d + d^3 \log d + nd \log(1/\varepsilon),
\]

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 \( m \approx d \), Algorithm 2 yields a complexity no larger than

\[
C_{\text{bas}} \approx nd \log d + d^3 + nd \log(1/\varepsilon),
\]

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

\[
\frac{C_{\text{bas}}}{C_{\text{cg}}} \approx \frac{1}{\log d}, \quad d \to \infty.
\]

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 & Nguyen, 2013)). We also note that the \( O(d^3) \) term can be improved to \( O(d^2) \), where \( \omega \) is the exponent of matrix multiplication.

It has also been shown by Lacotte et al. (2020) that the Heavy-ball update (3) with refreshed SRHT embeddings yields a complexity \( C_{\text{hs}} \) such that \( C_{\text{hs}}/C_{\text{cg}} \approx \frac{1}{\log^2 d} \), provided that \( m \approx d \). 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

\[
C_{\text{hs}} \approx nd \log m + md^2 + nd \frac{\log(1/\varepsilon)}{\log \rho_h^\text{ref}},
\]

whereas the former algorithm yields

\[
C_{\text{hs}} \approx (nd \log m + md^2 + nd) \frac{\log(1/\varepsilon)}{\log \rho_h^\text{ref}}.
\]

Since, in particular, \( \rho_h \) is uniformly smaller than \( \rho_h^\text{ref} \), it always holds that

\[
C_{\text{hs}} \leq C_{\text{hs}}.
\]

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.

---

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

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

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

References


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


