Optimal Embedding Dimension for Sparse Subspace Embeddings

Shabarish Chenakkod
University of Michigan
Ann Arbor, USA
shabari@umich.edu

Michał Dereziński
University of Michigan
Ann Arbor, USA
derezin@umich.edu

Xiaoyu Dong
University of Michigan
Ann Arbor, USA
xydong@umich.edu

Mark Rudelson
University of Michigan
Ann Arbor, USA
rudelson@umich.edu

ABSTRACT

A random $m \times n$ matrix S is an oblivious subspace embedding (OSE) with parameters $\epsilon > 0$, $\delta \in (0, 1/3)$ and $d \leq m \leq n$, if for any d-dimensional subspace $W \subseteq \mathbb{R}^n$,

$$\mathbb{P}\left(\forall_{x \in W}\ (1+\epsilon)^{-1}\|x\| \leq \|Sx\| \leq (1+\epsilon)\|x\|\right) \geq 1 - \delta.$$

It is known that the embedding dimension of an OSE must satisfy $m \geq d$, and for any $\theta > 0$, a Gaussian embedding matrix with $m \geq (1+\theta)d$ is an OSE with $\epsilon = O_\theta(1)$. However, such optimal embedding dimension is not known for other embeddings. Of particular interest are sparse OSEs, having $s \ll m$ non-zeros per column (Clarkson and Woodruff, STOC 2013), with applications to problems such as least squares regression and low-rank approximation.

We show that, given any $\theta > 0$, an $m \times n$ random matrix S with $m \geq (1+\theta)d$ consisting of randomly sparsified $\pm 1/\sqrt{s}$ entries and having $s = O(\log^4(d))$ non-zeros per column, is an oblivious subspace embedding with $\epsilon = O_\theta(1)$. Our result addresses the main open question posed by Nelson and Nguyen (FOCS 2013), who conjectured that sparse OSEs can achieve $m = O(d)$ embedding dimension, and it improves on $m = O(d\log(d))$ shown by Cohen (SODA 2016). We use this to construct the first oblivious subspace embedding with $O(d)$ embedding dimension that can be applied faster than current matrix multiplication time, and to obtain an optimal single-pass algorithm for least squares regression.

We further extend our results to Leverage Score Sparsification (LESS), which is a recently introduced non-oblivious embedding technique. We use LESS to construct the first subspace embedding with low distortion $\epsilon = o(1)$ and optimal embedding dimension $m = O(d/\epsilon^2)$ that can be applied in current matrix multiplication time, addressing a question posed by Cherapanamjeri, Silwal, Woodruff and Zhou (SODA 2023).

CCS CONCEPTS

• **Theory of computation → Sketching and sampling**; **Random projections and metric embeddings**.

STOC '24, June 24–28, 2024, Vancouver, BC, Canada

ACM ISBN 979-8-4007-0383-6/24/06
https://doi.org/10.1145/3618260.3649762

KEYWORDS

Matrix sketching, Subspace embeddings, Randomized Numerical Linear Algebra

ACM Reference Format:
Shabarish Chenakkod, Michał Dereziński, Xiaoyu Dong, and Mark Rudelson. 2024. Optimal Embedding Dimension for Sparse Subspace Embeddings. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC '24), June 24–28, 2024, Vancouver, BC, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3618260.3649762

1 INTRODUCTION

Since their first introduction by Sarlós [35], subspace embeddings have been used as a key dimensionality reduction technique in designing fast approximate randomized algorithms for numerical linear algebra problems, including least squares regression, l_p regression, low-rank approximation, and approximating leverage scores, among others [9–11, 21, 29, 33, 38]; we refer to the surveys [19, 22, 24, 28, 30, 39] for an overview. The subspace embedding property can be viewed as an extension of the classical Johnson-Lindenstrauss (JL) lemma [26], which provides a transformation that reduces the dimensionality of a finite set of n-dimensional vectors while preserving their pairwise distances (i.e., an embedding). Here, instead of a finite set, we consider a d-dimensional subspace of $\mathbb{R}^n$, where $d \ll n$. One way to define such a transformation is via an $m \times n$ random matrix S, where $m \ll n$ is the *embedding dimension*. Remarkably, when the embedding dimension m and the distribution of S are chosen correctly, then matrix S can provide an embedding for any d-dimensional subspace W with high probability. Such a distribution of S is *oblivious* to the choice of the subspace. Since any d-dimensional subspace of $\mathbb{R}^n$ can be represented as the range of an $n \times d$ matrix U with orthonormal columns, we arrive at the following definition.

Definition 1.1. A random $m \times n$ matrix S is an (ϵ, δ)-*subspace embedding* (SE) for an $n \times d$ matrix U with orthonormal columns, where $\epsilon > 0$, $\delta \in (0, 1/3)$ and $d \leq m \leq n$, if

$$\mathbb{P}\left(\forall_{x \in \mathbb{R}^d}\ (1+\epsilon)^{-1}\|x\| \leq \|SUx\| \leq (1+\epsilon)\|x\|\right) \geq 1 - \delta.$$

If S is an (ϵ, δ)-SE for all such U, then it is an (ϵ, δ, d)-*oblivious subspace embedding* (OSE).

The embedding dimension of any OSE must satisfy $m \geq d$, so ideally we would like an embedding with m as close to d as possible, while making sure that the distortion ϵ is not too large. For many

applications, a constant distortion, i.e., $\epsilon = O(1)$, is sufficient, while for some, one aims for low-distortion embeddings with $\epsilon \ll 1$.

One of the most classical families of OSE distributions are Gaussian matrices S with i.i.d. entries $s_{i,j} \sim \mathcal{N}(0, 1/m)$. Thanks to classical results on the spectral distribution of Gaussian matrices [34] combined with their rotational invariance, there is a sharp characterization of the distortion factor ϵ for Gaussian subspace embeddings, which implies that the embedding dimension can be arbitrarily close to d, i.e., $m = (1+\theta)d$ for any constant $\theta > 0$, while ensuring the OSE property with $\epsilon = O(1)$ and $\delta = \exp(-\Omega(d))$, where the big-O notation hides the dependence on θ. These guarantees have been partly extended, although only with a sub-optimal constant $\theta = O(1)$, to a broader class of random matrices that satisfy the Johnson-Lindenstrauss property, including dense subgaussian embeddings such as matrices with i.i.d. random sign entries scaled by $1/\sqrt{m}$.

Dense Gaussian and subgaussian embeddings are too expensive for many applications, due to the high cost of dense matrix multiplication. One of the ways of addressing this, as proposed by Clarkson and Woodruff [11], is to use very sparse random sign matrices, where the sparsity is distributed uniformly so that there are $s \ll m$ non-zero entries per column of S (we refer to s as the column-sparsity of S). Remarkably, choosing column-sparsity $s = 1$ (which is the minimum necessary sparsity for any OSE [29]) is already sufficient to obtain a constant distortion OSE, but only if we increase the embedding dimension to $m = O(d^2)$. On the other hand, Nelson and Nguyen [31], along with follow up works [3], showed that if we allow $s = \text{polylog}(d)$, then we can get an OSE with $m = d\,\text{polylog}(d)$. This was later improved by Cohen [12] to $m = O(d\log(d))$ with column-sparsity $s = O(\log(d))$. Nevertheless, the embedding dimension of sparse OSEs remains sub-optimal, not just by a constant, but by an $O(\log(d))$ factor, due to a fundamental limitation of the matrix Chernoff analysis employed by [12]. Thus, we arrive at the central question of this work:

What is the optimal embedding dimension for sparse oblivious subspace embeddings?

This question is essentially the main open question posed by Nelson and Nguyen [31, Conjecture 14]: They conjectured that a sparse random sign matrix with $s = O(\log(d))$ non-zeros per column achieves a constant distortion OSE with embedding dimension $m = O(d)$, i.e., within a constant factor of the optimum. We go one step further and ask whether a sparse OSE can recover the *optimal* embedding dimension, i.e., $m = (1+\theta)d$ for any $\theta > 0$, achieved by Gaussian embeddings. Note that this version of the question is open for *any* sparsity s and any distortion ϵ, even including dense random sign matrices (i.e., $s = m$).

1.1 Main results

In our main result, we show that embedding matrices with column-sparsity s polylogarithmic in d can recover the optimal Gaussian embedding dimension $m = (1+\theta)d$, while achieving constant distortion ϵ. The below result applies to several standard sparse embedding constructions, including a construction considered by Nelson and Nguyen (among others), where, we split each column of S into s sub-columns and sample a single non-zero entry in each sub-column, assigning a random $\pm 1/\sqrt{s}$ to each of those entries.

Theorem 1.2 (Sparse OSEs; informal Theorem 3.3). *Given any constant $\theta > 0$, an $m \times n$ sparse embedding matrix S with $n \geq m \geq (1+\theta)d$ and $s = O(\log^4(d))$ non-zeros per column is an oblivious subspace embedding with distortion $\epsilon = O(1)$. Moreover, given any ϵ, δ, it suffices to use $s = O(\log^4(d/\delta)/\epsilon^6)$ to get an (ϵ,δ)-OSE with $m = O((d+\log 1/\delta)/\epsilon^2)$.*

The embedding dimension of $m = O(d/\epsilon^2)$ exactly matches a known lower bound of $m = \Omega(d/\epsilon^2)$, given by Nelson and Nguyen [32]. To our knowledge, this result is the first to achieve the optimal embedding dimension for a sparse OSE with any sparsity $s = o(m)$, or indeed, for any OSE that can be applied faster than dense $d \times d$ matrix multiplication, including recent efforts [7, 8] (see Theorem 1.4 for our fast OSE algorithm).

