Effective Dimension Adaptive Sketching Methods for Faster Regularized Least-Squares Optimization

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

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

Abstract

We propose a new randomized algorithm for solving L2-regularized least-squares problems based on sketching. We consider two of the most popular random embeddings, namely, Gaussian embeddings and the Subsampled Randomized Hadamard Transform (SRHT). While current randomized solvers for least-squares optimization prescribe an embedding dimension at least greater than the data dimension, we show that the embedding dimension can be reduced to the effective dimension of the optimization problem, and still preserve high-probability convergence guarantees. In this regard, we derive sharp matrix deviation inequalities over ellipsoids for both Gaussian and SRHT embeddings. Specifically, we improve on the constant of a classical Gaussian concentration bound whereas, for SRHT embeddings, our deviation inequality involves a novel technical approach. Leveraging these bounds, we are able to design a practical and adaptive algorithm which does not require to know the effective dimension beforehand. Our method starts with an initial embedding dimension equal to 1 and, over iterations, increases the embedding dimension up to the effective one. Finally, we prove that our algorithm improves the state-of-the-art computational complexity for solving regularized least-squares problems. Further, we show numerically that it outperforms standard least-squares solvers such as the conjugate gradient method and its pre-conditioned version on several standard machine learning datasets.

1 Introduction

We study the performance of a randomized method, namely, the Hessian sketch [22], in the context of regularized least-squares problems,

\[ x^* = \arg\min_{x \in \mathbb{R}^d} \left\{ f(x) : = \frac{1}{2} \|Ax - b\|_2^2 + \frac{\nu}{2} \|x\|_2^2 \right\}, \tag{1} \]

where \( A \in \mathbb{R}^{n \times d} \) is a (dense) data matrix and \( b \in \mathbb{R}^n \) is a vector of observations. For clarity purposes and without loss of generality (by considering instead the dual problem of (1)), we make the assumption that the problem is over-determined, i.e., \( n \geq d \) and that \( \text{rank}(A) = d \).

The regularized solution \( x^* \) can be obtained using direct methods based on factorization methods (e.g., Cholesky or QR decomposition) and this has computational complexity \( O(nd^2) \). In the large-scale setting \( n, d \gg 1 \), this is prohibitively large. A linear dependence \( O(nd) \) is preferable and this can be obtained by using first-order iterative solvers [10] such as the conjugate gradient method (CG) for which the per-iteration complexity scales as \( O(nd) \). Using the standard prediction (semi)-norm error \( \|\overline{A}(\overline{x} - x^*)\|_2^2 \) where \( \overline{A} := \begin{bmatrix} A \\ \nu \mathcal{I}_d \end{bmatrix} \) as the evaluation criterion for an estimator \( \overline{x} \), these iterative methods
We denote by $\|A\|$ (or $\sqrt{\kappa}$ with acceleration) in order to find a solution $\hat{x}$ with acceptable accuracy. This also becomes prohibitively large when $\kappa \gg 1$. Besides the computational complexity, the number of iterations of an iterative solver is also a relevant performance metric in the large-scale setting, as distributed computation may be necessary at each iteration. In this regard, randomized preconditioning methods [24, 2, 19] involve using a random matrix $S \in \mathbb{R}^{m \times n}$ with $m \ll n$ to project the data $A$, and then improve the condition number of $A$ based on a spectral decomposition of $SA$. On the other hand, the iterative Hessian sketch (IHS) introduced by [22] and considered in [20, 14, 15, 21] addresses the conditioning issue differently. Given $x_0, x_1 \in \mathbb{R}^d$, it uses a pre-conditioned Heavy-ball update with step size $\mu$ and momentum parameter $\beta$, given by

$$x_{t+1} = x_t - \mu H_S^{-1} \nabla f(x_t) + \beta(x_t - x_{t-1})$$  \hspace{1cm} (2)

where the Hessian $A^T A$ of $f(x)$ is approximated by $H_S := A^T S^T S A$ and $S$ is a sketching matrix. We refer to the update (2) as the Polyak-IHS method, and, in the absence of acceleration ($\beta = 0$), we call it the Gradient-IHS method. In contrast to preconditioning methods [24, 2, 19], the IHS does not need to pay the full cost $O(md \min\{m, d\})$ for decomposing the matrix $SA$. Although solving exactly the linear system $H_S \cdot z = \nabla f(x_t)$ also takes time $O(md \min\{m, d\})$, approximate solving (using for instance CG) is also efficient and faster in practice [21, 20].

The choice of the sketching matrix $S$ is critical for statistical and computational performances. A classical sketch is a matrix $S$ with independent and identically distributed (i.i.d.) Gaussian entries $\mathcal{N}(0, m^{-1})$ for which forming $SA$ requires in general $O(nd \log m)$ time, and this is much faster than Gaussian projections. Consequently, along with the statistical benefits of orthogonal projections, this suggests to use the SRHT as a reference point for comparing sketching algorithms.

In the context of unregularized least-squares problems ($\nu = 0$), [14] showed that the error $\|A(x_t - x^*):m\|^2$ of the Polyak-IHS method is smaller than $(d/m)^t$ for both Gaussian and SRHT matrices provided that $m \approx d$. More recently, it has been shown in [15] that the scaling $(d/m)^t$ is exact for Gaussian embeddings in the asymptotic regime where we let the relevant dimensions go to infinity, whereas the exact scaling for the SRHT is slightly smaller than $(d/m)^t$.

In the regularized case ($\nu > 0$), more relevant than the matrix rank is the effective dimension $d_e \approx \sum_{i=1}^d \sigma_i^2 / \sigma_{i+1}$ (where the $\sigma_i$’s are the singular values of $A$) which always satisfies $d_e \leq d$, and it is significantly smaller than $d$ when the $\sigma_i$’s have a fast decay. It has been shown in [21] that one can pick $m \approx d_e$ and achieve the error rate $(d_e/m)^t$ by using the well-structured approximate Hessian

$$H_S := A^T S^T S A + \nu^2 \cdot I.$$  \hspace{1cm} (3)

However, it is necessary to estimate $d_e$ (which is usually unknown) to be able to pick $m \approx d_e$ and achieve significant computational and memory space savings. The randomized technique proposed by [3] can be used to estimate $d_e$, but under the restrictive assumption that $d_e$ is very small (e.g., see Theorem 60 in [3]). In [21], the authors propose to use a heuristic Hutchinson-type trace estimator [4] and do not provide any guarantee on the estimation accuracy of $d_e$. Consequently, our main goal in this paper is to design an adaptive algorithm which does not require the knowledge of $d_e$, but is still able to use a sketch size $m \lesssim d_e$ and achieve an error rate $(d_e/m)^t$.

Except for the work [21] that also considers the update (2) with $m \approx d_e$ (and a heuristic estimation method for $d_e$), state-of-the-art randomized preconditioning methods [24, 2, 19] prescribe to use $m$ proportional to $d$. Since it seems non-trivial to either adapt and analyze these methods for sketch sizes $m \approx d_e$ in the context of regularized least-squares, nor to design an adaptive scheme which does not require the knowledge of $d_e$, we focus our attention to the Polyak-IHS method in this work.

1.1 Notations

We denote by $\|z\|$ or $\|z\|_2$ 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 define its error at time
We start with providing deterministic convergence guarantees for the Polyak- and Gradient-IHS methods, and we relate the convergence rates to the extreme eigenvalues of the matrix $C_S$. Similar guarantees were established in [21].

Let $S \in \mathbb{R}^{m \times d}$ be any sketching matrix with arbitrary sketch size $m$, and denote by $\gamma_1$ (resp. $\gamma_d$) the largest (resp. smallest) eigenvalue of $C_S$. Since the matrix $D^T U^T S^T S U D$ is positive semi-definite

$t$ as $\delta_t := \frac{1}{2} \|A(x_t - x^*)\|^2$, and its convergence rate as $\limsup_{t \to \infty} (\delta_t/\delta_0)^{\frac{1}{t}}$. We introduce the diagonal matrix $D := \text{diag} \left( \frac{\sigma_1}{\sqrt{\sigma_1^2 + \nu^2}}, \ldots, \frac{\sigma_d}{\sqrt{\sigma_d^2 + \nu^2}} \right)$ where $\sigma_1 \geq \ldots \geq \sigma_d$ are the singular values of the matrix $A$. We define the effective dimension as $d_e := \frac{\|D\|_F^2}{\|D\|_2^2}$. Finally, we denote by $U \in \mathbb{R}^{n \times d}$ a matrix of left singular vectors of $A$, $\overline{A} := \begin{bmatrix} A & \nu \cdot I_d \end{bmatrix}$ and $\overline{U} \in \mathbb{R}^{(n+d) \times d}$ a matrix of left singular vectors of $\overline{A}$. Critical to our convergence analysis is the matrix $C_S := D(U^T S^T S U - I_d)D + I_d$.

1.2 Overview of our contributions

Our main contribution is to propose an algorithm based on the Polyak-IHS update (2) that does not require the knowledge of $d_e$, and is still able to achieve the error rate $O\left((d_e/m)^4\right)$. Our algorithm starts with an arbitrary $m$ (e.g. $m = 1$) and, at each iteration, it uses an improvement criterion to decide if it should increase $m$ or not. We prove that the adaptive sketch size satisfies at each iteration $m \leq d_e$ and that our algorithm improves on the state-of-the-art computational complexity for solving regularized least-squares problems.

Our algorithmic parameters and improvement criterion depend on the extreme eigenvalues of $C_S$, and it is then critical for optimal performance to have a sharp estimation of these. For Gaussian embeddings, we improve on the constant of the best known concentration bound [13]. Our constant is sharp in a worst-case sense, and our analysis is based on a recent extension [25] of Gordon’s min-max theorem [9]. In the SRHT case, although a similar concentration bound was obtained by [6] (see Theorem 1), we provide a novel technical approach that generalizes the analysis techniques from the work [26].

We evaluate numerically our adaptive algorithm on several standard datasets. We consider two settings: (i) the regularization parameter $\nu$ is fixed; (ii) one aims to compute the several solutions along a regularization path. The latter setting is more relevant to many practical applications [29, 12] where estimating a proper regularization parameter is essential. In both cases, we show that our method is faster than the standard conjugate gradient method and one of the state-of-the-art randomized preconditioning methods [24].

Finally, we consider the underdetermined case $d \geq n$. By taking the dual problem of (1) which is itself an overdetermined least-squares problems, we show that our adaptive algorithm and theoretical guarantees easily extend to this setting. We defer the presentation of these results to Appendix B.

1.3 Other related work

Other versions of the IHS have been proposed in the literature, especially in the context of unregularized least-squares. A fundamentally different version uses the same update (2) but with refreshed sketching matrices, i.e., a new matrix $S$ is sampled at each iteration and independently of the previous ones, and the approximate Hessian $H_S$ is re-computed. Surprisingly, refreshing embeddings does not improve on using a fixed embedding: it results in the same convergence rate in the Gaussian case [14, 15] and in a slower convergence rate in the SRHT case [15].

In contrast to randomized preconditioning methods, classical sketching algorithms project both $A$ and $b$, and then computes $\overline{x} := \text{argmin}_x \frac{1}{2} \|SAx - Sb\|_2^2 + \frac{1}{2} \|x\|_2^2$. In the closely related work [3], the authors showed that for $m \approx d_e/\varepsilon$, the estimate $\overline{x}$ satisfies $f(\overline{x}) \leq (1 + \varepsilon)f(x^*)$. This can result in large $m$ for even medium accuracy, whereas our method yields an $\varepsilon$-approximate solution with $m \approx d_e$ under the mild requirement that the number of iterations $T$ satisfies $T \approx \log(1/\varepsilon)$.

2 Preliminaries

We start with providing deterministic convergence guarantees for the Polyak- and Gradient-IHS methods, and we relate the convergence rates to the extreme eigenvalues of the matrix $C_S$. Similar guarantees were established in [21].
and \( \| D \|_2 < 1 \), it holds that \( C_S \) is positive definite. Given two real numbers \( \Lambda > \lambda > 0 \), we define the \( S \)-measurable event \( \mathcal{E}_S := \{ \lambda \leq \gamma_d \leq \gamma_1 \leq \Lambda \} \). The proofs of the two next results are essentially based on standard analyses of gradient methods \([23]\), and they are deferred to Appendix C.1.

**Theorem 1.** Consider the step size \( \mu = 2/(\sqrt{\lambda} + \frac{1}{\lambda}) \), where \( \kappa = \lambda / \mu \). Then, conditional on \( \mathcal{E}_S \), the Gradient-IHS method satisfies at each iteration

\[
\frac{\delta_{t+1}}{\delta_t} \leq c_{gd}(\lambda, \Lambda), \quad \text{where} \quad c_{gd}(\lambda, \Lambda) := \left( \frac{\kappa - 1}{\kappa + 1} \right)^2.
\]

**Theorem 2.** Consider the step size \( \mu = 4/(\sqrt{\kappa} + \frac{1}{\sqrt{\kappa}})^2 \) and momentum parameter \( \beta = (\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)^2 \), where \( \kappa = \Lambda / \lambda \). Then, conditional on \( \mathcal{E}_S \), the Polyak-IHS satisfies

\[
\limsup_{t \to \infty} \left( \frac{\delta_t}{b_0} \right) \leq c_p(\lambda, \Lambda), \quad \text{where} \quad c_p(\lambda, \Lambda) := \left( \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^2.
\]

The above rates \( c_{gd}(\lambda, \Lambda) \) and \( c_p(\lambda, \Lambda) \) will play a critical role in the design of our adaptive method. For the Gradient-IHS method, it should be noted that we are able to monitor the improvement ratio between two consecutive iterates. However, for the Polyak-IHS method, we only obtain an asymptotic guarantee as \( t \to +\infty \). This standard result regarding the Heavy-ball method \([23]\) essentially follows from the fact that the iterates obey a non-symmetric linear dynamical system so that the spectral and operator norms of the linear dynamics coincide asymptotically, according to Gelfand’s formula (we refer to the proof in Appendix C.1 for a formal explanation).

## 3 Sharp convergence rates for Gaussian and SRHT embeddings

According to Theorems 1 and 2, we need sharp estimates of the extreme eigenvalues of \( C_S \) in order to pick optimal parameters for the Polyak- and Gradient-IHS methods, and this will be necessary to design our adaptive method.

### 3.1 The Gaussian case

We provide a concentration bound on the edge eigenvalues \( \gamma_1 \) and \( \gamma_d \) of the matrix \( C_S \) in terms of the aspect ratio \( d_e / m \). Our analysis is based on a generalized Gordon’s Gaussian comparison theorem \([9, 25]\) and it provides sharper constants than existing results. We defer the proof to Appendix C.1.

**Theorem 3.** Let \( S \in \mathbb{R}^{m \times n} \) be a Gaussian embedding and \( \rho \in (0, 1) \). If \( m \geq \frac{d_e}{\rho^2} \) then for any \( \eta \in (0, (1 - \sqrt{\eta})^2 / 4) \), it holds that \( \lambda_\rho \leq \gamma_d \leq \gamma_1 \leq \Lambda_\rho \) with probability at least \( 1 - 16e^{-\frac{n^2 \rho^2}{4}} \)

where \( c(\eta) := \left( \frac{1 + \sqrt{\eta}}{1 - \sqrt{\eta}} \right)^2 \) and

\[
\begin{aligned}
\lambda_\rho &= 1 - \| D \|^2_2 + \| D \|^2_2 \left( 1 - c(\eta) \rho \right)^2 \\
\Lambda_\rho &= 1 - \| D \|^2_2 + \| D \|^2_2 \left( 1 + \sqrt{\rho} \right)^2 \left( 1 + \sqrt{\eta} \right)^2.
\end{aligned}
\]

If we let \( d_e, m \to +\infty \) while keeping \( \rho := \frac{d_e}{m} \) fixed and take \( \eta = o(1/\sqrt{m}) \), then our lower and upper bounds converge to the respective limits \( 1 - \| D \|^2_2 + \| D \|^2_2 (1 - \sqrt{\rho})^2 \) and \( 1 - \| D \|^2_2 + \| D \|^2_2 (1 + \sqrt{\rho})^2 \). When \( D = \| D \|_2 \cdot I_d \), these limits are exact as they correspond to the edges of the support of the Marchenko-Pastur distribution \([13]\), so that our bounds are tight in a worst-case sense.

It follows from Theorem 3 that \( \| C_S - I_d \|_2 \leq \| D \|^2_2 (1 + c_1 m^{-\frac{1}{4}}) (2\sqrt{\rho} + \rho) \) with probability at least \( 1 - 16e^{-\frac{n^2 \rho^2}{4}} \) where \( c_1 > 0 \) is a universal constant. Koltchinskii and Lounici \([13]\) established a similar inequality, i.e., \( \| C_S - I_d \|_2 \leq \| D \|^2_2 (1 + c_0) (2\sqrt{\rho} + \rho) \) with high probability for some universal constant \( c_0 > 0 \). In contrast, our factor \( (1 + c_1 m^{-\frac{1}{4}}) \) is sharper for \( m \) large enough.

The operator norm \( \| D \|_2 \) might be unknown in practice, and so would be the bounds \( \lambda_\rho \) and \( \Lambda_\rho \). However, if \( \rho \) is known and \( m \geq d_e / \rho \), one can use instead the following bounds.
Given $\rho \in (0, 1)$ (and $\eta \in (0, 1)$), we define the lower and upper bounds $\bar{\lambda}_\rho := (1 - \sqrt{c(\eta)^2})^2$ and $\Lambda_\rho := (1 + \sqrt{\eta})^2$, which always satisfy $\bar{\lambda}_\rho \leq \lambda_\rho$ and $\Lambda_\rho \leq \bar{\lambda}_\rho$ since $\|D\|_2 < 1$. Further, we define the target convergence rates $\tau_{gd}(\rho) := c_{gd}(\bar{\lambda}_\rho, \Lambda_\rho)$ and $\tau_p(\rho) := c_p(\bar{\lambda}_\rho, \Lambda_\rho)$.

3.2 The SRHT case

For the SRHT, our concentration bound is given in terms of the aspect ratio $C(n, d_e) \cdot \frac{d_e \log(d_e)}{m}$ where we introduced the relevant factor $C(n, d_e) := \frac{16}{3} (1 + \sqrt{\frac{8 \log(d_e n)}{d_e}})^2$. Under the mild requirement $d_e \geq \log(n)$, this factor satisfies $C(n, d_e) = O(1)$, so that the latter ratio scales as $\frac{d_e \log(d_e)}{m}$.

Our proof generalizes the results and analysis techniques from the work of J. Tropp [26], who treated the specific case $D = I_d$, and it relies on two powerful matrix inequalities, namely, Lieb’s and the matrix Bernstein inequalities [27, 28]. We defer it to Appendix D. We note that similar concentration bounds were obtained by the authors of [6] in the context of approximate matrix multiplication, using different analysis techniques.

**Theorem 4.** Let $\rho \in (0, 1)$ and $m \geq C(n, d_e) \cdot \frac{d_e \log(d_e)}{\rho}$. Then it holds with probability at least $1 - 9/d_e$ that $\lambda_\rho \leq \gamma_d \leq \gamma_1 \leq \Lambda_\rho$ where $\lambda_\rho := 1 - \|D\|_2^2 \sqrt{\rho}$ and $\Lambda_\rho := 1 + \|D\|_2^2 \sqrt{\rho}$.

As already mentioned in the previous section, the operator norm $\|D\|_2$ might be unknown in practice, and so would be the lower and upper bounds $\lambda_\rho$ and $\Lambda_\rho$. However, if $\rho$ is known and $m \geq C(n, d_e) \cdot \frac{d_e \log(d_e)}{\rho}$, one can use instead the following bounds.

**Definition 3.2.** Given $\rho \in (0, 1)$, we define the lower and upper bounds $\bar{\lambda}_\rho := 1 - \sqrt{\rho}$ and $\bar{\lambda}_\rho := 1 + \sqrt{\rho}$, which always satisfy $\bar{\lambda}_\rho \leq \lambda_\rho$ and $\Lambda_\rho \leq \bar{\lambda}_\rho$ since $\|D\|_2 < 1$. Further, we define the target convergence rates $\tau_{gd}(\rho) := c_{gd}(\bar{\lambda}_\rho, \Lambda_\rho)$ and $\tau_p(\rho) := c_p(\bar{\lambda}_\rho, \Lambda_\rho)$.

4 An adaptive method free of the knowledge of the effective dimension

One usually does not have access to the effective dimension $d_e$, and accurate estimates of this quantity can be efficiently obtained only in some restrictive settings (e.g., $d_e$ very small [3]).

We now propose a novel adaptive method with time-varying sketch size. Our algorithm does not require the knowledge of $d_e$, but still achieves a fast rate of convergence while keeping $m \leq d_e$.

Suppose first that we are able to compute the per-iteration improvement ratio $C_t := \frac{\delta_{t+1}}{\delta_t}$. Starting from an arbitrary initial sketch size $m$ (say $m = 1$) and fixing an improvement ratio threshold $C$, our adaptive algorithm based, for instance, on the Gradient-IHS update and Gaussian embeddings would proceed at each iteration as follows. If $C_t \leq C$ then it accepts the update $x_{t+1}$. Otherwise it rejects the update, increases the sketch size by a constant factor (say $m \leftarrow 2m$) and re-computes the sketched matrix $S \cdot A$. Since only updates with sufficient improvement are accepted, this method achieves a convergence rate at least as good as the chosen threshold $C$. Furthermore there is only a finite number of steps $K$ where the updates get rejected. Indeed, according to Theorems 1 and 3 as soon as the sketch size becomes larger than $\Omega(d_e/C)$ then all the updates are accepted with high probability. Thus, the number $K$ of rejected steps satisfies $K \lesssim \log(d_e/C)/\log(2)$.

However, the main issue with the previous method is that we are not able to compute the ratio $C_t$ in practice since this requires the knowledge of $\bar{x}^*$. We alleviate this difficulty by considering a quantity closely related to $C_t$, as follows. First, we introduce the error vector $e_t := \bar{U}^\top \bar{A}(x_t - x^*)$, its re-scaled version $E_t := C_{S^{-rac{1}{2}}} e_t$ and the re-scaled error $r_t := \frac{1}{2} \|E_t\|^2$. We need the following technical result whose proof is deferred to Appendix D.1.

**Lemma 1.** For any sequence of iterates $\{x_t\}$, it holds that $\delta_t = \frac{1}{2} \|e_t\|^2$ and $r_t = \frac{1}{2} g_t^\top H_S^{-1} g_t$ where $g_t := \nabla f(x_t)$.
Hence, the re-scaled error \( r_t \) is equal to \( \frac{1}{2} g_t^T H_S^{-1} g_t \) which can be easily computed. Consequently, we can efficiently monitor the ratio \( c_t := r_{t+1}/r_t \) as opposed to \( C_t \) in order to adapt the sketch size. Provided that \( c_t \) and \( C_t \) are close enough, this would yield the desired performance. We describe our proposed method in Algorithm 1.

**Algorithm 1: Adaptive Polyak-IHS method.**

| Line | Description |
|------|-------------|
| 1    | Sample \( S \in \mathbb{R}^{m \times n} \) and compute \( S_A := S \cdot A \). |
| 2    | Compute and cache \( g_1 = A^T (Ax_1 - b) + \nu^2 x_1 \) and \( \bar{g}_1 = H_S^{-1} g_1 \). |
| 3    | for \( t = 1, 2, \ldots, T - 1 \) do |
| 4    | Compute \( x_t^+ = x_t - \mu g_t \bar{g}_t \) and \( x_t^+ = x_t - \mu \bar{g}_t + \beta_p (x_t - x_{t-1}) \). |
| 5    | Compute the gradients \( \{ g_{gd}^+ = A^T (Ax_{gd}^+ - b) + \nu^2 x_{gd}^+, g_p^+ = A^T (Ax_p^+ - b) + \nu^2 x_p^+ \} \). |
| 6    | Compute the improvement ratios \( \bar{c}_{gd} = \frac{g_{gd}^+ - g_{gd}}{g_{gd}^+} \) and \( \bar{c}_p = \frac{g_p^+ - g_p}{g_p^+} \). |
| 7    | if \( \bar{c}_p < \tau_p \) then |
| 8    | Set \( g_{t+1} = g_{t+1} = g_p^+ \) and \( \bar{g}_{t+1} = \bar{g}_p^+ \). |
| 9    | else |
| 10   | if \( \bar{c}_{gd} < \bar{c}_{gd} \) then |
| 11   | Set \( x_{t+1} = x_{gd}^+, g_{t+1} = g_{gd}^+ \) and \( \bar{g}_{t+1} = \bar{g}_{gd}^+ \). |
| 12   | else |
| 13   | Set \( m := 2m \). |
| 14   | Sample \( S \in \mathbb{R}^{m \times n} \) and compute \( S_A = S \cdot A \). |
| 15   | Set \( \bar{g}_1 := H_S^{-1} g_1 \) and return to Step 4. |
| 16   | end |
| 17   | end |
| 18   | Return \( x_T \). |

At each iteration, Algorithm 1 uses the improvement ratio between Newton decrements instead of the ratio \( C_t \). According to Theorems 1 and 2, we can monitor exactly the improvement between two successive iterates of the Gradient-IHS method as opposed to the Polyak-IHS method. Algorithm 1 computes both updates in order to benefit either from the acceleration of the latter or from the stronger convergence guarantees of the former. If both updates do not make enough progress then the sketch size is increased.

Algorithm 1 requires to specify the target convergence rates \( \tau_{gd} \) and \( \tau_p \). It holds then that our method converges at least as fast as \( \tau_{gd} \). Importantly, the time-varying sketch size \( m \) satisfies \( m = \mathcal{O}(d_\epsilon/\tau_{gd}) \) with Gaussian embeddings and \( m = \mathcal{O}(d_\epsilon \log(d_\epsilon)/\tau_{gd}) \) with the SRHT. We formalize these guarantees in Theorems 5 and 6 whose proofs are deferred to Appendices C.2 and C.3.