A known lower bound on the level of sparsity achievable by any oblivious subspace embedding is a single non-zero entry per column ($s = 1$) of the embedding matrix S [29]. However, this limit can be circumvented by *non-oblivious* sparse embeddings, i.e., when we have additional information about the orthonormal matrix $U \in \mathbb{R}^{n\times d}$ that represents the subspace. Of particular significance is the distribution of the squared row norms of U (also known as leverage scores), which encode the relative importance of the rows of U in constructing a good embedding. Knowing accurate approximations of the leverage scores lies at the core of many subspace embedding techniques, including approximate leverage score sampling [21, 23] and the Subsampled Randomized Hadamard Transform [1, 37]. These approaches rely on the fact that simply sub-sampling $m = O(d\log d)$ rows of U proportionally to their (approximate) leverage scores is a constant distortion subspace embedding. This corresponds to the $m \times n$ embedding matrix S having one non-zero entry per *row* (a.k.a. a sub-sampling matrix), which is much sparser than any OSE since $n \gg m$. However, the sub-sampling embeddings are bound to the sub-optimal $O(d\log d)$ embedding dimension m, and it is not known when we can achieve the optimal $m = (1+\theta)d$ or even $m = O(d)$. We address this in the following result, showing that a non-oblivious sparse embedding knowing leverage score estimates requires only polylogarithmic in d row-sparsity (non-zeros per row). To do this, we construct embedding matrices with a non-uniform sparsity pattern that favors high-leverage rows, inspired by recently proposed Leverage Score Sparsified embeddings [16–18].

Theorem 1.3 (Sparser Non-oblivious SE; informal Theorem 4.3). *Consider $\alpha \geq 1$ and any matrix $U \in \mathbb{R}^{n\times d}$ such that $U^\top U = I$. Given α-approximations of all squared row norms of U, we can construct a $(1+\theta)d \times n$ subspace embedding for U having $O(\alpha\log^4 d)$ non-zeros per row and $\epsilon = O(1)$. Moreover, for any ϵ, δ, we can construct an (ϵ,δ)-SE for U with $O(\alpha\log^4(d)/\epsilon^4)$ non-zeros per row and embedding dimension $m = O((d+\log 1/\delta)/\epsilon^2)$.*

Even though the above result focuses on non-oblivious embeddings, one of its key implications is a new guarantee for a classical family of oblivious embeddings known as Fast Johnson-Lindenstrauss Transforms (FJLT), introduced by Ailon and Chazelle [1]. An FJLT is defined as $S = \Phi HD$, where Φ is an $m \times n$ uniformly sparsified embedding matrix, H is an $n \times n$ orthogonal matrix with

Ref.	Oblivious	Dimension m		Runtime
[7]	✓	$O(d \cdot \mathrm{pll}(d))$	–	$O(\mathrm{nnz}(A) + d^{2+\gamma})$
[7]	–	$O(d \log(d)/\epsilon^2)$	–	$O(\mathrm{nnz}(A) + d^{\omega} \mathrm{pll}(d))$
[8]	–	$O(d)$	✓	$O(\mathrm{nnz}(A) + d^{2+\gamma})$
[8]	–	$O(d \log(d)/\epsilon^2)$	–	$O(\mathrm{nnz}(A) + d^{\omega})$
Thm. 1.4	✓	$O(d)$	✓	$O(\mathrm{nnz}(A) + d^{2+\gamma})$
Thm. 1.6	–	$O(d/\epsilon^2)$	✓	$O(\mathrm{nnz}(A) + d^{\omega})$

Table 1: Comparison of our results to recent prior works towards obtaining fast subspace embeddings with optimal embedding dimension. For clarity of presentation, we assume that the distortion satisfies $\epsilon = \Omega(d^{-c})$ for a small constant $c > 0$, and we use $\mathrm{pll}(d)$ to denote $\mathrm{poly}(\log \log d)$. We use a checkmark ✓ to indicate which embeddings are oblivious, and which of them achieve optimal dependence of dimension m relative to the distortion ϵ.

fast matrix-vector products (e.g., a Fourier or Walsh-Hadamard matrix), and D is a diagonal matrix with random ± 1 entries. This embedding is effectively a two-step procedure: first, we use HD to randomly rotate the subspace defined by U, obtaining $\tilde{U} = HDU$ which has nearly-uniform leverage scores [37]; then we apply a uniformly sparsified embedding Φ to $\tilde{U}$, knowing that the uniform distribution is a good approximation for the leverage scores of $\tilde{U}$. Theorem 1.3 implies that an FJLT with $O(\log^4(d)/\epsilon^4)$ non-zeros per row is an (ϵ, δ, d)-OSE with the optimal embedding dimension $m = O((d + \log 1/\delta)/\epsilon^2)$ (see Theorem 4.4 for details). To our knowledge, this is the first optimal dimension OSE result for FJLT matrices.

Yet, the application to FJLTs does not leverage the full potential of Theorem 1.3, which is particularly useful for efficiently constructing optimal subspace embeddings when only coarse leverage score estimates are available, which has arisen in recent works on fast subspace embeddings [7, 8]. In the following section, we use it to construct a new fast low-distortion SE (i.e., $\epsilon \ll 1$) with optimal embedding dimension, addressing a question posed by Cherapanamjeri, Silwal, Woodruff and Zhou [8] (see Theorem 1.6 for details).

1.2 Fast subspace embeddings

Next, we illustrate how our main results can be used to construct fast subspace embeddings with optimal embedding dimension. In most applications, OSEs are used to perform dimensionality reduction on an $n \times d$ matrix A by constructing the smaller $m \times d$ matrix SA. The subspace embedding condition ensures that $\|SAx\| \approx \|Ax\|$ for all $x \in \mathbb{R}^d$ up to a multiplicative factor $1+\epsilon$, which has numerous applications, including fast linear regression (see Section 1.3). The key computational bottleneck here is the cost of computing SA. We aim for input sparsity time, i.e., $O(\mathrm{nnz}(A))$, where $\mathrm{nnz}(A)$ is the number of non-zeros in A, possibly with an additional small polynomial dependence on d. Our results for computing a fast subspace embedding SA with optimal embedding dimension are summarized in Table 1, alongside recent prior works.

In the following result, we build on Theorem 1.2 to provide an input sparsity time algorithm for constructing a fast oblivious subspace embedding with constant distortion $\epsilon = O(1)$, and optimal embedding dimension $m = O(d)$. This is the first optimal OSE construction that is faster than current matrix multiplication time $O(d^{\omega})$.

Theorem 1.4 (Fast oblivious subspace embedding). *Given $d \le n$ and any $\gamma > 0$, there is a distribution over $m \times n$ matrices S where $m = O(d)$, such that for any $A \in \mathbb{R}^{n \times d}$, with probability at least 0.9:*

$$\frac{1}{2}\|Ax\| \le \|SAx\| \le 2\|Ax\| \qquad \forall x \in \mathbb{R}^d,$$

and SA can be computed in $O(\gamma^{-1} \mathrm{nnz}(A) + d^{2+\gamma} \mathrm{polylog}(d))$ time. Moreover, for $\gamma = \Omega(1)$, we can generate such a random matrix S using only $\mathrm{polylog}(nd)$ many uniform random bits.

Remark 1.5. The problem of constructing an OSE using $\mathrm{polylog}(nd)$ many random bits was also brought up by Nelson and Nguyen, who obtained this with $m = d\ \mathrm{polylog}(d)$. To achieve it with $m = O(d)$, we introduce a new sparse construction, likely of independent interest, where the non-zeros are distributed along the diagonals instead of the columns of S.

The runtime of our method matches the best known non-oblivious SE with $m = O(d)$, recently obtained by [8], while at the same time being much simpler to implement: their construction requires solving a semidefinite program to achieve the optimal dimension, while we simply combine several sparse matrix multiplication steps. Moreover, thanks to its obliviousness, our embedding can be easily adapted to streaming settings. For example, consider numerical linear algebra in the turnstile model [10], where we wish to maintain a sketch of A while receiving a sequence of updates $A_{i,j} \leftarrow A_{i,j} + \delta$. Using the construction from Theorem 1.4, we can maintain a constant-distortion subspace embedding of A in the turnstile model with optimal space of $O(d^2 \log(nd))$ bits, while reducing the update time exponentially, from $O(d)$ (for a dense OSE matrix) to $\mathrm{polylog}(d)$ time.

In the next result, we build on Theorem 1.3 to provide the first subspace embedding with low distortion $\epsilon = o(1)$ and optimal embedding dimension $m = O(d/\epsilon^2)$ that can be applied in current matrix multiplication time. This addresses the question posed by Cherapanamjeri, Silwal, Woodruff and Zhou [8], who gave a current matrix multiplication time algorithm for a low-distortion subspace embedding, but with a sub-optimal embedding dimension $m = O(d \log(d)/\epsilon^2)$. We are able to improve upon this embedding dimension by replacing leverage score sampling (used by [8]) with our leverage score sparsified embedding construction, developed as part of the proof of Theorem 1.3.

Theorem 1.6 (Fast low-distortion subspace embedding). *Given $A \in \mathbb{R}^{n \times d}$ and $\epsilon > 0$, we can compute an $m \times d$ matrix SA such that $m = O(d/\epsilon^2)$, and with probability at least 0.9:*

$$(1+\epsilon)^{-1}\|Ax\| \le \|SAx\| \le (1+\epsilon)\|Ax\| \qquad \forall x \in \mathbb{R}^d,$$

in time $O(\gamma^{-1} \mathrm{nnz}(A) + d^{\omega} + \mathrm{poly}(1/\epsilon) d^{2+\gamma} \mathrm{polylog}(d))$ for any $0 < \gamma < 1$.