**Theorem 5.** Let \( \rho, \eta \in (0, 1) \) be user’s choice parameters and consider Gaussian (resp. SRHT) embeddings. Let \( \lambda_p, \Lambda_p, \tau_{gd}(\rho) \) and \( \tau_p(\rho) \) be as given in Definition 3.1 (resp. Definition 3.2), and choose the target convergence rates \( \tau_{gd} = \tau_{gd}(\rho) \) and \( \tau_p = \tau_p(\rho) \). Let \( \mu_{gd}, \mu_p \) and \( \beta_p \) be the corresponding parameters as given in Theorems 1 and 2. Fix an arbitrary initial sketch size \( m_{\text{initial}} \geq 1 \). Denote by \( K = K_t \), the number of rejected updates at iteration \( t \). Then, for Gaussian embeddings, with \( a(\rho, \eta) := \frac{1 + \sqrt{\rho}}{(1 - \sqrt{\rho})^2} \), it holds with probability at least \( 1 - 16 e^{-\eta^2 d_\epsilon/2} \) that at any iteration, \( m \leq 2 a(\rho, \eta) d_\epsilon \rho \) and \( K \leq \log \left( a(\rho, \eta) d_\epsilon \rho \right) \). Further, for

\footnote{The quantity \( r_t = g_t^T H_S^{-1} g_t \) is usually referred to as a Newton decrement in the optimization literature.}
the SRHT, with \( a(\rho) := \frac{1 + \sqrt{\pi}}{2} \epsilon \), it holds with probability at least \( 1 - \frac{9}{\epsilon^2} \) that at any iteration, \( m \leq 2 a(\rho) C(n, d_c) \cdot \frac{d \log(d_c)}{\rho} \) and \( K \leq \log \left( a(\rho) C(n, d_c) \frac{d \log(d_c)}{m_{\text{mul}} \rho} \right) / \log(2) + 1. \)

**Theorem 6.** Consider the same conditions of Theorem 5. For Gaussian embeddings, we have for any \( t \geq 1 \) that \( \frac{d}{m} \leq 16 \cdot (1 + \frac{\sigma^2}{\rho^2}) \cdot \max\{1, \frac{d_c}{m_{\text{mul}}}\} \cdot \epsilon^2 \) with probability at least \( 1 - 16e^{-n^2 d_c/2} \). For the SRHT, we have that \( \frac{d}{m} \leq \frac{1 + \frac{\sigma^2}{\rho^2}}{\epsilon} \cdot \max\{1, \frac{d_c}{m_{\text{mul}}}\} \cdot \epsilon^2 \) with probability at least \( 1 - 9/d_c \).

### 4.1 Time and space complexity

We consider here the SRHT for which computing \( SA \) is faster than Gaussian projections. We have the following complexity result, whose proof is deferred to Appendix C.4.

**Theorem 7.** Consider the same conditions of Theorem 5 with the SRHT. Let \( \epsilon \in (0, 1) \) be a given precision such that \( \epsilon \leq \min\{\frac{\sigma^2}{\rho^2}, \frac{1}{d_c}\} \). Then it holds with probability at least \( 1 - \frac{9}{\epsilon^2} \) that the number of iterations to reach a solution \( x_T \) such that \( \|A(x_T - x^*)\|^2 \leq \epsilon \) satisfies \( T = O\left( \frac{1}{\epsilon \log(1/\rho)} \right) \).

Thus the total time complexity \( C_T \) of Algorithm 1 verifies

\[
C_T = O\left( \log(d_c/\epsilon) (nd \log(d_c/\epsilon)) + \frac{d^2 \log^2(d_c)}{\rho^2} d + nd \log(1/\epsilon) \log(1/\rho) \right).
\]

The time complexity \( C_T \) is decomposed into three terms. Sketching the data matrix takes \( O(nd \log(d_c/\epsilon)) \) time. The cost \( O\left( \frac{d^2 \log^2 d}{\rho^2} \right) \) corresponds to computing a factorization of \( H_S \) using the Woodbury identity (see Appendix C.4 for details). These two costs are multiplied by an extra factor \( O(\log(d_c/\epsilon)) \) which is the number of rejected steps. The last term is the per-iteration complexity \( O(nd) \) times the number of iterations \( T = O\left( \frac{1}{\epsilon \log(1/\rho)} \right) \). In contrast, other state-of-the-art randomized preconditioning methods \([24, 2, 19]\) prescribe the sketch size \( m = \frac{d \log d}{\rho} \) and they are also decomposed into three steps: sketching, factoring, and iterating. Sketching with the SRHT also costs \( O(nd \log(d/\rho)) \) and the factoring step takes \( O\left( \frac{d^2 \log^2 d}{\rho^2} \right) \) time. The iteration part costs \( O(nd \log(1/\rho)) \). This yields the total complexity \( C_{\text{other}} = O(nd \log(d/\rho)) + \frac{d^2 \log^2 d}{\rho^2} + nd \log(1/\rho) \). Thus, even with the extra factor \( \log(d_c/\epsilon) \) due to the rejected steps, our adaptive method improves on the sketching plus factor costs especially when the effective dimension \( d_c \) is much smaller than the data dimension \( d \) and thus on the total complexity.

Regarding space complexity, our method requires \( O(d \cdot d_c \log d_c/\rho) \) space to store the sketched matrix \( SA \) whereas the other preconditioning methods needs \( O(d^2 \log d/\rho) \), and this is a significant improvement when \( d_c \) is much smaller than \( d \).

### 5 Numerical experiments

We carry out numerical simulations of Algorithm 1 and we compare it to standard least-squares solvers, that is, the CG method and the randomized preconditioned CG (pCG) \([24]\). Numerical simulations were carried out a 512Gb desktop station and implemented in Python using its standard numerical linear algebra modules (NumPy and SciPy).

We consider two evaluation criteria: (i) the cumulative time to compute the solutions up to a given precision \( \epsilon > 0 \) along an entire regularization path (several values of \( \nu \) in decreasing order) and the memory space required by each sketching-based algorithm as measured by the sketch size \( m \), and, (ii) the same criteria but for a fixed value of \( \nu > 0 \).

We present in Figures 1 and 2 results for two standard datasets (see Appendix A for additional experiments): (i) one-vs-all classification of MNIST digits and (ii) one-vs-all classification of CIFAR10 images.

Except for very large values of the regularization parameter \( \nu > 0 \) for which the regularized least-squares problem (1) is well-conditioned so that the conjugate gradient is very efficient, we observe that our method is the fastest and requires less memory space than pCG for computing both the solutions of the entire regularization path and for a fixed value of \( \nu \). In particular, pCG uses \( m = \frac{d}{\rho} \).
for Gaussian embeddings and $m = \frac{d \log d}{\rho}$ for the SRHT. Thus pCG is especially slow at the beginning because the factorization cost scales as $O(d^3)$ and it requires memory space $O(d^2)$. In contrast, our method starts with $m = 1$ and $m$ does not exceed $O(d \log d / \rho)$ for Gaussian embeddings and $O(d \log d \log d / \rho)$ for the SRHT, as predicted by Theorem 5. Our adaptive sketch size actually remains significantly smaller than these theoretical upper bounds, and we still have a fast rate of convergence.

We observe in practice that, in Algorithm 1, the Polyak-IHS update is often rejected compared to the Gradient-IHS update. We consider an entire regularization path, where we initialize each algorithm at the previous solution of Algorithm 1 which only computes Gradient-IHS updates. This variant enjoys the same convergence guarantees as presented in Theorems 5 and 6. Since it computes only a single candidate update, this variant is faster than Algorithm 1 in the case where the Polyak-IHS update is often rejected.

![Figure 1: CIFAR10 and MNIST datasets: comparison of CG, pCG, Algorithm 1 and a variant of it which does not consider the Polyak-IHS update but only the Gradient-IHS update.](image)

Figure 1: CIFAR10 and MNIST datasets: comparison of CG, pCG, Algorithm 1 and a variant of Algorithm 1 which only computes Gradient-IHS updates. We consider an entire regularization path $\nu \in \{10^j \mid j = 4, \ldots, -2\}$. For each algorithm, we start with the largest value $\nu = 10^4$. For $j \leq 3$, we initialize each algorithm at the previous solution $\tilde{x}$ found for $j + 1$. For each value of $\nu$, we stop the algorithm once $\varepsilon = 10^{-10}$-precision is met. We observe that pCG is slow at the beginning due to forming and factoring the $m \times d$ sketched matrix $S \cdot A$ with $m \approx d$. In contrast, our methods start with $m = 1$ and the varying sketch size remains much smaller than that of pCG. Each run is averaged over 30 independent trials. Mean standard deviations are reported in the form of error bars.

![Figure 2: CIFAR10 and MNIST datasets: comparison of CG, pCG, Algorithm 1 and a variant of Algorithm 1 which only computes Gradient-IHS updates.](image)

Figure 2: CIFAR10 and MNIST datasets: comparison of CG, pCG, Algorithm 1 and a variant of Algorithm 1 which only computes Gradient-IHS updates. We fix the value of the regularization parameter $\nu = 10$. We observe that pCG is slow at the beginning due to forming and factoring the $m \times d$ sketched matrix $S \cdot A$ with $m \approx d$. In contrast, our methods start with $m = 1$ and the varying sketch size remains much smaller than that of pCG. On the other hand, we observe that the (per-iteration) convergence rate of pCG is faster and we believe this is due to using a larger sketch size (which is the best statistical lower bound known for this method). Each run is averaged over 30 independent trials.
Acknowledgements

This work was partially supported by the National Science Foundation under grant IIS-1838179.

References

[1] Ailon, Nir and Chazelle, Bernard. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pp. 557–563, 2006.

[2] Avron, Haim and Maymounkov, Petar and Toledo, Sivan. Blendenpik: Supercharging LAPACK’s least-squares solver. *SIAM Journal on Scientific Computing*, vol. 32, no. 3, pp. 1217–1236, 2010.

[3] Avron, Haim and Clarkson, Kenneth and Woodruff, David. Sharper bounds for regularized data fitting. [arXiv:1611.03225] 2016.

[4] Avron, Haim and Toledo, Sivan. Randomized algorithms for estimating the trace of an implicit symmetric positive semi-definite matrix. *Journal of the ACM*, vol. 58, no. 2, pp. 1–34, 2011.

[5] Boyd, Stephen and Vandenberghe, Lieven. Convex optimization. *Cambridge University Press*, 2004.

[6] Cohen, Michael and Nelson, Jelani and Woodruff, David. Optimal approximate matrix product in terms of stable rank. [arXiv:1507.02268] 2015.

[7] Dobriban, Edgar and Liu, Sifan. Asymptotics for Sketching in Least Squares Regression. *Advances in Neural Information Processing Systems*, pp. 3670–3680, 2019.

[8] Drineas, Petros and Mahoney, Michael. RandNLA: randomized numerical linear algebra. *Communications of the ACM*, vol. 59, no. 6, pp. 80–90, 2019.

[9] Gordon, Yehoram. Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics*, vol. 50, no. 4, pp. 265–289, 1985.

[10] Grewal, Mohinder and Andrews, Angus. Kalman filtering: Theory and Practice with MATLAB. *John Wiley & Sons*, 2014.

[11] Gross, David and Nesme, Vincent. Note on sampling without replacing from a finite collection of matrices. [arXiv:1001.2738] 2010.

[12] Hastie, Trevor and Tibshirani, Robert and Friedman, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. *Springer Science & Business Media*, 2009.

[13] Koltchinskii, Vladimir and Lounici, Karim. Concentration inequalities and moment bounds for sample covariance operators. [arXiv:1405.2468] 2014.

[14] Lacotte, Jonathan and Pilanci, Mert. Faster Least Squares Optimization. [arXiv:1911.02675] 2019.

[15] Lacotte, Jonathan and Pilanci, Mert. Optimal Randomized First-Order Methods for Least-Squares Problems. *Proceedings of the Thirty-seventh International Conference on Machine Learning*, 2020.

[16] Mahoney, Michael. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, vol. 3, no. 1, pp. 123–224, 2011.

[17] Mahoney, Michael and Drineas, Petros. Structural Properties Underlying High-Quality Randomized Numerical Linear Algebra Algorithms. 2016.

[18] Marchenko, Vladimir and Pastur, Leonid. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, vol. 1, no. 4, pp. 457, 1967.

[19] Meng, Xiangrui and Saunders, Michael and Mahoney, Michael. LSRN: A parallel iterative solver for strongly over-or underdetermined system. *SIAM Journal on Scientific Computing*, vol. 36, no. 2, pp. 95–118, 2014.

[20] Ozaslan, Ibrahim and Pilanci, Mert and Arikan, Orhan. Iterative hessian sketch with momentum. *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7470–7474, 2019.
[21] Ozaslan, Ibrahim and Pilanci, Mert and Arikan, Orhan. Regularized Momentum Iterative Hessian Sketch for Large Scale Linear System of Equations. [arXiv:1912.03514], 2019.

[22] Pilanci, Mert and Wainwright, Martin. Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1842–1879, 2016.

[23] Polyak, Boris. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.

[24] Rokhlin, Vladimir and Tygert, Mark. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, vol. 105, no. 36, pp. 13212–13217, 2008.