1.3 Applications to linear regression

Our fast subspace embeddings can be used to accelerate numerous approximation algorithms in randomized numerical linear algebra,

including for linear regression, low-rank approximation, rank computation and more. Here, we illustrate this with the application to linear regression tasks. In the following result, we use Theorem 1.4 to provide the first single pass algorithm for a relative error least squares approximation with optimal both time and space complexity.

Theorem 1.7 (Fast least squares). *Given an $n \times d$ matrix A and an $n \times 1$ vector b, specified with $O(\log nd)$-bit numbers, consider the task of finding $\tilde{x}$ such that:*

$$\|A\tilde{x} - b\|_2 \le (1+\epsilon) \min_x \|Ax - b\|_2.$$

The following statements are true for this task:

1. *For $\epsilon = \Theta(1)$, we can find $\tilde{x}$ with a single pass over A and b in $O(\mathrm{nnz}(A) + d^\omega)$ time, using $O(d^2 \log(nd))$ bits of space.*
2. *For arbitrary $\epsilon > 0$, we can compute $\tilde{x}$ in $O(\epsilon^{-1}\,\mathrm{nnz}(A) + d^\omega + d^{2+\gamma}/\epsilon)$ time, using $O(d^2 \log(nd))$ bits of space, for any $0 < \gamma < 1$.*

For part (1) of the claim, we note that the obtained space complexity matches the lower bound $\Omega(d^2 \log(nd))$ of Clarkson and Woodruff [10]. Moreover, it is clear that solving a least squares problem with any worst-case relative error guarantee requires at least reading the entire matrix A and solving a $d \times d$ linear system, which implies that the $O(\mathrm{nnz}(A) + d^\omega)$ time is also optimal. For part (2) of the claim, we note that a similar time complexity for a $1+\epsilon$ (non-single-pass) least squares approximation was shown by [8], except they had an additional $O(\epsilon^{-1} d^2\,\mathrm{polylog}(d)\log(1/\epsilon))$. We avoid that extra term, thereby obtaining the correct $O(1/\epsilon)$ dependence on the relative error, by employing a carefully tuned preconditioned mini-batch stochastic gradient descent with approximate leverage score sampling. This approach is of independent interest, as it is very different from that of [8], who computed a sketch-and-solve estimate by running preconditioned gradient descent on the sketch.

Finally, we point out that our fast low-distortion subspace embeddings (Theorem 1.6) can be used to construct reductions for a wider class of constrained/regularized least squares problems, which includes Lasso regression among others [3]. The following result provides the first $O(d/\epsilon^2) \times d$ such reduction for $\epsilon = o(1)$ in current matrix multiplication time.

Theorem 1.8 (Fast reduction for constrained/regularized least squares). *Given $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$ and $\epsilon > 0$, consider an $n \times d$ linear regression task $T(A, b, \epsilon)$ of finding $\tilde{x}$ such that:*

$$f(\tilde{x}) \le (1+\epsilon) \min_{x\in C} f(x), \quad \text{where} \quad f(x) = \|Ax - b\|_2^2 + g(x),$$

for some $g : \mathbb{R}^d \to \mathbb{R}_{\ge 0}$ and a set $C \subseteq \mathbb{R}^d$. We can reduce this task to solving an $O(d/\epsilon^2) \times d$ instance $T(\tilde{A}, \tilde{b}, 0.1\epsilon)$ in $O(\gamma^{-1}\,\mathrm{nnz}(A) + d^\omega + \mathrm{poly}(1/\epsilon) d^{2+\gamma}\,\mathrm{polylog}(d))$ time.

1.4 Overview of techniques

One of the key ingredients in our analysis involves establishing the universality of a class of random matrices, building on the techniques of Brailovskaya and Van Handel [4], by characterizing when the spectrum of a sum of independent random matrices is close to that of a Gaussian random matrix whose entries have the same mean and covariance. We adapt these techniques to a class of nearly-square random matrices that arise from applying an $m \times n$ sparse random matrix S to an $n \times d$ isometric embedding matrix U, showing high probability bounds for the Hausdorff distance between the spectrum of SU and the spectrum of a corresponding Gaussian random matrix.

A key limitation of the results of [4] is that they require full independence between the random matrices in a sum (which correspond to sub-matrices of the matrix S), unlike, for instance, the analysis of Nelson and Nguyen [31] which uses a moment method that only requires $O(\log(d))$-wise independence. We address this with the *independent diagonals* construction: we propose a distribution over $m \times n$ sparse random matrices S where the non-zeros are densely packed into a small number of diagonals (see Figure 1) so that, while the diagonals are fully independent, the entries within a single diagonal only need to be 2-wise independent. As a consequence, the resulting construction requires only $n/m \cdot \mathrm{polylog}(n)$ uniform random bits to generate, and we further improve that to $\mathrm{polylog}(n)$ by combining it with the Nelson-Nguyen embedding.

Standard sparse embedding matrices are not very effective at producing low-distortion subspace embeddings, i.e., with $\epsilon = o(1)$, because their density (non-zeros per column) has to grow with $1/\epsilon$, so that their complexity is no longer input sparsity time. Prior work has dealt with this problem by using a constant distortion subspace embedding as a preconditioner for computing the leverage score estimates $l_1, ..., l_n$ of the input matrix A [7], and then constructing a subspace embedding in a non-oblivious way out of a sub-sample of $m = O(d \log d/\epsilon^2)$ rows of A. This leverage score sampling scheme is effectively equivalent to using an extremely sparse embedding matrix S which has a single non-zero entry $S_{i,I_i} \sim \pm 1/\sqrt{l_i}$ in each row, with its index I_i sampled according to the leverage score distribution $(l_1/Z, ..., l_n/Z)$, where $Z = \sum_i l_i$. Unfortunately due to the well-known coupon collector problem, such a sparse embedding matrix cannot achieve the optimal embedding dimension $m = O(d/\epsilon^2)$. We circumvent this issue by making the embedding matrix S slightly denser, with $\alpha\,\mathrm{poly}(1/\epsilon)\,\mathrm{polylog}(d)$ non-zeros per row, where α is the approximation factor in the leverage score distribution (i.e., *leverage score sparsification*, see Figure 2). Unlike the oblivious sparse embedding, here it is the row-density (instead of column-density) that grows with $1/\epsilon$, which means that the overall algorithm can still run in input sparsity time. We note that our algorithms use $\alpha = O(d^\gamma)$ approximation factor for the leverage scores, where $0 < \gamma < 1$ is a parameter that can be chosen arbitrarily. This parameter reflects a trade-off in the runtime complexity, between the $O(\gamma^{-1}(\mathrm{nnz}(A) + d^2))$ cost of estimating the leverage scores, and the density of the leverage score sparsified embedding.

To construct a least squares approximation with $\epsilon = o(1)$ (Theorem 1.7 part 2), we use our constant distortion subspace embedding to compute a preconditioner for matrix A. That preconditioner is then used first to approximate the leverage scores, as well as to compute a constant factor least squares approximation, and then to improve the convergence rate of a gradient descent-type algorithm. However, unlike prior works [7, 8, 40], which either use a full gradient or a stochastic gradient based on a single row-sample, we observe that the computationally optimal strategy is to use a stochastic gradient based on a mini-batch of $O(\alpha d)$ rows, where

α is the leverage score approximation factor. With the right sequence of decaying step sizes, this strategy leads to the optimal balance between the cost of computing the gradient estimate and the cost of preconditioning it, while retaining fast per-iteration convergence rate, leading to the $O(\alpha d^2/\epsilon)$ overall complexity of stochastic gradient descent.

In what follows, we provide sketches for our main results about embedding guarantees for oblivious and non-oblivious embeddings. We direct the reader to the full version for the full proofs and applications.

1.5 Related work

Our results follow a long line of work on matrix sketching techniques, which have emerged as part of the broader area of randomized linear algebra; see [19, 22, 24, 28, 30, 39] for comprehensive overviews of the topic. These methods have proven pivotal in speeding up fundamental linear algebra tasks such as least squares regression [11, 33, 35], l_p regression [6, 15, 29, 38], low-rank approximation [14, 27], linear programming [13], and more [25, 36]. Many of these results have also been studied in the streaming and turnstile models [10].

Subspace embeddings are one of the key algorithmic tools in many of the above randomized linear algebra algorithms. Using sparse random matrices for this purpose was first proposed by Clarkson and Woodruff [11], via the CountSketch which has a single non-zero entry per column, and then further developed by several other works [12, 29, 31] to allow multiple non-zeros per column as well as refining the embedding guarantees. Non-uniformly sparsified embedding constructions have been studied recently, including Leverage Score Sparsified embeddings [16–18, 20], although these works use much denser matrices than we propose in this work, as well as relying on somewhat different constructions. There have also been recent efforts on achieving the optimal embedding dimension for subspace embeddings, including [5], who also rely on sparse embeddings, but require additional assumptions on the dimensions of the input matrix as well as its leverage score distribution; and [7, 8], who do not rely on sparse embedding matrices, and therefore do not address the conjecture of Nelson and Nguyen (see Table 1 for a comparison).