[25] Thrampoulidis, Christos and Oymak, Samet and Hassibi, Babak. A tight version of the gaussian min-max theorem in the presence of convexity. [arXiv:1408.4837], 2014.

[26] Tropp, Joel. Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis*, vol. 3, no. 01n02, pp. 115–126, 2011.

[27] Tropp, Joel. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, vol. 8, no. 01n02, pp. 1–230, 2015.

[28] Vershynin, Roman. High-dimensional probability: An introduction with applications in data science. *Cambridge University Press*, vol. 47, 2018.

[29] Vogel, Curtis. Computational methods for inverse problems. *SIAM*, vol. 24, 2002.
A Additional numerical experiments

Here, we consider a synthetic dataset with $A$ having exponential spectral decay $\sigma_j = 0.95^j$ for $j = 1, \ldots, d$. The observation vector is generated as follows, $b = Ax_{pl} + \eta$, where $x_{pl}$ is a planted vector with $\frac{1}{\sqrt{d}}N(0, 1)$ independent entries and $\eta$ is a vector of Gaussian noise $\frac{1}{\sqrt{n}}N(0, I_n)$. We also consider the similar synthetic dataset but with polynomially decaying singular values $\sigma_j = 1/j$ for $j = 1, \ldots, d$. Results are reported in Figure 3.

![Figure 3](image_url)

**Figure 3:** Exponential and polynomial spectral decays: comparison of CG, pCG, Algorithm [1] and a variant of Algorithm [1] which only computes Gradient-IHS updates. We consider an entire regularization path $\nu \in \{10^j \mid j = 0, \ldots, -4\}$. For each algorithm, we start with the largest value $\nu = 1$. For $j \leq 3$, we initialize each algorithm at the previous solution $\tilde{x}$ found for $j + 1$. For each value of $\nu$, we stop the algorithm once $\varepsilon = 10^{-10}$-precision is met. We observe that pCG is slow at the beginning due to forming and factoring the $m \times d$ sketched matrix $S \cdot A$ with $m \approx d$. In contrast, our methods start with $m = 1$ and the varying sketch size remains much smaller than that of pCG. This leads to better time and memory space performance, except for the case of Gaussian embeddings and polynomial decays. In the latter case, our method is slowed down by Gaussian projections which are expensive. But with the SRHT, our method has the best performance. Each run is averaged over 30 independent trials. Mean standard deviations are reported in the form of error bars.

B The underdetermined case $n \leq d$

A dual of the problem (1) is

$$z^* := \arg\min_{z \in \mathbb{R}^n} \left\{ g(z) := \frac{1}{2} \| A^T z \|^2 + \frac{\nu}{2} \| z \|^2 - b^T z \right\},$$

and one can map the optimal dual solution $z^*$ to the primal one using the relationship

$$x^* = A^T z^*.$$  

The dual problem fits into the primal overdetermined framework we considered in the main body of this manuscript. Note that there exists a priori the following technical difficulty. In order to formulate (7) as a regularized least-squares problem, one needs to consider the observation vector $\hat{b} = A^T b$ (where $A^T$ is the pseudo-inverse of $A$) and then, we have

$$z^* := \arg\min_{z \in \mathbb{R}^n} \left\{ \frac{1}{2} \| A^T z - \hat{b} \|^2 + \frac{\nu}{2} \| z \|^2 \right\}.$$  

One might wonder if $\hat{b}$ needs to be computed in order to apply the previous framework to the dual overdetermined case: this is not the case. Indeed, in Algorithm [1] the observation vector $\hat{b}$ only appears in the gradient formula, as $\nabla f(x_t) = A^T (Ax_t - \hat{b})$. Equivalently, here we obtain

$$\nabla g(z_t) = A (A^T z_t - \hat{b}) = A A^T z_t - b.$$
We denote by $\mathcal{D}$ where $\sigma$ is the largest singular value of $\mathcal{A}$. That is, the gradient is easily computed and Algorithm 1 can be applied to the dual problem (7). Further, this yields guarantees on the error

$$\frac{1}{2} \| \mathcal{A}(z_t - z^*) \|^2,$$

and using the dual-primal map (8), we would get guarantees on the error $\frac{1}{2} \| x_t - x^* \|^2$. Regarding the semi-norm prediction error, we have

$$\frac{1}{2} \| \mathcal{A}(x_t - x^*) \|^2 \leq \sigma_1^2 \cdot \frac{1}{2} \| x_t - x^* \|^2 = \sigma_1^2 \cdot \frac{1}{2} \| \mathcal{A}(z_t - z^*) \|^2,$$

where $\sigma_1$ is the largest singular value of $\mathcal{A}$. Thus, even with the extra factor $\sigma_1^2$ which is negligible in our complexity analysis, we obtain the same guarantees as in Theorems 5, 6 and 7.

### C Proof of main results

#### C.1 Proof of Theorems 1 and 2

We denote by $A = USV^T$ a singular value decomposition of the matrix $\mathcal{A}$, where $U = [u_1, \ldots, u_d] \in \mathbb{R}^{n \times d}$ has orthonormal columns, $V = [v_1, \ldots, v_d] \in \mathbb{R}^{d \times d}$ has orthonormal columns, and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_d)$, with $\sigma_1 \geq \ldots \geq \sigma_d > 0$.

We denote $D = \text{diag} \left( \frac{\sigma_1}{\sqrt{\sigma_1^2 + \nu^2}}, \ldots, \frac{\sigma_d}{\sqrt{\sigma_d^2 + \nu^2}} \right)$, $D' = \text{diag} \left( \frac{\nu}{\sqrt{\sigma_1^2 + \nu^2}}, \ldots, \frac{\nu}{\sqrt{\sigma_d^2 + \nu^2}} \right)$, and further,

$$\bar{U} := \begin{bmatrix} UD \\ VD' \end{bmatrix}, \quad \bar{\Sigma} := \text{diag} \left( \sqrt{\sigma_1^2 + \nu^2}, \ldots, \sqrt{\sigma_d^2 + \nu^2} \right).$$

Note that $\bar{A} = \bar{U} \bar{\Sigma} V^T$. Indeed,

$$\bar{U} \bar{\Sigma} V^T = \begin{bmatrix} UD \bar{\Sigma} V^T \\ VD' \bar{\Sigma} V^T \end{bmatrix} = \begin{bmatrix} U \Sigma V^T \\ V(\nu \cdot I_d) V^T \end{bmatrix} = \begin{bmatrix} A \\ \nu \cdot I_d \end{bmatrix}.$$

Further, the columns of $\bar{U}$ are orthonormal, and the matrix $\bar{\Sigma}$ is diagonal with non-negative entries, so that $\bar{U} \bar{\Sigma} V^T$ is a singular value decomposition of $\bar{A}$.

Given an embedding $S \in \mathbb{R}^{m \times n}$, denote by $\bar{S}$ the $(m + d) \times (n + d)$ block-diagonal matrix $\begin{bmatrix} S & 0 \\ 0 & I_d \end{bmatrix}$.

Denote $\bar{b} = \begin{bmatrix} b \\ 0 \end{bmatrix}$. We have that $\bar{A}^T \bar{S} \bar{A} = A^T S^T \mathcal{A} + \nu^2 I_d = H_S$. Consequently, given a step size $\mu \in \mathbb{R}$ and a momentum parameter $\beta \in \mathbb{R}$, the update formula (2) of the Polyak-IHS method can be equivalently written as

$$x_{t+1} = x_t - \mu (\bar{A}^T \bar{S} \bar{A})^{-1} \bar{A}^T (\bar{A}x_t - \bar{b}) + \beta (x_t - x_{t-1}) \quad (9).$$

Multiplying the update formula (9) by $\bar{U}^T \bar{A}$, subtracting $\bar{U}^T \bar{A} x^*$, using the normal equation $\bar{A}^T \bar{b} = \bar{A}^T \bar{A} x^*$ and using the notation $\epsilon_t := \bar{U}^T \bar{A}(x_t - x^*)$, we obtain that

$$\epsilon_{t+1} = \epsilon_t - \mu \bar{U}^T \bar{A} (\bar{A}^T \bar{S} \bar{A})^{-1} \bar{A}^T \epsilon_t + \beta (\epsilon_t - \epsilon_{t-1})$$

$$= (I - \mu (\bar{U}^T \bar{S} \bar{U})^{-1}) \epsilon_t + \beta (\epsilon_t - \epsilon_{t-1}).$$

Further, unrolling the expression $\bar{U}^T \bar{S} \bar{U} = D(U^T S^T SU - I_d) + I_d = C_S$, we find the error recursion

$$\begin{bmatrix} \epsilon_{t+1} \\ \epsilon_t \end{bmatrix} = \begin{bmatrix} 1 + \beta & -\beta \\ \beta & 0 \end{bmatrix} \begin{bmatrix} C_S^{-1} & 0 \\ 0 & \epsilon_t \end{bmatrix} + \begin{bmatrix} 1 + \beta & -\beta \\ \beta & 0 \end{bmatrix} \begin{bmatrix} C_S^{-1} & 0 \\ 0 & \epsilon_t \end{bmatrix} = \begin{bmatrix} (1 + \beta) I_d - \mu C_S^{-1} & -\beta I_d \\ \beta I_d & 0 \end{bmatrix} \begin{bmatrix} \epsilon_{t+1} \\ \epsilon_t \end{bmatrix}.$$

(10)
C.1.1 Gradient-IHS method

For the Gradient-IHS method, we have that $\beta = 0$ so that the dynamics (10) simplifies to \[ e_{t+1} = (I_d - \mu C_S^{-1}) e_t . \]

Using the fact that $\delta_t = \frac{1}{2} \| e_t \|^2$, we obtain that for any $t \geq 0$,

\[ \frac{\delta_{t+1}}{\delta_t} \leq \| I_d - \mu C_S^{-1} \|^2_2 . \]

The eigenvalues of the matrix $I_d - \mu C_S^{-1}$ are given by $1 - \frac{\mu}{\gamma_i}$ where the $\gamma_i$'s are the eigenvalues of $C_S$ indexed in non-increasing order. Then,

\[ \| I_d - \mu C_S^{-1} \|_2 = \max \left\{ | 1 - \frac{\mu}{\gamma_1} |, | 1 - \frac{\mu}{\gamma_d} | \right\} . \]

If $\lambda, \Lambda > 0$ are two real numbers such that $\lambda \leq \gamma_d \leq \gamma_1 \leq \Lambda$, then it holds that for any $\mu \geq 0$,

\[ \max \left\{ | 1 - \frac{\mu}{\gamma_1} |, | 1 - \frac{\mu}{\gamma_d} | \right\} \leq \max \left\{ | 1 - \frac{\mu}{\Lambda} |, | 1 - \frac{\mu}{\lambda} | \right\} . \]

Picking $\mu = 2/(\frac{1}{\lambda} + \frac{1}{\Lambda})$ yields that

\[ \| I_d - \mu C_S^{-1} \|_2 \leq \left( \frac{\Lambda - \lambda}{\Lambda + \lambda} \right) = \frac{\kappa - 1}{\kappa + 1} , \]

where $\kappa = \Lambda/\lambda$, and this implies the claimed result.

C.1.2 Polyak-IHS method

Using (10) and the fact that $\delta_t = \frac{1}{2} \| e_t \|^2$, we immediately find by recursion that

\[ \left( \frac{\delta_{t+1} + \delta_t}{\delta_1 + \delta_0} \right)^{\frac{1}{2}} \leq \| M(\mu, \beta) \|_2^{\frac{1}{2}} . \]

From Gelfand formula, we obtain that

\[ \limsup_{t \to \infty} \left( \frac{\delta_t}{\delta_0} \right)^{\frac{1}{2}} \leq \rho^2(M(\mu, \beta)) , \]

where $\rho(M(\mu, \beta))$ is the spectral radius of the matrix $M(\mu, \beta)$ (the largest module of its eigenvalues), which remains to be computed. Let $C_S = T \Lambda T^T$ be an eigenvalue decomposition of the positive definite matrix $C_S$—where $\Lambda = \text{diag}(\gamma_1, \ldots, \gamma_d)$ and $\gamma_1 \geq \cdots \geq \gamma_d > 0$—, and define the $(2d) \times (2d)$ permutation matrix $\Pi$ as

\[ \Pi_{i,j} = \begin{cases} 1 \text{ if } i \text{ odd}, j = i \\ 1 \text{ if } i \text{ even}, j = n + i \\ 0 \text{ otherwise} \end{cases} \]

Then, it holds that

\[ \Pi \begin{bmatrix} T & 0 \\ 0 & T \end{bmatrix}^T M(\mu, \beta) \begin{bmatrix} T & 0 \\ 0 & T \end{bmatrix} \Pi^T = \begin{bmatrix} M_1(\mu, \beta) & 0 & \cdots & 0 \\ 0 & M_2(\mu, \beta) & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & M_d(\mu, \beta) \end{bmatrix} \]

where $M_i(\mu, \beta) = \begin{bmatrix} 1 + \beta - \mu \gamma_i^{-1} & -\beta \\ 1 & 0 \end{bmatrix}$. That is, $M(\mu, \beta)$ is similar to the block diagonal matrix with $2 \times 2$ diagonal blocks $M_i(\mu, \beta)$. To compute the eigenvalues of $M(\mu, \beta)$, it suffices to compute the eigenvalues of all of the $M_i(\mu, \beta)$. For fixed $i$, the eigenvalues of the $2 \times 2$ matrix are roots of the equation $u^2 - (1 + \beta - \mu/\gamma_i)u + \beta = 0$. In the case that $1 \geq \beta \geq (1 - \sqrt{\mu/\gamma_i})^2$, the roots of the characteristics equations are imaginary, and both have magnitude $\sqrt{\beta}$. Pick $\mu = \mu^* : = 4/(1/\sqrt{\lambda} + 1/\sqrt{\Lambda})^2$ and $\beta = \beta^* : = \left( \frac{\sqrt{\lambda} - \sqrt{\Lambda}}{\sqrt{\lambda} + \sqrt{\Lambda}} \right)^2$, where $\lambda, \Lambda > 0$ are respectively any lower and upper bounds of $\gamma_d$ and $\gamma_1$. Then, we have that $\beta \geq (1 - \sqrt{\mu/\gamma_i})^2$ for all $i = 1, \ldots, d$, so that $\rho(M(\mu, \beta)) \leq \sqrt{\beta}$, and this yields the claimed result.
Algorithm 1 then satisfies
\[ r \text{ Gradient-IHS update always verifies the improvement criterion} \]

Consequently, as soon as \( m > a(\rho, \eta) \frac{d_x}{\rho} \), we have \( m \leq 2 \cdot a(\rho, \eta) \frac{d_x}{\rho} \) and \( K \leq \log(a(\rho, \eta) \frac{d_x}{\rho}) / \log(2) + 1 \).

Denote \( x_t \) the current iterate, \( S \) the current sketching matrix, \( \delta_t = \frac{1}{2} \| A(x_t - x^*) \|_2^2 \) and \( r_t = \frac{1}{2} \| C^{-\frac{1}{2}} S^{-\frac{1}{2}} A(x_t - x^*) \|_2^2 \). According to Theorem 3, it holds with probability at least 1 – 16e\(-\eta^2d_x/2\) that

\[
\left(1 - \sqrt{\frac{c(\eta) \rho}{a(\rho, \eta)}}\right)^2 \leq \sigma_{\min}(C_S) \leq \sigma_{\max}(C_S) \leq \left(1 + \sqrt{\frac{\rho}{a(\rho, \eta)}}\right)^2 (1 + \sqrt{\eta})^2.
\]

In particular, it immediately follows from the above bounds that \( \frac{\sigma_{\max}(C_S)}{\sigma_{\min}(C_S)} \leq a(\rho, \eta) \).

According to Theorem 4 and using that \( \tau_{gd}(\frac{\rho}{a(\rho, \eta)}) \leq \tau_{gd}(\rho) \), the Gradient-IHS candidate \( x^*_{gd} \) of Algorithm then satisfies \( \frac{\delta^+}{\delta_t} \leq \frac{\tau_{gd}(\rho)}{a(\rho, \eta)} \) with probability at least 1 – 16e\(-\eta^2d_x/2\), where we introduced the notation \( \delta^+ = \frac{1}{2} \| A(x^*_gd - x^*) \|_2^2 \). Denoting \( r^+ = \frac{1}{2} \| C^{-\frac{1}{2}} S^{-\frac{1}{2}} A(x^*_gd - x^*) \|_2^2 \) and using \( r^+ \leq \frac{\delta^+}{\sigma_{\min}(C_S)} \) and \( r_t \geq \frac{\delta_t}{\sigma_{\min}(C_S)} \), it follows that

\[
\frac{r^+}{r_t} \leq \frac{\sigma_{\max}(C_S)}{\sigma_{\min}(C_S)} \cdot \frac{\delta^+}{\delta_t} \leq \frac{\sigma_{\max}(C_S)}{\sigma_{\min}(C_S)} \cdot \frac{\tau_{gd}(\rho)}{a(\rho, \eta)} \leq a(\rho, \eta) \cdot a(\rho, \eta)
\]

\[
= \tau_{gd}(\rho)
\]

\[
= \tau_{gd}.
\]

Consequently, as soon as \( m > a(\rho, \eta) \frac{d_x}{\rho} \), it holds with probability at least 1 – 16e\(-\eta^2d_x/2\) that the Gradient-IHS update always verifies the improvement criterion \( \frac{c(\eta) \rho}{a(\rho, \eta)} \leq \tau_{gd} \). Thus, future updates are not rejected and the sketch size remains constant.

The proof for the SRHT follows similar steps. Either the sketch size always remains smaller than \( a(\rho) \cdot C(n, d_x) \frac{d_x \log(d_x)}{\rho} \), which is equivalent to \( K \leq \log(a(\rho) \cdot C(n, d_x) \frac{d_x \log(d_x)}{\rho}) / \log(2) \). In this case, the statement of the theorem holds. Otherwise, suppose that at some iteration \( t \geq 1 \), we have \( m > a(\rho) \cdot C(n, d_x) \frac{d_x \log(d_x)}{\rho} \). Let \( t \geq 1 \) be the first such iteration so that we also have \( m \leq 2 \cdot a(\rho) \cdot C(n, d_x) \frac{d_x \log(d_x)}{\rho} \) and \( K \leq \log(a(\rho) \cdot C(n, d_x) \frac{d_x \log(d_x)}{\rho}) / \log(2) + 1 \).

Denote \( x_t \) the current iterate, \( S \) the current sketching matrix, \( \delta_t = \frac{1}{2} \| A(x_t - x^*) \|_2^2 \) and \( r_t = \frac{1}{2} \| C^{-\frac{1}{2}} S^{-\frac{1}{2}} A(x_t - x^*) \|_2^2 \). According to Theorem 4, it holds with probability at least 1 – \( \frac{a}{d_x} \) that

\[
1 - \frac{\rho}{a(\rho)} \leq \sigma_{\min}(C_S) \leq \sigma_{\max}(C_S) \leq 1 + \sqrt{\frac{\rho}{a(\rho)}}.
\]

In particular, it immediately follows from the above bounds that \( \frac{\sigma_{\max}(C_S)}{\sigma_{\min}(C_S)} \leq a(\rho) \).

According to Theorem 4 and using that \( \tau_{gd}(\rho/a(\rho)) = \tau_{gd}(\rho) \), the Gradient-IHS candidate \( x^*_gd \) of Algorithm then satisfies \( \frac{\delta^+}{\delta_t} \leq \frac{\tau_{gd}(\rho)}{a(\rho, \eta)} \) with probability at least 1 – \( \frac{a}{d_x} \), where we introduced the notation \( \delta^+ = \frac{1}{2} \| A(x^*_gd - x^*) \|_2^2 \). Denoting \( r^+ = \frac{1}{2} \| C^{-\frac{1}{2}} S^{-\frac{1}{2}} A(x^*_gd - x^*) \|_2^2 \) and using \( r^+ \leq \frac{\delta^+}{\sigma_{\min}(C_S)} \) and \( r_t \geq \frac{\delta_t}{\sigma_{\min}(C_S)} \), it follows that

\[
\frac{r^+}{r_t} \leq \frac{\sigma_{\max}(C_S)}{\sigma_{\min}(C_S)} \cdot \frac{\delta^+}{\delta_t} \leq \frac{\sigma_{\max}(C_S)}{\sigma_{\min}(C_S)} \cdot \frac{\tau_{gd}(\rho)}{a(\rho)} \leq a(\rho) \cdot \frac{\tau_{gd}(\rho)}{a(\rho)}
\]

\[
= \tau_{gd}(\rho)
\]

\[
= \tau_{gd}.
\]
Consequently, as soon as \( m > \alpha(\rho) C(n, d_c) \frac{d \log(d_e)}{\rho} \), it holds with probability at least \( 1 - \frac{\rho}{d_c} \) that the Gradient-IHS update always verifies the improvement criterion \( \frac{r_t}{r_0} \leq \tau_{gd} \). Thus, future updates are not rejected and the sketch size remains constant. \( \square \)

C.3 Proof of Theorem 6

At every iteration, we observe at least the improvement

\[
\frac{r_t}{r_0} \leq \max\{c_{gd}^t, \sigma_t \} = \sigma_t^{gd}.
\]

It holds that \( \delta_t \leq \sigma_{\max}(C_S) \cdot r_t \) and \( \delta_0 \geq \sigma_{\min}(C_{S_{\text{initial}}}) \cdot r_0 \) so that

\[
\frac{\delta_t}{\delta_0} \leq \frac{\sigma_{\max}(C_S)}{\sigma_{\min}(C_{S_{\text{initial}}})}, \quad r_t \leq \frac{\sigma_{\max}(C_S)}{\sigma_{\min}(C_{S_{\text{initial}}})} \cdot \sigma_t^{gd}.
\]

Hence it suffices to control the ratio \( \frac{\sigma_{\max}(C_S)}{\sigma_{\min}(C_{S_{\text{initial}}})} \).

For a Gaussian embedding, according to Lemma 2, we have that for any sketch size \( m \geq m_{\text{initial}} \),

\[
\sigma_{\max}(C_S) \leq \frac{\nu^2}{\sigma_1^2 + \nu^2} + \frac{\sigma_1^2}{\sigma_1^2 + \nu^2} \cdot \left( 1 + \sqrt{\frac{d_e}{m}} \right)^2 \cdot (1 + \sqrt{\eta})^2.
\]

with probability at least \( 1 - 8 \cdot e^{-\eta^2 d_e / 2} \). We have \( (1 + \sqrt{d_e / m})^2 \leq 4 \max\{1, \frac{d_e}{m_{\text{initial}}} \} \). Using \( (1 + \sqrt{\eta})^2 \leq 4 \), this further yields

\[
\sigma_{\max}(C_S) \leq \frac{\nu^2}{\sigma_1^2 + \nu^2} + 16 \frac{\sigma_1^2}{\sigma_1^2 + \nu^2} \max\{1, \frac{d_e}{m_{\text{initial}}} \}.
\]

For a SRHT embedding, according to Lemma 6, we have that for any sketch size \( m_{\text{initial}} \leq m \leq d_c \),

\[
\sigma_{\max}(C_S) \leq \frac{\nu^2}{\sigma_1^2 + \nu^2} + \frac{\sigma_1^2}{\sigma_1^2 + \nu^2} \cdot \max\{1, \frac{d_e \log(d_e) C(n, d_c)}{m_{\text{initial}}} \},
\]

with probability at least \( 1 - 9 / d_e \). On the other hand, it holds that for any embedding \( S \),

\[
\sigma_{\min}(C_S) \geq 1 - \|D\|_2^2 = \frac{\nu^2}{\sigma_1^2 + \nu^2}.
\]

Thus, for Gaussian embeddings, we obtain

\[
\frac{\sigma_{\max}(C_S)}{\sigma_{\min}(C_{S_{\text{initial}}})} \leq 16 \left( 1 + \frac{\sigma_1^2}{\nu^2} \cdot \max\{1, \frac{d_e}{m_{\text{initial}}} \} \right) \leq 16 \cdot \left( 1 + \frac{\sigma_1^2}{\nu^2} \right) \cdot \max\{1, \frac{d_e}{m_{\text{initial}}} \},
\]

and for SRHT embeddings,

\[
\frac{\sigma_{\max}(C_S)}{\sigma_{\min}(C_{S_{\text{initial}}})} \leq 1 + \frac{\sigma_1^2}{\nu^2} \cdot \max\{1, \frac{d_e \log(d_e) C(n, d_c)}{m_{\text{initial}}} \}
\]

\[
\leq (1 + \frac{\sigma_1^2}{\nu^2}) \cdot \max\{1, \frac{d_e \log(d_e) C(n, d_c)}{m_{\text{initial}}} \}.
\]

C.4 Proof of Theorem 7

According to Theorem 6 we have for any \( t \geq 1 \)

\[
\frac{\delta_t}{\delta_0} \leq (1 + \frac{\sigma_1^2}{\nu^2}) \cdot d_e \log(d_e) C(n, d_c) \cdot \tau_{gd}.
\]

It is immediate to verify that \( \tau_{gd} = \rho \) for the SRHT. Therefore, the total number of iterations \( T \) to reach an \( \varepsilon \)-accurate solution is given by

\[
T = \mathcal{O} \left( \frac{\log(1 + \frac{\sigma_1^2}{\nu^2}) + \log d_e + \log(1/\varepsilon)}{\log(1/\rho)} \right).
\]
For $\varepsilon \leq \min\{\frac{\nu^2}{\sigma^2 n + \nu^2}, \frac{1}{\nu}\}$, this reduces to

$$T = \mathcal{O}\left(\frac{\log(1/\varepsilon)}{\log(1/\rho)}\right).$$

According to Theorem 5, the number of rejected steps satisfies

$$K = \mathcal{O}(\log(d_c/\rho)).$$

At the beginning of the algorithm and for each rejected step, one needs to form the sketched matrix $SA$ which costs $\mathcal{O}(nd \log d_c)$, and, compute a factorization of $H_S$ as follows. According to the Woodbury matrix identity, the inverse of $H_S$ verifies

$$H_S^{-1} = ((SA)^\top SA + \nu^2 I_d)^{-1} = \frac{1}{\nu^2} (I_d - (SA)^\top (\nu^2 I_m + SA(SA)^\top)^{-1} SA).$$

To reduce the complexity of solving at each iteration the linear system $H_S \cdot z = \nabla f(x_t)$, one can simply cache the matrix $(\nu^2 I_m + SA(SA)^\top)^{-1}$ which takes time $\mathcal{O}(\frac{d^2 \log^2 d}{\rho})$ to compute. Since $K = \mathcal{O}(\log(d_c/\rho))$, it follows that the total sketching and factor costs scale as $\mathcal{O}(\log(d_c/\rho) \cdot \frac{d^2 \log^2 d}{\rho} + nd \log(d_c/\rho))$.