2 PRELIMINARIES

Notation. The following notation and terminology will be used in the paper. The notation $[n]$ is used for the set $\{1, 2, ..., n\}$. All matrices considered in this work are real valued and the space of $m \times n$ matrices with real valued entries is denoted by $M_{m\times n}(\mathbb{R})$. The operator norm of a matrix X as $\|X\|$ and its condition number by $\kappa(X)$. For clarity, the operator norm is also denoted by $\|X\|_{op}$ in some places where other norms appear. We shall denote the spectrum of a matrix X, which is the set of all eigenvalues of X, by $\mathrm{spec}(X)$. The standard probability measure is denoted by $\mathbb{P}$, and the symbol $\mathbb{E}$ means taking the expectation with respect to the probability measure. The standard L_p norm of a random variable ξ is denoted by $\|\xi\|_p$, for $1 \le p \le \infty$. Throughout the paper, the symbols $c_1, c_2, ...,$ and $Const, Const', ...$ denote absolute constants.

Oblivious Subspace Embeddings. We define an oblivious subspace embedding, i.e., an (ϵ, δ, d)-OSE, following Definition 1.1, to be a random $m \times n$ matrix S such that for any $n \times d$ matrix U with orthonormal columns (i.e., $U^\top U = I_d$),

$$\mathbb{P}\left(\forall_{x\in\mathbb{R}^d}\ (1+\epsilon)^{-1}\|x\| \le \|SUx\| \le (1+\epsilon)\|x\|\right) \ge 1-\delta. \quad (2.1)$$

For computational efficiency, we usually consider sparse OSEs. A standard construction for a sparse OSE involves i.i.d. rademacher entries in each position, sparsified by multiplication with independent Bernoulli random variables. More precisely, S has i.i.d. entries $s_{i,j} = \delta_{i,j}\xi_{i,j}$ where $\delta_{i,j}$ are independent Bernoulli random variables taking value 1 with probability $p_{m,n,d} \in (0, 1]$ and $\xi_{i,j}$ are i.i.d. random variables with $\mathbb{P}(\xi_{i,j} = 1) = \mathbb{P}(\xi_{i,j} = -1) = 1/2$. Note that this results in S having $s = pm$ many non zero entries per column and pn many non zero entries per row on average. We shall call this the oblivious subspace embedding with independent entries distribution.

Definition 2.1 (OSE-IID-ENT)**.** A $m \times n$ random matrix S is called an oblivious subspace embedding with independent entries (OSE-IID-ENT) if S has i.i.d. entries $s_{i,j} = \delta_{i,j}\xi_{i,j}$ where $\delta_{i,j}$ are independent Bernoulli random variables taking value 1 with probability $p_{m,n,d} \in (0, 1]$ and $\xi_{i,j}$ are i.i.d. random variables with $\mathbb{P}(\xi_{i,j} = 1) = \mathbb{P}(\xi_{i,j} = -1) = 1/2$.

Another example comes from a class of sparse sketching matrices proposed by Nelson and Nguyen [31], called OSNAPs. They define a sketching matrix S as an *oblivious sparse norm-approximating projection* (OSNAP) if it satisfies the following properties -

(1) $s_{ij} = \delta_{ij}\sigma_{ij}/\sqrt{s}$ where σ are i.i.d. ± 1 random variables, and δ_{ij} is an indicator random variable for the event $S_{ij} \neq 0$.
(2) $\forall j \in [n], \sum_{i=1}^{m} \delta_{i,j} = s$ with probability 1, i.e. every column has exactly s non-zero entries.
(3) For any $T \subset [m] \times [n], \mathbb{E}\,\Pi_{(i,j)\in T}\delta_{ij} \le (s/m)^{|T|}$.
(4) The columns of S are i.i.d.

One example of an OSNAP can be constructed as follows when s divides m. In this case, we divide each column of S into s many blocks, with each block having $\frac{m}{s}$ many rows. For each block, we randomly and uniformly select one nonzero entry and set its value to be ± 1 with probability $1/2$. Note that the blocks in each column are i.i.d., and the columns of S are i.i.d. We then see that $S/\sqrt{s}$ satisfies the properties of an OSNAP. For convenience, in this work we will refer to this as the OSNAP distribution, and we will define it using the parameter $p = s/m$ instead of s. To define such a distribution formally, we first define the one hot distribution.

Definition 2.2 (One Hot Distribution)**.** Let M be an $a \times b$ random matrix. Let γ be a random variable taking values in $[a] \times [b]$ with $\mathbb{P}(\gamma = (i, j)) = (1/ab)$. Let ξ be a Rademacher random variable ($\mathbb{P}(\xi = -1) = \mathbb{P}(\xi = 1) = \frac{1}{2}$). M is said to have the one hot distribution if $M = \xi(\sum_{(i,j)\in[a]\times[b]} \mathbf{1}_{\{(i,j)\}}(\gamma)E_{i,j})$ where $E_{i,j}$ is an $a \times b$ matrix with 1 in $(i, j)^{th}$ entry and 0 everywhere else.

Definition 2.3 (OSNAP-IND-COL)**.** An $m \times n$ random matrix S is called an oblivious sparse norm-approximating projection with independent subcolumns distribution (OSNAP-IND-COL) with parameter p such that $s = pm$ divides m, if each submatrix

$$S_{[(m/s)(i-1)+1:(m/s)i]\times\{j\}}$$

of S for $i \in [s], j \in [n]$ has the one hot distribution, and all these submatrices are jointly independent.

Below we collect the existing subspace embedding results for OSNAP matrices, which are relevant to this work.

Lemma 2.4 (Existing sparse embedding guarantees). *The following are some of the known guarantees for OSNAP embedding matrices:*

- *[11] showed that there is an OSNAP matrix S with*
$$m = O(\epsilon^{-2}d^2)$$
rows and 1 non-zero per column (i.e., CountSketch) which is an OSE with distortion ϵ.
- *[12] showed that there is an OSNAP matrix S with*
$$m = O(\epsilon^{-2}d^{1+\gamma}\log d)$$
rows and $s = O(1/\gamma\epsilon)$ non-zero entries per column which is an OSE with distortion ϵ. Note that setting $\gamma = 1/\log(d)$, we get $m = O(\epsilon^{-2}d\log d)$ and $O(\log(d)/\epsilon)$ non-zeros per column.
- *[31] showed that there is an OSNAP matrix using*
$$O(\log(d)\log(nd))$$
uniform random bits with
$$m = O(\epsilon^{-2}d^{1+\gamma}\log^8(d))$$
and $O(1/\gamma^3\epsilon)$ non-zero entries per column.

Non-oblivious subspace embeddings. Following Definition 1.1, we say that an $m \times n$ random matrix S is a (non-oblivious) subspace embedding for a given $n \times d$ matrix U with orthonormal columns if it satisfies (2.1) for that matrix U. In this case, to obtain subspace embedding guarantees with even sparser random matrices, we can use the information about the subspace in the form of its leverage scores. For $i = 1, ..., n$, the ith leverage score of a d-dimensional subspace of $\mathbb{R}^n$ is the squared norm of the ith row of its orthonormal basis matrix U, i.e., $\|e_i^\top U\|_2^2$ (this definition is in fact independent of the choice of basis).

We note that in most applications (e.g., Theorem 1.6), subspace embedding matrices are typically used to transform an arbitrary $n \times d$ matrix A (not necessarily with orthonormal columns), constructing a smaller $m \times d$ matrix SA. In this case, we seek an embedding for the subspace of vectors $\{z : z = Ax \text{ for } x \in \mathbb{R}^d\}$. Here, the corresponding U matrix has columns that form an orthonormal basis for the column-span of A. Thus, in practice we do not have access to matrix U or its leverage scores. Instead, we may compute leverage score approximations [21].

Definition 2.5 (Approximate Leverage Scores)**.** For $\beta_1 \geq 1, \beta_2 \geq 1$, a tuple $(l_1, \ldots, l_n)$ of numbers are (β_1, β_2)-approximate leverage scores for U if, for $1 \leq i \leq n$,
$$\frac{\|e_i^T U\|^2}{\beta_1} \leq l_i \qquad \text{and} \qquad \sum_{i=1}^n l_i \leq \beta_2(\sum_{i=1}^n \|e_i^T U\|^2) = \beta_2 d.$$
And in this case, we also say that they are α-approximations of squared row norms of U with $\alpha = \beta_1\beta_2$.

Uniformizing leverage scores by preconditioning. Another way of utilizing information about the leverage scores to get embedding guarantees with sparser matrices is to precondition the matrix U using the randomized Hadamard transform to uniformize the row norms, resulting in $(d/n, d/n, \ldots d/n)$ becoming approximate leverage scores for the preconditioned matrix. To this end, we first define the Walsh-Hadamard matrix.

Definition 2.6. The Walsh-Hadamard matrix H_{2^k} of dimension $2^k \times 2^k$ for $k \in \mathbb{N} \cup \{0\}$ is the matrix obtained using the recurrence relation
$$H_0 = [1], \qquad H_{2n} = \begin{bmatrix} H_n & H_n \\ H_n & -H_n \end{bmatrix}.$$

In what follows, we drop the subscript of H_{2^k} when the dimension is clear.

Definition 2.7. The *randomized Hadamard transform* (RHT) of an $n \times d$ matrix U is the product $\frac{1}{\sqrt{n}}HDU$, where D is a random $n \times n$ diagonal matrix whose entries are independent random signs, i.e., random variables uniformly distributed on $\{\pm 1\}$. Here, by padding U with zero rows if necessary, we may assume that n is a power of 2.