The per-iteration cost is that of computing the matrix-vector products $Ax_t$ and $A^\top (Ax_t - b)$, which is given by $\mathcal{O}(nd)$. Note that the other main numerical operation consists in solving the linear system $H_S \cdot z = \nabla f(x_t)$ using the cached matrix $(\nu^2 I_m + SA(SA)^\top)^{-1}$ and the identity $H_S^{-1} = \frac{1}{\nu^2} (I_d - (SA)^\top (\nu^2 I_m + SA(SA)^\top)^{-1} SA)$, the term $H_S^{-1} \nabla f(x_t)$ can be computed in time $\mathcal{O}(\frac{d \log^2 d}{\rho})$, which is negligible compared to $\mathcal{O}(nd)$.

Thus, we obtain the total time complexity

$$C_{\varepsilon} = \mathcal{O}\left(\log(d_c/\rho) \cdot \frac{d^2 \log^2 d}{\rho^2} d + nd \log(d_c/\rho) + nd \frac{\log(1/\varepsilon)}{\log(1/\rho)}\right),$$

which is the claimed result.

## D Proofs of auxiliary results

### D.1 Proof of Lemma 1

Let $\{x_t\}$ be a sequence of iterates. Let $\bar{U} \Sigma \bar{V}^\top$ be a singular value decomposition of $\bar{A}$. Denote $S = \begin{bmatrix} S & 0 \\ 0 & I_d \end{bmatrix}$, so that $H_S = (SA)^\top SA + \nu^2 I_d = (S A)^\top (S A)$.

By definition, we have $\delta_t = \frac{1}{2} \|\bar{A}(x_t - x^*)\|^2$. Since $\bar{A}(x_t - x^*)$ belongs to the range of $\bar{U}$, it follows that $\|\bar{A}(x_t - x^*)\| = \|\bar{U}^\top \bar{A}(x_t - x^*)\|$. Thus, $\delta_t = \frac{1}{2} \|e_t\|^2$.

On the other hand, we have that $g_t = \bar{A}^\top (\bar{A} x_t - \bar{b}) = \bar{A}^\top \bar{A}(x_t - x^*)$ and thus,

$$g_t^\top H_S^{-1} g_t = \langle \bar{A}^\top \bar{A}(x_t - x^*), (\bar{A}^\top \bar{S}^\top \bar{S} \bar{A})^{-1} \bar{A}^\top \bar{A}(x_t - x^*) \rangle$$

$$= \langle \bar{A}(x_t - x^*), (\bar{A}^\top \bar{S}^\top \bar{S} \bar{A})^{-1} \bar{A}^\top \bar{A}(x_t - x^*) \rangle$$

$$= \langle \bar{A}(x_t - x^*), \bar{U} \Sigma \bar{V}^\top (\Sigma \Sigma \bar{V}^\top \bar{S}^\top \bar{S} \Sigma \Sigma \bar{V}^\top)^{-1} \bar{V} \Sigma \bar{U}^\top \bar{A}(x_t - x^*) \rangle$$

$$= \langle \bar{A}(x_t - x^*), (\bar{U}^\top \bar{S}^\top \bar{S} \bar{U})^{-1} \bar{U}^\top \bar{A}(x_t - x^*) \rangle$$

$$= \langle \bar{U}^\top \bar{A}(x_t - x^*), (\bar{U}^\top \bar{S}^\top \bar{S} \bar{U})^{-1} \bar{U}^\top \bar{A}(x_t - x^*) \rangle.$$


E Proofs of concentration inequalities

E.1 Gaussian concentration over ellipsoids – Proof of Theorem \[2\]

Fix a sketch size \( m \geq 1 \), define \( \rho = \frac{d}{m} \), and let \( S \in \mathbb{R}^{m \times n} \) be a Gaussian embedding, i.e., a random matrix with i.i.d. entries \( \mathcal{N}(0, 1/m) \). Let \( \gamma_1 \) and \( \gamma_d \) be respectively the largest and smallest eigenvalue of the matrix \( C_S = D(U^T S^T SU - I_d)D + I \), i.e.,

\[
\gamma_1 = \sup_{\|x\| = 1} 1 + \langle x, D(U^T S^T SU - I_d)Dx \rangle
\]

\[
\gamma_d = \inf_{\|x\| = 1} 1 + \langle x, D(U^T S^T SU - I_d)Dx \rangle.
\]

We show separately the following two results which, combined together, yields Theorem \[2\]

**Lemma 2.** For any \( \eta \in (0, 1) \), it holds with probability at least \( 1 - 8 \cdot e^{-m\rho^2/2} \) that

\[
\gamma_1 \leq 1 - \|D\|_2^2 + \|D\|_2^2 (1 + \sqrt{\rho})^2 (1 + \sqrt{\eta})^2.
\]

**Lemma 3.** Suppose that \( m > d_e \). Then, for any \( \eta \in (0, (1 - \sqrt{\rho})^2/4) \), it holds with probability at least \( 1 - 8e^{-m\rho^2/2} \) that

\[
\gamma_d \geq 1 - \|D\|_2^2 + \|D\|_2^2 \left(1 - \sqrt{c(\eta)\rho}\right)^2,
\]

where \( c(\eta) := \left(\frac{1 + \sqrt{\eta}}{1 - \sqrt{\eta}}\right)^2 \).

We prove Lemma \[2\] in Appendix E.1.1 and Lemma \[3\] in Appendix E.1.2

E.1.1 Controlling the largest eigenvalue \( \gamma_1 \) (proof of Lemma \[2\])

Suppose that \( m \geq 1 \), and denote, as previously, \( \rho := \frac{d}{m} \). Define the matrix \( \bar{D} = \frac{D}{\|D\|_2} \). Note that \( \|\bar{D}\|_2^2 = d_e \) and \( \|\bar{D}\|_2 = 1 \). We aim to study the upper tail of the largest eigenvalue \( \gamma_1 \) of the matrix \( C_S \), which satisfies the variational equation

\[
\frac{\gamma_1 - 1}{\|D\|_2^2} = \sup_{\|x\| = 1} \langle x, D(\frac{1}{m}G^TG - I)Dx \rangle = \sup_{\|x\| = 1} \frac{1}{m} ||G\bar{D}x||^2 - ||\bar{D}x||^2
\]

\[
= \frac{2}{m} \sup_{z \in \mathbb{C}} \sup_{u \in \mathbb{R}^m} u^T G z + \psi(u, z),
\]

where we introduced the random matrix \( G \in \mathbb{R}^{m \times d} \) with i.i.d. Gaussian entries \( \mathcal{N}(0, 1) \) and the first equality holds since \( SU = \frac{1}{\sqrt{m}} G \). We also use the notations \( C = \{\bar{D}x \mid \|x\| = 1\} \) and \( \psi(u, z) := -\frac{1}{2} (\|u\|^2 + m\|z\|^2) \). We introduce the auxiliary random variable

\[
Y := \frac{2}{m} \sup_{z \in \mathbb{C}} \sup_{u \in \mathbb{R}^m} ||g^T u + ||u||h^T z + \psi(u, z),
\]

where \( g \in \mathbb{R}^m \) and \( h \in \mathbb{R}^d \) are random vectors with i.i.d. entries \( \mathcal{N}(0, 1) \). Using Theorem \[3\] (see Appendix E.1.3), it holds that for any \( c \in \mathbb{R} \),

\[
P\left(\frac{\gamma_1 - 1}{\|D\|_2^2} \geq c\right) \leq 2P(Y \geq c).
\]

(12)

Consequently, it suffices to control the upper tail of \( Y \) in order to control that of \( \gamma_1 \). First, we recall a few basic facts on the concentration of Gaussian random vectors (see, for instance, Theorems 3.1.1 and 6.3.2 in [28]). That is, for any \( \eta \in (0, 1) \), the following event holds with probability at least \( 1 - 4e^{-m\rho^2/2} \),

\[
\mathcal{E}_\eta := \left\{ m \frac{||g|| - \sqrt{m}}{\sqrt{m\eta}}, \quad m \frac{||g||^2 - m}{m\eta}, \quad m \frac{||\bar{D}h|| - \sqrt{d_e}}{\sqrt{d_e\eta}} \right\}.
\]

(13)
On the event $E_\eta$, we have

$$Y = \frac{1}{2} \sup_{m} \sup_{z \in C} \sup_{u \in \mathbb{R}^n} \|z\|g^\top u + \|u\|h^\top z - \frac{1}{2} \|u\|^2 - \frac{m}{2} \|z\|^2$$

\[ \text{(i)} \quad \frac{1}{2} \sup_{m} \sup_{z \in C} \sup_{t \geq 0} t \|z\| \|g\| + t \ h^\top z - \frac{1}{2} t^2 - \frac{m}{2} \|z\|^2 \]

\[ \text{(ii)} \quad \frac{1}{2} \sup_{m} \sup_{z \in C} \sup_{t \in \mathbb{R}} t (\|z\| \|g\| + |h^\top z|) - \frac{1}{2} t^2 - \frac{m}{2} \|z\|^2 \]

\[ \text{(iii)} \quad \frac{1}{2} \sup_{m} \sup_{z \in C} \|z\|^2 \|g\|^2 - m |+ \frac{1}{2} \|h^\top z\|^2 + \|z\| \|g\| \|h^\top z\| \]

\[ \leq \frac{1}{2} \sup_{m} \sup_{z \in C} \|z\|^2 \|g\|^2 - m |+ \frac{1}{2} \|h^\top z\|^2 + \|g\| \|h^\top z\| \]

\[ \text{(iv)} \quad \frac{1}{2} \sup_{m} \sup_{z \in C} \|z\|^2 \|g\|^2 - m |+ \frac{1}{2} \|h^\top z\|^2 + \|g\| \|h^\top z\| \]

\[ \text{(vii)} \quad \|g\|^2 - m \| + \frac{2}{m} \sup_{\|x\| = 1} \left\{ \frac{1}{2} \langle Dh, x \rangle^2 + \|g\| \langle Dh, x \rangle \right\} \]

\[ \leq \eta + \frac{2}{m} \sup_{0 \leq t \leq \|Dh\|} \left\{ \frac{t^2}{2} + \|g\| t \right\} \]

\[ \text{(viii)} \quad \eta + \left( \frac{d_e}{m} + 2 \sqrt{\frac{d_e}{m}} \right) (1 + \sqrt{\eta})^2 \]

\[ \text{(ix)} \quad \eta + (1 + \sqrt{\rho})^2 - 1 (1 + \sqrt{\eta})^2 \]

\[ \leq -1 + (1 + \sqrt{\rho})^2 (1 + \sqrt{\eta})^2 \]

In equality (i), we used the fact that for a vector $u$ with fixed norm $\|u\| = t$, the maximum of $g^\top u$ is equal to $\|g\| t$. In inequality (ii), we bounded $h^\top z$ by $|h^\top z|$ and then relaxed the constraint $t \geq 0$ to $t \in \mathbb{R}$. In inequality (iii), we plugged-in the value of the maximizer $t^* = \|z\| \|g\| + |h^\top z|$. In inequality (iv), we used the fact that for $z \in C$, $\|z\| \leq 1$. In (v), we used the change of variable $z = Dx$ with $\|x\| = 1$. In (vi), we used the fact that for $\|x\| = 1$, $\langle Dh, x \rangle \leq \|Dh\|$ and the equality is attained for some $x$, and then, we used the change of variable $t = \|Dh, x\|$. In (vii), we used the fact that on $E_\eta$, $\|g\|^2 - m \| \leq \eta$, $\|Dh\| \leq \sqrt{\eta} (1 + \sqrt{\eta})$ and $\|g\| \leq \sqrt{\eta} (1 + \sqrt{\eta})$. In (viii), we plugged into the objective the value of the maximizer $t^* = \sqrt{\eta} (1 + \sqrt{\eta})$. Equality (ix) follows from the definition of $\rho = d_e / m$. Inequality (x) follows from the fact that for any $\eta > 0$, we have $-1 \geq \eta - (1 + \sqrt{\eta})^2$.

Thus, we have that

$$\mathbb{P} \left[ \frac{\gamma - 1}{\|\bar{D}\|^2} \geq -1 + (1 + \sqrt{\rho})^2 (1 + \sqrt{\eta})^2 \right] \leq 2 \mathbb{P} \left[ Y \geq -1 + (1 + \sqrt{\rho})^2 (1 + \sqrt{\eta})^2 \right] \leq 2(1 - \mathbb{P}[E_\eta]) \leq 8 \cdot e^{-m \rho n^2 / 2},$$

which is the claimed inequality (11).