The key property of the randomized Hadamard transform that we use is that it uniformizes the row norms of U with high probability. More precisely, we have,

Lemma 2.8 (Lemma 3.3, [37]). *Let U be an $n \times d$ matrix with orthonormal columns. Then, $\frac{1}{\sqrt{n}}HDU$ is an $n \times d$ matrix with orthonormal columns, and, for $\delta > 0$*
$$\mathbb{P}\left(\max_{j=1,\ldots,n} \|e_j^T(\frac{1}{\sqrt{n}}HDU)\| \geq \sqrt{\frac{d}{n}} + \sqrt{\frac{8\log(n/\delta)}{n}}\right) \leq \delta$$

Universality. In this paragraph, we describe the random matrix universality result of [4], which is central to our analysis of sparse subspace embedding matrices. The object of study here is a random matrix model given by
$$X := Z_0 + \sum_{i=1}^n Z_i \tag{2.2}$$
where Z_0 is a symmetric deterministic $d \times d$ matrix and $Z_1, \ldots, Z_n$ are symmetric independent random matrices with $\mathbb{E}[Z_i] = 0$. We shall compare the spectrum of X to the spectrum of a gaussian model G that has the same mean and covariance structure as X. More precisely, denoting by $\mathrm{Cov}(X)$ the $d^2 \times d^2$ covariance matrix of the entries of X,
$$\mathrm{Cov}(X)_{i,j,k,l} := \mathbb{E}[(X - \mathbb{E}X)_{ij}(X - \mathbb{E}X)_{kl}]$$
G is the $d \times d$ symmetric random matrix such that:

(1) $\{G_{ij} : i, j \in [d]\}$ are jointly Gaussian
(2) $\mathbb{E}[G] = \mathbb{E}[X]$ and $\mathrm{Cov}(G) = \mathrm{Cov}(X)$.

The above two properties uniquely define the distribution of G. We next define the notion of Hausdorff distance, which will be used in the universality result below.

Definition 2.9 (Hausdorff Distance)**.** Let $A, B \subset \mathbb{R}^n$. Then the Hausdorff distance between A and B is given by,

$$d_H(A,B) = \inf\{\varepsilon \geq 0; A \subseteq B_\varepsilon \text{ and } B \subseteq A_\varepsilon\}$$

where A_ε (resp. B_ε) denotes the ε-neighbourhood of A.

Lemma 2.10 (Theorem 2.4 [4]). *Given the random matrix model (2.2), define the following:*

$$\sigma(X) = \|\mathbb{E}[(X-\mathbb{E}X)^2]\|_{op}^{\frac{1}{2}}$$
$$\sigma_*(X) = \sup_{\|v\|=\|w\|=1} \mathbb{E}[|\langle v, (X-\mathbb{E}X)w\rangle|^2]^{\frac{1}{2}}$$
$$\text{and} \quad R(X) = \| \max_{1\leq i\leq n} \|Z_i\|_{op}\|_\infty.$$

There is a universal constant $C > 0$ such that for any $t \geq 0$,

$$\mathbb{P}\left(d_H\big(\operatorname{spec}(X), \operatorname{spec}(G)\big) > C\epsilon(t)\right) \leq de^{-t},$$
$$\text{where} \quad \epsilon(t) = \sigma_*(X)t^{\frac{1}{2}} + R(X)^{\frac{1}{3}}\sigma(X)^{\frac{2}{3}}t^{\frac{2}{3}} + R(X)t.$$

This result can be viewed as a sharper version of the Matrix Bernstein inequality [37] for the concentration of sums of random matrices. To see this, note that for the random matrix model (2.2), Matrix Bernstein implies that:

$$\mathbb{E}\|X\| \lesssim \sigma(X)\sqrt{\log d} + R(X)\log d,$$

which can be recovered by Lemma 2.10 (see Example 2.12 in [4]). However, Lemma 2.10 together with Theorem 1.2 in [2] implies that:

$$\mathbb{E}(\|X\|) \leq C(\sigma(X) + v(X)^{1/2}\sigma(X)^{1/2}(\log d)^{3/4} + R(X)^{\frac{1}{3}}\sigma(X)^{\frac{2}{3}}(\log d)^{2/3} + R(X)\log d)$$

where $v(X) = \|\mathrm{Cov}(X)\|$ is the norm of the covariance matrix of the d^2 scalar entries. This result can be sharper than the Matrix Bernstein inequality because when $v(X)$ and $R(X)$ are small enough, then we will have $\mathbb{E}\|X\| \lesssim \sigma(X)$, which improves the Matrix Bernstein inequality by removing the $\sqrt{\log d}$ factor.

Spectrum of Gaussian Matrices. To leverage the universality properties of the random matrix model, we shall rely on the following result about the singular values of Gaussian matrices, which in particular can be used to recover the optimal subspace embedding guarantee for Gaussian sketches.

Lemma 2.11 ((2.3), [34]). *Let G be an $m \times n$ matrix whose entries are independent standard normal variables. Then,*

$$\mathbb{P}(\sqrt{m} - \sqrt{n} - t \leq s_{\min}(G) \leq s_{\max}(G) \leq \sqrt{m} + \sqrt{n} + t) \geq 1 - 2e^{-t^2/2}$$

3 ANALYSIS OF OBLIVIOUS SPARSE EMBEDDINGS

In this section, we state and provide a sketch of the proof of our main OSE result, Theorem 3.3 (given as Theorem 1.2 in Section 1). Before we get to the proof, however, we propose a new model for sparse OSEs that is designed to exploit the strength of our proof in dealing with the number of independent random bits required.

To illustrate the issue, consider that in the OSE-IID-ENT model (Definiton 2.1), we need mn many random bits to determine the nonzero entries of the matrix S, even though the matrix will have far fewer non-zero entries. Naturally, there are many known strategies for improving this, including OSNAP-IND-COL (Definition 2.3) and even more elaborate hashing constructions based on polynomials over finite fields [31], which allow reducing the random bit complexity to polylogarithmic in the dimensions. However, these constructions do not provide sufficient independence needed in the random matrix model (2.2) to apply the universality result of Brailovskaya and Van Handel (Lemma 2.10). We address this with the *independent diagonals* distribution family, defined shortly.

In order to use universality for establishing a subspace embedding guarantee, we must analyze a symmetrized version of the matrix SU, for an $n \times d$ orthogonal matrix U, with S being a sum of sparse independent random matrices, say Y_i's. Naturally, to compare the spectra of SU and an appropriate Gaussian model, the matrix S cannot be too sparse. However, since the individual entries of each Y_i need not be independent, we can reduce the number of independent summands in S by making each individual summand Y_i denser. At the same time, we need to control $\|Y_iU\|$, just as we would when using the standard matrix Bernstein inequality.

Both these goals can be achieved by placing non-zero entries along a diagonal of Y_i. Placing ± 1 entries along a diagonal of a matrix keeps its norm bounded by 1 whereas other arrangements (say, along a row or column) do not. Moreover, these ± 1 entries along the diagonal need not be independent, they can simply be uncorrelated. As a result, an instance of Y_i can be generated with just $O(1)$ random bits.

3.1 Independent Diagonals Construction

With this motivation, we define the independent diagonals distribution formally (Figure 1 illustrates this construction).

Definition 3.1 (OSE-IND-DIAG)**.** An $m \times n$ random matrix S is called an oblivious subspace embedding with independent diagonals (OSE-IND-DIAG) with parameter p if it is constructed in the following way. Assume that np is an integer. Let $W = (w_1, \ldots, w_m)$ be a random vector whose components are ± 1 valued and uncorrelated, i.e. $\mathbb{E}[w_i] = 0, \mathbb{E}[w_i^2] = 1, \mathbb{E}[w_iw_j] = 0$. We define γ to be a random variable uniformly distributed in $[n]$. Let $\gamma_1, \ldots, \gamma_{np}$ be i.i.d. copies of γ. Let $W_1, \ldots, W_{np}$ be i.i.d. copies of W. Let $F_j(x)$ be a function that transforms a m dimensional vector x to the $m \times n$ matrix putting x on the j^{th} diagonal, i.e. positions $(1, j)$ through $(m, j+m \bmod n)$ with all other entries zero (See Fig 1 for an illustration). Let $S = \sum_{l\in[np]} F_{\gamma_l}(W_l)$.

Universality results show that the properties of a general random matrix are similar to the properties of a gaussian random matrix with the same covariance profile. Therefore, to analyze the OSE models, we need to first calculate the covariances between entries.

Lemma 3.2 (Variance and uncorrelatedness). *Let $p = p_{m,n} \in (0,1]$ and $S = \{s_{ij}\}_{i\in[m],j\in[n]}$ be a $m \times n$ random matrix distributed according to the OSE-IID-ENT, OSNAP-IND-COL, or OSE-IND-DIAG distributions. Then, $\mathbb{E}(s_{ij}) = 0$ and $\mathrm{Var}(s_{ij}) = p$ for all $i \in [m], j \in [n]$, and $\mathrm{Cov}(s_{i_1j_1}, s_{i_2j_2}) = 0$ for any $\{i_1, i_2\} \subset [m], \{j_1, j_2\} \subset [n]$ and $(i_1, j_1) \neq (i_2, j_2)$*

Proof. (see the full version) □

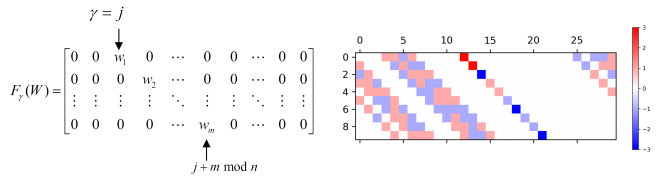

Figure 1: Left: Structure of the random matrix $F_{\gamma_l}(W_l)$. Right: Illustration of an embedding matrix S with independent diagonals with parameters $d = 8$, $m = 10$, and $n = 30$, showing how the nonzero entries of S occur along diagonals. The number of diagonals is controlled by the parameter p.

3.2 Proof of Theorem 1.2

We now state the main theorem of this section, which is the detailed version of Theorem 1.2.

Theorem 3.3 (Analysis of OSE by Universality). *Let S be an $m \times n$ matrix distributed according to the OSE-IID-ENT, OSNAP-IND-COL, or OSE-IND-DIAG distributions with parameter p. Let U be an arbitrary $n \times d$ deterministic matrix such that $U^T U = I$. Then, there exist constants $c_{3.3.1}$ and $c_{3.3.2}$ such that for any $\varepsilon, \delta > 0$, we have*

$$\mathbb{P}\left(1-\varepsilon \le s_{\min}((1/\sqrt{pm})SU) \le s_{\max}((1/\sqrt{pm})SU) \le 1+\varepsilon\right) \ge 1-\delta$$

when $m \ge c_{3.3.1}\max(d, \log(4/\delta))/\varepsilon^2$ and $pm > c_{3.3.2}(\log(d/\delta))^4/\varepsilon^6$.

Alternatively, given a fixed $\theta < 3$, there exist constants $c_{3.3.3}$, $c_{3.3.4}$ and $c_{3.3.5}$ such that for $m \ge \max\{(1+\theta)d, c_{3.3.3}\log(4/\delta))/\theta^2\}$ and $pm > c_{3.3.4}(\log(d/\delta))^4/\theta^6$,

$$\mathbb{P}\left(\kappa\left(\frac{1}{\sqrt{pm}}SU\right) \le \frac{c_{3.3.5}}{\theta}\right) \ge 1-\delta$$

The proof of Theorem 3.3 follows by applying the universality result 2.10 to an augmented and symmetrized version of SU. More precisely, following [2, p.443], for a $m \times d$ matrix Y, we define the augmented and symmetrized version of Y as

$$\mathrm{augsym}(Y,\lambda) = \begin{bmatrix} 0_{d\times d} & 0_{d\times m} & Y^T & \mathrm{aug}(Y,\lambda)^{1/2} \\ 0_{m\times d} & 0_{m\times m} & 0_{m\times m} & 0_{m\times d} \\ Y & 0_{m\times m} & 0_{m\times m} & 0_{m\times d} \\ \mathrm{aug}(Y,\lambda)^{1/2} & 0_{d\times m} & 0_{d\times m} & 0_{d\times d} \end{bmatrix}$$

where

$$\mathrm{aug}(Y,\lambda) = (\|\mathbb{E}\, Y^T Y\| + 4\lambda^2)\cdot \mathrm{Id} - \mathbb{E}\, Y^T Y$$

Then we set $X_\lambda = \mathrm{augsym}(Y,\lambda)$.

There are two reasons for using the matrix $\mathrm{augsym}(SU,\lambda)$. First, using Lemma 3.2 requires symmetric matrices, so we need to symmetrize the matrix SU. Second, after symmetrization, we obtain the matrix

$$\begin{bmatrix} & (SU)^T \\ SU & \end{bmatrix}$$

and the spectrum of this matrix the the union of $\mathrm{spec}(SU)$ and $\{0\}$. By universality results, we can only claim that $\mathrm{spec}(SU) \cup \{0\}$ is close to $\mathrm{spec}(\sqrt{p}G) \cup \{0\}$, which does not directly imply the desired result that $s_{\min}(SU)$ is close to $s_{\min}(\sqrt{p}G)$. Therefore, we need to introduce look at the perturbed matrix $\mathrm{aug}(SU,\lambda)$ to show that $s_{\min}(SU)$ is not close to zero.

To use Lemma 2.10 we need to find out the corresponding guassian model and bound the parameters $\sigma(X_\lambda), \sigma_*(X_\lambda)$ and $R(X_\lambda)$ defined in Lemma 2.10.

By Lemma 3.2, we know that, in all of OSE-IID-ENT, OSNAP-IND-COL, and OSE-IND-DIAG distributions, each entry of SU has variance p and different entries have zero covariances. Therefore, we know that the corresponding gaussian model for X_λ is $\mathrm{augsym}(\sqrt{p}G,\lambda)$.

By the covariance structure of SU, we have $\mathbb{E}[U^T S^T SU] = pm\cdot\mathrm{Id}$, and therefore

$$\begin{aligned}\mathrm{aug}(SU,\lambda) =& (\|\mathbb{E}(SU)^T(SU)\| + 4\lambda^2)\cdot\mathrm{Id} - \mathbb{E}(SU)^T(SU)\\ =& (pm + 4\lambda^2)\cdot\mathrm{Id} - pm\cdot\mathrm{Id}\\ =& 4\lambda^2\cdot\mathrm{Id}\end{aligned}$$

Similarly, we also have $\mathrm{aug}(\sqrt{p}G,\lambda) = 4\lambda^2\cdot\mathrm{Id}$.

We observe that $\sigma(X_\lambda)$ and $\sigma_*(X_\lambda)$ do not depend on the decomposition of X_λ as a sum of independent random matrices and can be calculated explicitly using the covariance structure of X_λ. Using this idea, we derive the following lemma that bounds $\sigma(X_\lambda)$ and $\sigma_*(X_\lambda)$.

Lemma 3.4 (Covariance Parameters). *Let $S = \{s_{ij}\}_{i\in[m],j\in[n]}$ be a $m\times n$ random matrix such that $\mathbb{E}(s_{ij}) = 0$ and $\mathrm{Var}(s_{ij}) = p$ for all $i \in [m]$, $j \in [n]$, and $\mathrm{Cov}(s_{ij}, s_{kl}) = 0$ for any $\{i,k\} \subset [m]$, $\{j,l\} \subset [n]$ and $(i,j) \ne (k,l)$. Let $\sigma^* : L_\infty(\mathbb{R}) \otimes M_{2(m+d)}(\mathbb{R}) \to \mathbb{R}$ and $\sigma : L_\infty(\mathbb{R}) \otimes M_{2(m+d)}(\mathbb{R}) \to \mathbb{R}$ be the functions defined in Lemma 2.10. Then for any $\lambda > 0$, we have*

$$\sigma_*(\mathrm{augsym}(SU,\lambda)) \le 2\sqrt{p} \text{ and } \sigma(\mathrm{augsym}(SU,\lambda)) \le \sqrt{pm}$$

Proof. (see the full version) □

Proof of Theorem 3.3 (Sketch). Using Lemma 3.4 and Lemma 3.2, we have

$$\sigma_*(\mathrm{augsym}(SU,\lambda)) \le 2\sqrt{p} \text{ and } \sigma(\mathrm{augsym}(SU,\lambda)) \le \sqrt{pm}$$

for all the three distributions.

$R(X_\lambda)$ depends on the decomposition of X_λ as a sum of independent random matrices, so we write the matrix SU as a sum of independent random matrices in each of the three distributions as follows.

For the OSE-IID-ENT distribution, we observe that

$$SU = \sum_{i,j} s_{i,j}(e_i e_j^T)U = \sum_{i,j} s_{i,j}(e_i u_j^T)$$

where u_j is the jth row vector of the matrix U.

For the OSE-IND-DIAG distribution, we have $SU = \sum_{l=1}^{pn} Y_l U$, where $Y_l = F_{\gamma_l}(W_l)$ as in Definition 3.1.

For the OSNAP-IND-COL distribution, the sum is similar to the OSE-IND-DIAG case. More precisely, for $k \in [s], l \in [n]$, we define $Y_{k,l}$ to be the $m \times n$ matrix such that

$$(Y_{k,l})_{i,j} = \mathbf{1}_{[(m/s)(i-1)+1:(m/s)i]\times\{j\}}((i,j))S_{i,j}$$

In conclusion, we have $R(\mathrm{augsym}(SU,\lambda)) \le 1$ for all the three distributions. (See the full version for details.)

Using lemma 2.10 with $t = \log(2d/\delta)$, we have

$$\mathbb{P}(d_H(\text{spec}(\text{augsym}(SU)), \text{spec}(\text{augsym}(\sqrt{p}G))) > c_1\zeta(\log(2d/\delta))) \le \delta/2$$

where

$$\zeta(t) = \sigma_*(X_\lambda)t^{\frac{1}{2}} + R(X_\lambda)^{\frac{1}{3}}\sigma(X_\lambda)^{\frac{2}{3}}t^{\frac{2}{3}} + R(X_\lambda)t$$

for some constant c_1. Without loss of generality, assume $c_1 > 1$.

Using lemma 2.11, we have

$$\mathbb{P}\left(\sqrt{pm} - \sqrt{pd} - \sqrt{2p\log(4/\delta)} \le s_{\min}(\sqrt{p}G) \le s_{\max}(\sqrt{p}G) \le \sqrt{pm} + \sqrt{pd} + \sqrt{2p\log(4/\delta)}\right) \ge 1 - \delta/2$$

Let $\mathcal{E}$ be the event

$$\mathcal{E} = \{\sqrt{pm} - \sqrt{pd} - \sqrt{2p\log(4/\delta)} \le s_{\min}(\sqrt{p}G) \le s_{\max} \le (\sqrt{p}G) \le \sqrt{pm} + \sqrt{pd} + \sqrt{2p\log(4/\delta)}\} \cap \{d_H(\text{spec}(\text{augsym}(SU, \lambda)), \text{spec}(\text{augsym}(\sqrt{p}G, \lambda))) \le c_1\zeta(\log(2d/\delta))\}$$

Then, we have $\mathbb{P}(\mathcal{E}) \ge 1 - \delta$ by the union bound. Assume that the event $\mathcal{E}$ happens, then $\mathcal{E}$ implies that the spectrum of $\text{augsym}(SU, \lambda)$ and $\text{augsym}(\sqrt{p}G, \lambda)$ are close in the Hausdorff distance and also implies bounds on the extreme singular values of $\sqrt{p}G$.