### E.1.2 Controlling the smallest eigenvalue $\gamma_d$ (proof of Lemma 3)

Suppose that $m > d_e$, and denote, as previously, $\rho := \frac{d_e}{m}$ and $\bar{D} := \frac{D}{\|D\|^2}$. Note that $\|\bar{D}\|^2 = d_e$ and $\|\bar{D}\| = 1$. We aim to study the lower tail of the smallest eigenvalue $\gamma_d$ of the matrix $C_S$, which satisfies the variational equation

$$\gamma_d - \frac{1}{\|\bar{D}\|^2} \inf_{\|x\| = 1} \langle x, D \left( \frac{1}{m} G^\top G - I \right) Dx \rangle = \inf_{\|x\| = 1} \frac{1}{m} \|GDx\|^2 - \|Dx\|^2 \leq \frac{2}{m} \sup_{z \in C} \sup_{u \in \mathbb{R}^m} u^\top Gz + \psi(u, z),$$

18
where we introduced the random matrix \( G \in \mathbb{R}^{m \times d} \) with i.i.d. Gaussian entries \( \mathcal{N}(0, 1) \) and the first equality holds since \( SU \overset{d}{=} \frac{1}{\sqrt{m}} G \). We also use the notations \( \mathcal{C} = \{ D x \mid \| x \| = 1 \} \) and 
\[
\psi(u, z) := -\frac{1}{2}(\|u\|^2 + m\|z\|^2).
\]
We introduce the auxiliary random variable
\[
Y := \frac{2}{m} \inf_{z \in \mathcal{C}} \sup_{u \in \mathbb{R}^m} \| z \| g^\top u + \| u \| h^\top z + \psi(u, z),
\]
where \( g \in \mathbb{R}^m \) and \( h \in \mathbb{R}^d \) are random vectors with i.i.d. entries \( \mathcal{N}(0, 1) \). Using Theorem II.1 from [28], it holds that for any \( c \in \mathbb{R} \),
\[
\mathbb{P}(\gamma_d - \frac{1}{\|D\|^2} < c) \leq 2 \mathbb{P}(Y < c).
\]
(14)
Consequently, it suffices to control the lower tail of \( Y \) in order to control that of \( \gamma_d \). First, we recall a few basic facts on the concentration of Gaussian random vectors (see, for instance, Theorems 3.1.1 and 6.3.2 in [28]). That is, for any \( \eta \in (0, 1) \), the following event holds with probability at least \( 1 - 4e^{-m\eta^2/2} \),
\[
\mathcal{E}_\eta := \left\{ \left\| g \right\| - \sqrt{m} \eta, \left\| g \right\|^2 - m \eta, \left\| \bar{D} h \right\| - \sqrt{d_e} \eta \leq \sqrt{d_e} \eta \right\}.
\]
We have
\[
Y = \frac{2}{m} \inf_{z \in \mathcal{C}} \sup_{u \in \mathbb{R}^m} \| z \| g^\top u + \| u \| h^\top z - \frac{1}{2}\| u \|^2 - \frac{m}{2}\| z \|^2
\]
\[
= \frac{2}{m} \inf_{z \in \mathcal{C}} \sup_{t \geq 0} t \| z \| \| g \| + t h^\top z - \frac{1}{2} t^2 - \frac{m}{2}\| z \|^2
\]
\[
= \inf_{z \in \mathcal{C}} \left( -\| z \|^2, \text{ if } \| z \| \| g \| + h^\top z \leq 0 \right)
\]
\[
\| z \|^2 \left( \| g \|^2 - m \right) + \frac{(h^\top z)^2}{m} + \frac{2}{m}\| z \| \| g \| \| h^\top z \|,
\]
otherwise.
Define
\[
Y_1 := \inf_{z \in \mathcal{C}, \| z \|^2 + h^\top z \leq 0} -\| z \|^2,
\]
\[
Y_2 := \inf_{z \in \mathcal{C}, \| z \|^2 + h^\top z \geq 0} \| z \|^2 \left( \| g \|^2 - m \right) + \frac{(h^\top z)^2}{m} + \frac{2}{m}\| z \| \| g \| \| h^\top z \|,
\]
so that \( Y = \min\{Y_1, Y_2\} \). For any \( z \in \mathcal{C} \), it holds that \( h^\top z \geq -\|Dh\| \), and consequently
\[
Y_1 \geq \inf_{z \in \mathcal{C}, \| z \|^2 + \|Dh\|^2 \leq \| g \|^2} -\| z \|^2 \geq -\frac{\|Dh\|^2}{\| g \|^2}.
\]
Hence, for \( \eta \in (0, 1) \), conditional on \( \mathcal{E}_\eta \), we have
\[
Y_1 \geq -\frac{d_e}{m} \left( \frac{1 + \sqrt{\eta}}{1 - \sqrt{\eta}} \right)^2 \geq \left( -2\sqrt{\frac{d_e}{m}} + \frac{d_e}{m} \right) \left( \frac{1 + \sqrt{\eta}}{1 - \sqrt{\eta}} \right)^2 = ((1 - \sqrt{\eta})^2 - 1) \left( \frac{1 + \sqrt{\eta}}{1 - \sqrt{\eta}} \right)^2,
\]
where the second inequality follows from the fact that \( m \geq d_e \) and thus, \( \sqrt{\frac{d_e}{m}} \geq \frac{d_e}{m} \). On the other hand, we have
\[
Y_2 \geq -\frac{1}{m}\| g \|^2 - m - \inf_{z \in \mathcal{C}} \left\{ \frac{(h^\top z)^2}{m} - \frac{2}{m}\| g \| \| h^\top z \| \right\}
\]
\[
\geq -\frac{1}{m}\| g \|^2 - m + \inf_{\| x \| = 1} \left\{ \frac{\| Dh, x \|^2}{m} - \frac{2}{m}\| g \| \| Dh, x \| \right\}
\]
\[
= -\frac{1}{m}\| g \|^2 - m + \frac{2}{m} \inf_{0 \leq t \leq \| Dh \|} \left\{ t^2 - \| g \| t \right\},
\]
where, in the first inequality, we relaxed the constraint set by removing the constraint $\|z\|g + h^Tz \geq 0$ and we used the fact that $\|z\| \leq 1$. In the second inequality, we used the change of variable $z = Dh$ with $\|x\| = 1$. In the third inequality, we used the fact that $\|Dh, x\| \leq \|Dh\|$ and used the change of variable $\langle Dh, x\rangle = t$ with $t \in [0, \|Dh\|]$. On the event $E_\eta$, it follows that

$$Y_2 \geq -\eta + \frac{2}{m} \inf_{0 \leq t \leq \sqrt{\bar{m}(1 + \sqrt{\eta})}} t^2 - \sqrt{\bar{m}}(1 + \sqrt{\eta})t$$

$$= -\eta + (1 + \sqrt{\eta})^2 \left( \frac{d_c}{m} - 2 \sqrt{\frac{d_c}{m}} \right)$$

$$= -\eta + (1 + \sqrt{\eta})^2((1 - \sqrt{\rho})^2 - 1).$$

For $\eta < c(\eta) := \left( \frac{1 - \sqrt{\eta}}{1 + \sqrt{\rho}} \right)^2$, further algebra yields that, on the event $E_\eta$,

$$Y \geq -1 + \left( 1 - \sqrt{c(\eta)\rho} \right)^2.$$

Hence, combining all of the above, we find that, for any $\eta \in (0, 1/2)$, with probability at least $1 - 8e^{-m\rho\eta^2}/2$,

$$\gamma_d \geq 1 - \|D\|^2_2 + \|D\|^2_2 \left( 1 - \sqrt{c(\eta)\rho} \right)^2.$$

### E.1.3 A new Gaussian comparison inequality

We start with the following well-known comparison inequality, which was first derived in [9].

**Theorem 8** (Gordon’s Gaussian comparison theorem). Let $I, J \subset \mathbb{N}^*$, and $\{X_{ij}\}, \{Y_{ij}\}$ be two centered Gaussian processes indexed on $I \times J$, such that for any $i, l \in I$ with $i \neq l$ and $j, k \in J$,

$$\begin{align*}
\mathbb{E}X_{ij}^2 &= \mathbb{E}Y_{ij}^2 \\
\mathbb{E}X_{ij}X_{lk} &\geq \mathbb{E}Y_{ij}Y_{lk} \\
\mathbb{E}X_{ij}X_{lk} &\leq \mathbb{E}Y_{ij}Y_{lk}.
\end{align*}$$

Then, for any $\{\lambda_{ij}\} \in \mathbb{R}^{t \times J}$, we have

$$\mathbb{P}\left( \bigcap_{i=1}^l \bigcup_{j=1}^J [Y_{ij} \geq \lambda_{ij}] \right) \geq \mathbb{P}\left( \bigcap_{i=1}^l \bigcup_{j=1}^J [X_{ij} \geq \lambda_{ij}] \right).$$

Our next result is a consequence of Gordon’s comparison inequality, and appears to be new. More specifically, it can be seen as a variant of the Sudakov-Fernique’s inequality (see, for instance, Theorem 7.2.11 in [28]).

**Theorem 9.** Let $S_1 \subset \mathbb{R}^n$ and $S_2 \subset \mathbb{R}^m$ be non-empty sets, and $\psi : S_1 \times S_2 \to \mathbb{R}$ be a continuous function. Then, for any $c \in \mathbb{R}$,

$$\mathbb{P}\left( \sup_{(x,y) \in S_1 \times S_2} y^T Gx + \psi(x, y) \geq c \right) \leq 2 \mathbb{P}\left( \sup_{(x,y) \in S_1 \times S_2} \|x\|^T g + \|y\|^T h^T x + \psi(x, y) \geq c \right).$$

**Proof.** The proof relies on several intermediate results, and is deferred to Section E.1.4.

**Lemma 4.** Let $G \in \mathbb{R}^{m \times n}$, $Z \in \mathbb{R}$, $g \in \mathbb{R}^m$ and $h \in \mathbb{R}^n$ have independent standard Gaussian entries. Let $I_1 \subset \mathbb{R}^n$ and $I_2 \subset \mathbb{R}^m$ be finite sets, and $\psi$ be a function defined over $I_1 \times I_2$. Then, for any $c \in \mathbb{R}$, we have

$$\mathbb{P}\left( \max_{(x,y) \in I_1 \times I_2} y^T Gx + Z\|x\|^T y + \psi(x, y) \geq c \right) \leq \mathbb{P}\left( \max_{(x,y) \in I_1 \times I_2} \|x\|^T g + \|y\|^T h^T x + \psi(x, y) \geq c \right).$$

20
We define $f$ for any sets $x,y$. By density arguments and monotone convergence, it also follows for any sets $x,y$. Thus, using the independence of $f_1$ and $Z$, we get

$$
\frac{1}{2} \mathbb{P} \left( \sup_{(x,y) \in S_1 \times S_2} f_1(x,y) \geq c, \ Z > 0 \right) \leq \mathbb{P} \left( \sup_{(x,y) \in S_1 \times S_2} f_1(x,y) \geq c \right).
$$

Consequently, using the independence of $f_1$ and $Z$, we get

$$
\mathbb{P} \left( \sup_{(x,y) \in S_1 \times S_2} f_1(x,y) \geq c, \ Z > 0 \right) \leq \mathbb{P} \left( \sup_{(x,y) \in S_1 \times S_2} f_1(x,y) \geq c \right).
$$

which yields the claim.
E.2 SRHT matrices – matrix deviation inequalities over ellipsoids

Let \( S \in \mathbb{R}^{m \times n} \) be a SRHT matrix, that is, \( S = RH\text{diag}(e) \) where \( R \) is a row-subsampling matrix of size \( m \times n \), \( H \) is the normalized Walsh-Hadamard transform of size \( n \times n \) and \( e \) is a vector of \( n \) independent Rademacher variables. We introduce the scaled diagonal matrix \( \bar{D} = \frac{D}{\|D\|_2} \). Note that \( \|\bar{D}\|_2^2 = d_e \) and \( \|\bar{D}\|_2 = 1 \).