Using the relationship between singular values of SU (resp. $\sqrt{p}G$) and $\text{augsym}(SU, \lambda)$ (resp. $\text{augsym}(\sqrt{p}G, \lambda)$) (see full version for details), we have,

$$\sqrt{pm} - \sqrt{pd} - \sqrt{2p\log(4/\delta)} - 5\lambda \le s_{\min}(SU) \le s_{\max}(SU) \le \sqrt{pm} + \sqrt{pd} + \sqrt{2p\log(4/\delta)} + 5\lambda$$

Therefore, we derive that

$$\mathbb{P}\left(\sqrt{pm} - \sqrt{pd} - \sqrt{2p\log(4/\delta)} - 5\lambda \le s_{\min}(SU) \le s_{\max}(SU) \le \sqrt{pm} + \sqrt{pd} + \sqrt{2p\log(4/\delta)} + 5\lambda\right) \ge 1 - \delta$$

Now, we choose $\lambda = \frac{1}{10}\varepsilon\sqrt{pm}$. This is possible when $\frac{1}{10}\epsilon\sqrt{pm} \ge c_1\zeta(\log(2d/\delta))$, and we will simplify this condition later.

Assuming that we can choose $\lambda = \frac{1}{10}\varepsilon\sqrt{pm}$ and

$$m > \max\{\frac{16d}{\varepsilon^2}, \frac{32\log(4/\delta)}{\varepsilon^2}\}$$

we have $\frac{\sqrt{2p\log(4/\delta)}}{\sqrt{pm}} < \frac{\varepsilon}{4}$ and $\sqrt{\frac{d}{m}} < \frac{\varepsilon}{4}$, and therefore we have

$$\mathbb{P}\left(1 - \frac{\varepsilon}{4} - \frac{\varepsilon}{4} - \frac{\varepsilon}{2} \le s_{\min}((1/\sqrt{pm})SU) \le s_{\max}((1/\sqrt{pm})SU) \le 1 + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{2}\right) \ge 1 - \delta$$

which is exactly what we want.

Then it suffices to translate the condition

$$\frac{1}{10}\epsilon\sqrt{pm} \ge c_1\zeta(\log(2d/\delta))$$

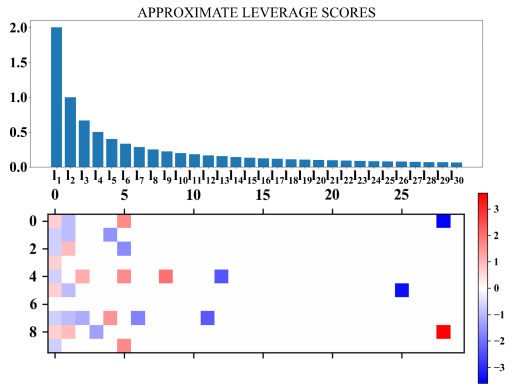

Figure 2: LESS-IND-ENT with decreasing leverage scores. Since the probability of an entry being non-zero is proportional to the corresponding leverage score, we see that the matrix becomes sparser as we move in the direction of decreasing leverage scores. Since the scaling of entries is inversely proportional to the square root of the corresponding leverage score, the magnitude of the non-zero entries becomes larger as we move to the right.

into the requirements for p, m, and d. To this end, we first calculate that

$$\zeta(t) = 2\sqrt{p}t^{1/2} + (\sqrt{pm})^{2/3}t^{2/3} + t$$

by the earlier bounds for $\sigma(X)$, $\sigma_*(X)$ and $R(X)$.

We claim that it is enough to require that,

$$(\sqrt{pm})^{2/3}(\log(2d/\delta))^{2/3} \le \frac{\varepsilon}{80c_1}\sqrt{pm}$$

Equivalently, we just need

$$\frac{(80c_1)^6(\log(2d/\delta))^4}{\varepsilon^6} \le pm$$

In fact, if $pm \ge c_2\frac{(\log(2d/\delta))^4}{\varepsilon^6}$ where $c_2 = (80c_1)^6$, we will also have $\log(2d/\delta) \le \frac{\varepsilon}{80c_1}\sqrt{pm}$ and $2(\log(2d/\delta))^{1/2} \le \frac{\varepsilon}{40c_1}\sqrt{pm}$, which gives us $c_1\zeta(\log(2d/\delta)) \le \frac{\varepsilon}{10}\sqrt{pm}$, and therefore we have proved the first part of the theorem.

The proof the second part follows from similar calculation (see the full version). □

4 ANALYSIS OF NON-OBLIVIOUS SPARSE EMBEDDINGS

In the non-oblivious setting, we are looking to embed a specific d-dimensional subspace of $\mathbb{R}^n$ represented by a matrix U, and we assume that we have access to (β_1, β_2)-approximate leverage scores of U as defined in Section 2. Given access to this information, we can modify the oblivious models of S to give more weight to certain coordinates of the ambient space. We refer to this approach as Leverage Score Sparsitication (LESS).

4.1 Leverage Score Sparsification

We propose two variants of a LESS embedding. First, we consider an extension of the OSE-IND-ENT model (with i.i.d. entries) studied

in Section 3, with the Bernoulli sparsifier for each entry of S being non-zero with a probability proportional to the leverage score of the corresponding component. However, we still want the variance of each entry of S to be p, so the entries are appropriately scaled copies of ± 1 (See Figure 2).

Definition 4.1 (LESS-IND-ENT)**.** An $m \times n$ random matrix S is called a leverage score sparsified embedding with independent entries (LESS-IND-ENT) corresponding to (β_1, β_2)-approximate leverage scores $(l_1, ..., l_n)$ with parameter p, if S has entries $s_{i,j} = \frac{1}{\sqrt{\beta_1 l_j}}\delta_{i,j}\xi_{i,j}$ where $\delta_{i,j}$ are independent Bernoulli random variables taking value 1 with probability $p_{ij} = \beta_1 l_j p$ and $\xi_{i,j}$ are i.i.d. random variables with $\mathbb{P}(\xi_{i,j} = 1) = \mathbb{P}(\xi_{i,j} = -1) = 1/2$.

In the next variant of a LESS embedding, we are able to reduce the computational and random bit complexity for generating sparsity by only generating as many non-zero entries as required. Here, the i^{th} row of S is the sum of np many i.i.d. random matrices Z_{ij} where each Z_{ij} is determined by choosing one entry from the n possible entries of the i^{th} row and setting the remaining entries to 0. Here, instead of choosing the positions uniformly at random, we choose them proportionally to the corresponding leverage score.

Definition 4.2 (LESS-IND-ROWS)**.** Assume that $(\beta_1 p \sum l_i)$ is an integer. An $m \times n$ random matrix S is called a leverage score sparsified embedding with independent rows (LESS-IND-ROWS) corresponding to (β_1, β_2)-approximate leverage scores $(l_1, ..., l_n)$ with parameter p, if the i^{th} row of S is a sum of $(\beta_1 p \sum l_j)$ i.i.d. copies $Z_{i1}, Z_{i2}, ...$, of a random variable Z_i, i.e.,

$$S = \sum_{i=1}^{m} \sum_{k=1}^{(\beta_1 p \sum l_i)} Z_{ik},$$

where Z_i is defined as follows. Let γ be a random variable taking values in $[n]$ with $\mathbb{P}(\gamma = j) = l_j/(\sum_{k=1}^{n} l_k)$. Let ξ be a Rademacher random variable, $\mathbb{P}(\xi = -1) = \mathbb{P}(\xi = 1) = \frac{1}{2}$. Then,

$$Z_i = \xi \sum_{j \in [n]} \mathbf{1}_{\{j\}}(\gamma) \frac{1}{\sqrt{\beta_1 l_j}} E_{i,j}$$

where $E_{i,j}$ is an $m \times n$ matrix with 1 in the $(i,j)^{th}$ entry and 0 everywhere else.

4.2 Proof of Theorem 1.3

These modifications allow us to prove subspace embedding guarantees for sparser matrices than in the oblivious case, thereby showing Theorem 1.3. In particular, we show that with LESS it suffices to use $O(\log^4(d/\delta))$ nonzero entries per row of S, instead of per column of S, which is much sparser since S is a wide matrix.