**Lemma 5.** Let \( e_j \) be the \( j \)-th vector of the canonical basis in \( \mathbb{R}^n \). Then,

\[
\mathbb{P} \left\{ \max_{j=1,\ldots,n} \|e_j^T H\text{diag}(e)U\bar{D}\| \geq \sqrt{\frac{d_e}{n} + \frac{8\log(\beta n)}{n}} \right\} \leq \frac{1}{\beta^2}.
\]  

**Proof.** We fix a row index \( j \in \{1,\ldots,n\} \), and define the function

\[
f(x) := \|e_j^T H\text{diag}(x)U\bar{D}\| = \|x^T EU\bar{D}\|,
\]

where \( E := \text{diag}(e_j^T H) \). Each entry of \( E \) has magnitude \( n^{-\frac{1}{2}} \). The function \( f \) is convex, and its Lipschitz constant is upper bounded as follows,

\[
|f(x) - f(y)| \leq \|(x - y)^T EV\bar{D}\| \leq \|x - y\| \|E\|_2 \|V\|_2 \|\bar{D}\|_2 = \frac{1}{\sqrt{n}}\|x - y\|.
\]

For a Rademacher vector \( \varepsilon \), we have

\[
\mathbb{E} f(\varepsilon) \leq \sqrt{\mathbb{E} f(\varepsilon)^2} = \|EU\bar{D}\|_F \leq \|E\|_2 \|\bar{D}\|_2 = \sqrt{\frac{d_e}{n}}.
\]

Applying Lipschitz concentration results for Rademacher variables, we obtain

\[
\mathbb{P} \left\{ \|e_j^T H\text{diag}(\varepsilon)U\bar{D}\| \geq \sqrt{\frac{d_e}{n} + \frac{8\log(\beta n)}{n}} \right\} \leq \frac{1}{n\beta^2}.
\]

Finally, taking a union bound over \( j \in \{1,\ldots,n\} \), we obtain the claimed result. \( \square \)

**Theorem 10 (Matrix Bernstein).** Let \( \mathcal{X} = \{X_1,\ldots,X_n\} \) be a finite set of squared matrices with dimension \( d \). Fix a dimension \( m \), and suppose that there exists a positive semi-definite matrix \( V \) and a real number \( K > 0 \) such that \( \mathbb{E}[X_1] = 0 \), \( \mathbb{E}[X_1^2] \leq V \), and \( \|X_i\|_2 \leq K \) almost surely, where \( f \) is a uniformly random index over \( \{1,\ldots,n\} \). Let \( T \) be a subset of \( \{1,\ldots,n\} \) with \( m \) indices drawn uniformly at random without replacement. Then, for any \( t \geq \sqrt{m\|V\|_2^2 + K/3} \), we have

\[
\mathbb{P} \left\{ \left\| \sum_{i \in T} X_i \right\|_2 \geq t \right\} \leq 8 \cdot d_e \cdot \exp \left( -\frac{t^2}{2m\|V\|_2 + Kt/3} \right),
\]

where \( d_e := \text{tr}(V)/\|V\| \) is the intrinsic dimension of the matrix \( V \).

**Proof.** We denote \( S_T := \sum_{i \in T} X_i \). Fix \( \theta > 0 \), define \( \psi(t) := e^{\theta t} - \theta t - 1 \), and use the Laplace matrix transform method (e.g., Proposition 7.4.1 in [27]) to obtain

\[
\mathbb{P} \{ \lambda_{\max}(S_T) \geq t \} \leq \frac{1}{\psi(t)} \mathbb{E} \text{tr} \psi(S_T)
\]

\[
= \frac{1}{e^{\theta t} - \theta t - 1} \mathbb{E} \text{tr} \left( e^{\theta S_T} - I \right),
\]

and the last equality holds due to the fact that \( \mathbb{E} S_T = m \mathbb{E} X_T = 0 \). Let \( T' = \{i_1,\ldots,i_m\} \) be a subset of \( \{1,\ldots,n\} \), drawn uniformly at random with replacement. In particular, the indices of \( T' \) are independent random variables, and so are the matrices \( \{X_{i_j}\}_{j=1}^m \). Write \( S_{T'} := \sum_{j=1}^m X_{i_j} \). Gross and Nesme [11] have shown that for any \( \theta > 0 \),

\[
\mathbb{E} \text{tr} \exp(\theta S_T) \leq \mathbb{E} \text{tr} \exp(\theta S_{T'}) .
\]
As a consequence of Lieb’s inequality (e.g., Lemma 3.4 in [27]), it holds that

$$E \, \text{tr} \exp (\theta S_T) \leq \text{tr} \exp \left( \sum_{j=1}^{m} \log E \, e^{\theta X_j} \right) = \text{tr} \exp \left( m \log E \, e^{\theta X_j} \right).$$

Thus, it remains to bound $E \, e^{\theta X_j}$. By assumption, $E[X_j] = 0$ and $\|X_j\|_2 \leq K$ almost surely. Then, using Lemma 5.4.10 from [28], we get $E \, e^{\theta X_j} \leq \exp(g(\theta) \, E \, X_j^2)$, for any $|\theta| < 3/K$ and where $g(\theta) = \frac{\theta^2/2}{1-|\theta|K/3}$. By monotonicity of the logarithm, $m \cdot \log E \, e^{\theta X_j} \leq m \cdot g(\theta) \, E \, X_j^2$. By assumption, $E \, X_j^2 \leq V$ and thus, $m \cdot \log E \, e^{\theta X_j} \leq m \cdot g(\theta) \, V$. By monotonicity of the trace exponential, it follows that $\text{tr} \exp(m \log E \, e^{\theta X_j}) \leq \text{tr} \exp(m \cdot g(\theta) \, V)$, and further,

$$P \{ \lambda_{\max}(S_T) \geq t \} \leq \frac{1}{e^{\theta t} - \theta t - 1} \cdot \text{tr} \left( e^{\theta t} V - I \right) = \frac{1}{e^{\theta t} - \theta t - 1} \cdot \text{tr} \varphi(m \cdot g(\theta) \, V),$$

where $\varphi(a) = e^a - 1$. The function $\varphi$ is convex, and the matrix $m \cdot g(\theta) \, V$ is positive semidefinite. Therefore, we can apply Lemma 7.5.1 from [27] and obtain

$$\text{tr} \varphi(m \cdot g(\theta) \, V) \leq \epsilon \cdot \|m \cdot g(\theta) \|_2 \leq \epsilon \cdot e^{m \cdot g(\theta) \|V\|_2},$$

which further implies that

$$P \{ \lambda_{\max}(S_T) \geq t \} \leq \epsilon \cdot \frac{e^{\theta t} \cdot \epsilon \cdot e^{-\theta t + m \cdot g(\theta) \|V\|_2}}{e^{\theta t} - \theta t - 1} \leq \epsilon \cdot \left( 1 + \frac{3}{\theta^2 t^2} \right) \cdot e^{-\theta t + m \cdot g(\theta) \|V\|_2}.$$  

For the last inequality, we used the fact that $\frac{e^a}{e^a - a - 1} = 1 + \frac{1 + a}{e^a - a - 1} \leq 1 + \frac{3}{a^2}$ for all $a \geq 0$. Picking $\theta = t/(m \|V\|_2 + Kt/3)$, we obtain

$$P \{ \lambda_{\max}(S_T) \geq t \} \leq \epsilon \cdot \left( 1 + 3 \cdot \frac{(m \|V\|_2 + Kt/3)^2}{t^4} \right) \cdot \exp \left( -\frac{t^2/2}{m \|V\|_2 + Kt/3} \right).$$

Under the assumption $t \geq m \|V\|_2 + Kt/3$, the parenthesis in the above right-hand side is bounded by four, which results in

$$P \{ \lambda_{\max}(S_T) \geq t \} \leq 4 \cdot \epsilon \cdot \text{exp} \left( -\frac{t^2/2}{m \|V\|_2 + Kt/3} \right).$$

Repeating the argument for $-S_T$ and combining the two bounds, we obtain the claimed result.  

\textbf{E.2.1 Proof of Theorem 4}  

We write $v_j := \sqrt{\frac{n}{m}} \, w_j$, where $w_j = e_j^T \, H \, \text{diag}(\varepsilon) \, U \, \bar{D}$, and $\varepsilon \in \{\pm 1\}^n$ is a fixed vector. We denote

$$\gamma := \max \left\{ \max_{j=1, \ldots, n} \|v_j\|, \, m^{-\frac{1}{2}} \right\} \quad \text{and} \quad X_i := v_i v_i^T - \frac{1}{m} \bar{D}^2.$$

Let $I$ be a uniformly random index over $\{1, \ldots, n\}$. We have

$$E[X_I] = \frac{n}{m} E[w_I w_I^T] - \frac{1}{m} \bar{D}^2 = \frac{n}{m} \left( \frac{1}{n} \sum_{i=1}^{n} \bar{D} U^T \text{diag}(\varepsilon) H e_i e_i^T H \text{diag}(\varepsilon) U \bar{D} \right) - \frac{1}{m} \bar{D}^2$$

$$= \frac{1}{m} \bar{D} U^T \text{diag}(\varepsilon) H \sum_{i=1}^{n} e_i e_i^T H \text{diag}(\varepsilon) U \bar{D} - \frac{1}{m} \bar{D}^2$$

$$= 0.$$

The last equality holds due to the fact that $H^2 = I$, $\text{diag}(\varepsilon)^2 = I$ and $U^T U = I$. Further, $\|v_I\|^2 \leq \gamma^2$ a.s., so that $\|v_I\|^2 v_I v_I^T \leq \gamma^2 v_I v_I^T$ a.s., and consequently, $E \left( \|v_I\|^2 v_I v_I^T \right) \leq \gamma^2 \cdot E[v_I v_I^T]$. Thus,

$$E \left[ X_I^2 \right] = E \left[ \|v_I\|^2 v_I v_I^T \right] - \frac{2}{m} \bar{D}^2 + \frac{1}{m^2} \bar{D}^4$$

$$\leq \frac{\gamma^2}{m} \bar{D}^2 - \frac{2}{m^2} \bar{D}^4 + \frac{1}{m^2} \bar{D}^4$$

$$= \frac{\gamma^2}{m} \bar{D}^2 - \frac{1}{m^2} \bar{D}^4$$

$$\leq \frac{\gamma^2}{m} \bar{D}^2.$$
The first inequality holds due to the fact that $E[v_j v_j^T] = m^{-1} D^2$. Further, we have

$$\|X_t\| = \|v_j v_j^T - \frac{1}{m} D^2\| \leq \max \left\{ \max_{j=1,...,n} \|v_j\|^2, m^{-1} \right\} = \gamma^2 .$$

Let $T$ be a subset of $m$ indices in $\{1, \ldots, n\}$ drawn uniformly at random, without replacement. Applying Theorem 10 with $V = m^{-1} \gamma^2 D^2$ and using the scale invariance of the effective dimension, we obtain that for any $t \geq \gamma + \gamma^2 / 3$,

$$\mathbb{P} \left\{ \left\| \sum_{i \in T} X_i \right\|_2 \geq t \right\} \leq 8d_e \cdot \exp \left( -\frac{t^2/2}{\gamma^2 (1 + t/3)} \right) .$$

Suppose now that $\varepsilon$ is a vector of independent Rademacher variables. Note that $\sum_{i \in T} X_i \overset{d}{=} DU^T (S^T S - I) UD$. From Lemma 5 we know that $\gamma \leq \sigma := \sqrt{\frac{d_e}{m} + \sqrt{\frac{8 \log(d_e n)}{m}}}$ with probability at least $1 - d_e^{-1}$. Consequently, with probability at least $1 - d_e^{-1} - 8d_e \cdot \exp \left( -\frac{t^2/2}{\sigma^2 (1+t/3)} \right)$, for $t \geq \sigma (1 + \sigma/3)$ we have

$$\left\| DU^T (S^T S - I) UD \right\|_2 \leq t . \quad (17)$$

We set $t = \sigma \sqrt{8/3 \log d_e}$, and $\rho = \frac{d_e \log(d_e)}{m} C(n,d_e)$ where $C(n,d_e) = \frac{16}{3} \left( 1 + \sqrt{\frac{\log(d_e n)}{d_e}} \right)^2$. We choose $m$ large enough so that $\rho \leq \left( 1 - (8/3 \log d_e)^{-\frac{2}{3}} \right)^2$. Then, we get that

$$\mathbb{P} \left\{ \left\| DU^T (S^T S - I) UD \right\|_2 \geq \|D\|_2 \cdot \sqrt{\rho} \right\} \leq \frac{9}{d_e} ,$$

which is the claimed result. \hfill \square

E.2.2 An additional technical result: the case $d_e > m$

**Lemma 6.** Suppose that $m < d_e$, define $C(n,d_e) = \frac{16}{3} \left( 1 + \sqrt{\frac{\log(d_e n)}{d_e}} \right)^2$, and set $\rho = \frac{d_e \log(d_e)}{m} C(n,d_e)$. Then, we have that

$$\left\| DU^T (S^T S - I) UD \right\|_2 \leq \|D\|_2 \cdot \rho ,$$

with probability at least $1 - \frac{9}{d_e}$.

**Proof.** We use the notations of the proof of Theorem 4. From the assumption $m < d_e$, we have that $\sigma > 1$. We set $t = \frac{3}{8} \sigma^2 \log(d_e^2)$, so that $t > 1$. Then,

$$\exp \left( -\frac{t^2/2}{\sigma^2 (1 + t/3)} \right) \leq \exp \left( -\frac{3}{8} \cdot \frac{t}{\sigma^2} \right) = \frac{1}{d_e^2} .$$

It is easy to verify that $t \geq \sigma (1 + \sigma/3)$. Thus, using (17), we find that

$$\mathbb{P} \left\{ \left\| DU^T (S^T S - I) UD \right\|_2 \geq \|D\|_2 \cdot \rho \right\} \leq \frac{9}{d_e} ,$$

which is the claimed result. \hfill \square