Theorem 4.3. *Let U be an arbitrary $n \times d$ deterministic matrix such that $U^T U = I$ with (β_1, β_2)-approximate leverage scores $(l_1, ..., l_n)$. There exist constants $c_{4.3.1}, c_{4.3.2}, c_{4.3.3}$, such that for any $0 < \varepsilon < 1, 0 < \delta < 1$, and any LESS-IND-ENT or LESS-IND-ROWS random matrix S corresponding to $(l_1, ..., l_n)$ with embedding dimension $m \geq c_{4.3.1} \max(d, \log(4/\delta))/\varepsilon^2$ and parameter $p \geq c_{4.3.2}(\log(d/\delta))^4/(m\varepsilon^6)$, we have*

$$\mathbb{P}\left(1 - \varepsilon \leq s_{\min}((1/\sqrt{pm})SU) \leq s_{\max}((1/\sqrt{pm})SU) \leq 1 + \varepsilon\right) \geq 1 - \delta,$$

and if we choose $m = c_{4.3.1} \max(d, \log(4/\delta))/\varepsilon^2$ and

$$p = c_{4.3.2}(\log(d/\delta))^4/(m\varepsilon^6),$$

then we have the following high probability bound for the maximum number of nonzero entries per row

$$\mathbb{P}\left(\max_{i \in [m]} (\text{card}(\{j \in [n] : s_{ij} \neq 0\})) \leq c_{4.3.3}\beta_1\beta_2(\log(d/\delta))^4/\varepsilon^4\right) \geq 1 - \delta$$

Alternatively, given a fixed $\theta < 3$, there exist constants $c_{4.3.4}$, $c_{4.3.5}$, $c_{4.3.6}$ and $c_{4.3.7}$ such that for $m \geq \max\{(1+\theta)d, c_{4.3.4}\log(4/\delta))/\theta^2\}$ and $p > c_{4.3.5}(\log(d/\delta))^4/(m\theta^6)$,

$$\mathbb{P}\left(\kappa(\frac{1}{\sqrt{pm}}SU) \leq \frac{c_{4.3.6}}{\theta}\right) \geq 1 - \delta$$

and if we choose $m = \max\{(1+\theta)d, c_{4.3.4}\log(4/\delta))/\theta^2\}$ and $p = c_{4.3.5}(\log(d/\delta))^4/(m\theta^6)$, then we have the following high probability bound for the maximum number of nonzero entries per row

$$\mathbb{P}\left(\max_{i \in [m]} (\text{card}(\{j \in [n] : s_{ij} \neq 0\})) \leq c_{4.3.7}\beta_1\beta_2(\log(d/\delta))^4/\theta^6\right) \geq 1 - \delta$$

Proof of Theorem 4.3 (sketch). Similarly to Lemma 3.4, we can conclude that $\sigma_*(X) \leq 2\sqrt{p}$ and $\sigma(\text{augsym}(SU)) \leq \sqrt{pm}$.

For the IND-ENT case, since $|s_{i,j}|$ is bounded by $\frac{1}{\sqrt{\beta_1 l_j}}$, we have

$$\begin{aligned} &R(\text{augsym}(SU)) \\ &\leq \max_{i,j} \frac{1}{\sqrt{\beta_1 l_j}} \left\| \begin{bmatrix} 0 & 0 & (e_i u_j{}^T)^T & 0 \\ 0 & 0 & 0 & 0 \\ (e_i u_j{}^T) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right\|_{op} \\ &\leq 1 \end{aligned}$$

And for the IND-POS case, we have

$$Z_k U = \sum_{i \in [m], j \in [n]} (Z_k)_{i,j}(e_i u_j{}^T)$$

Since this sum has only one nonzero term, we have

$$\|Z_k U\|_{op} = \max_{i \in [m], j \in [n]} |(Z_k)_{i,j}| \|e_i u_j{}^T\|_{op}$$

Using this decomposition to augsym(SU) as well as the fact that $|(Z_k)_{i,j}| \leq \frac{1}{\sqrt{\beta_1 l_j}}$, we conclude

$$\begin{aligned} &R(\text{augsym}(SU)) \\ &\leq \max_{\substack{i \in [m], j \in [n], \\ k \in [\beta_1 pm \sum l_i]}} (Z_k)_{i,j} \left\| \begin{bmatrix} 0 & 0 & (e_i u_j{}^T)^T & 0 \\ 0 & 0 & 0 & 0 \\ (e_i u_j{}^T) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right\|_{op} \\ &\leq 1 \end{aligned}$$

For both cases, since their matrix parameters $\sigma_*(X), \sigma(X)$ and $R(X)$ are the same as in Theorem 3.3, following the proof for Theorem 3.3, we conclude that there exist constants $c_1, c_2,$, such that for any $\varepsilon, \delta > 0$, we have

$$\mathbb{P}\left(1-\varepsilon \le s_{\min}((1/\sqrt{pm})SU) \le s_{\max}((1/\sqrt{pm})SU) \le 1+\varepsilon\right) \ge 1-\delta$$

when $m \ge c_1 \max(d, \log(4/\delta))/\varepsilon^2$ and $pm > c_2(\log(d/\delta))^4/\varepsilon^6$. The proof of the claim when $m \ge (1+\theta)d$ also follows in the same manner as Theorem 3.3.

Since p is just a parameter in the LESS models, we want to know what the requirement $pm > c_2(\log(d/\delta))^4/\varepsilon^6$ (similarly $pm > c_2(\log(d/\delta))^4/\theta^6$) means for the average number of nonzero entries in each row. First, the average number of nonzero entries in each row will be $\beta_1 p \sum_{j=1}^n l_j \le \beta_1\beta_2 pd$ for both cases, just by the construction of these matrices. Next, we observe that the condition $pm > c_2(\log(d/\delta))^4/\varepsilon^6$ is equivalent to

$$pd > c_2(\log(d/\delta))^4/\varepsilon^6(\frac{d}{m})$$

Since $m \ge d/\varepsilon^2$, we have $\frac{d}{m} \le \varepsilon^2$. Therefore, to meet the requirement $pd > c_2(\log(d/\delta))^4/\varepsilon^6\frac{d}{m}$, it suffices to have

$$pd > c_2(\log(d/\delta))^4/\varepsilon^4$$

So the optimal choice of p leads to $\beta_1\beta_2 c_2(\log(d/\delta))^4/\varepsilon^4$ many nonzero entries in each row on average.

In the case when $m \ge (1+\theta)d$ and $pm > c_2(\log(d/\delta))^4/\theta^6$, since we can only claim $d/m < 1$ in general, we need $pd > c_2(\log(d/\delta))^4/\theta^6$ so the average number of nonzero entries in each row would be $\beta_1\beta_2 c_2(\log(d/\delta))^4/\theta^6$.

The high probability bound for the number of nonzero entries in each row follows from Bernstein's Inequality. (see full version for details.) □

As an immediate corollary of Theorem 4.3 we give a new subspace embedding guarantee for the Fast Johnson-Lindenstrauss Transform (FJLT). Recall that an FJLT preconditions the matrix U with the Randomized Hadamard Transform (see Definition 2.7) to transform it into another matrix V whose row norms can be well controlled. In this way, we obtain approximate leverage scores for V by construction rather than by estimation. Then, we can apply LESS-IND-ENT or LESS-IND-ROWS random matrices to the preconditioned matrix V according to approximate leverage scores.

Theorem 4.4 (Analysis of Preconditioned Sparse OSE). *Let U be an arbitrary $n\times d$ deterministic matrix such that $U^TU = I$. There exist constants $c_{4.4.1}, c_{4.4.2}, c_{4.4.3}$, such that for any $0 < \varepsilon < 1$, $\frac{2n}{e^d} < \delta < 1$, the following holds. Let $S = \Phi(\frac{1}{\sqrt{n}})HD$ where H and D are as in definition 2.6 and 2.7, and Φ has LESS-IND-ENT or LESS-IND-ROWS distribution corresponding to uniform leverage scores $(d/n, ..., d/n)$ with embedding dimension $m \ge c_{4.4.1}\max(d, \log(8/\delta))/\varepsilon^2$ and average number of nonzero entries per row $\ge c_{4.4.2}(\log(2d/\delta))^4/\varepsilon^4$. Then,*

$$\mathbb{P}\left(1-\varepsilon \le s_{\min}((1/\sqrt{pm})SU) \le s_{\max}((1/\sqrt{pm})SU) \le 1+\varepsilon\right) \ge 1-\delta.$$

Proof of Theorem 4.4 (sketch). First, note that $\frac{1}{\sqrt{n}}HDU$ is an $n \times d$ matrix with orthonormal columns. Let $\mathcal{E}$ denote the event that the tuple $(l_1 = d/n, l_2 = d/n, \ldots, l_n = d/n)$ of numbers are $(16,1)$-approximate leverage scores for the matrix $\frac{1}{\sqrt{n}}HDU$. Clearly, $\sum_{i=1}^n l_i \le d$. Since $2n \le \delta e^d$, we have $\log(2n/\delta) \le d$, and the claim from Lemma 2.8 reads,

$$\mathbb{P}\left(\max_{j=1,\ldots,n}\|e_j^T(\frac{1}{\sqrt{n}}HDU)\|_{op} \ge \sqrt{\frac{d}{n}} + \sqrt{\frac{8d}{n}}\right) \le \frac{\delta}{2}.$$

Thus, with probability greater than $1-\delta/2$, we have, for all $j \in [n]$,

$$\|e_j^T(\frac{1}{\sqrt{n}}HDU)\|^2 < (1+2\sqrt{2})^2\left(\frac{d}{n}\right) < \frac{16d}{n} = 16l_j.$$

So, the conditions for $(d/n, d/n, \ldots, d/n)$ to be $(16,1)$-approximate leverage scores for the matrix $\frac{1}{\sqrt{n}}HDU$ are satisfied with probability greater than $1-\delta/2$, i.e.,

$$\mathbb{P}(\mathcal{E}) \ge 1-\delta/2.$$

Let $V = \frac{1}{\sqrt{n}}HDU$. Then the desired result follows by conditioning on the random matrix V and applying Theorem 4.3. □

ACKNOWLEDGMENTS

This work was partially supported by DMS 2054408.

REFERENCES

[1] Nir Ailon and Bernard Chazelle. 2009. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing* 39, 1 (2009), 302–322.

[2] Afonso S Bandeira, March T Boedihardjo, and Ramon van Handel. 2023. Matrix concentration inequalities and free probability. *Inventiones mathematicae* (2023), 1–69.

[3] Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. 2015. Toward a unified theory of sparse dimensionality reduction in euclidean space. In *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*. 499–508.

[4] Tatiana Brailovskaya and Ramon van Handel. 2022. Universality and sharp matrix concentration inequalities. *arXiv preprint arXiv:2201.05142* (2022).

[5] Coralia Cartis, Jan Fiala, and Zhen Shao. 2021. Hashing embeddings of optimal dimension, with applications to linear least squares. *arXiv preprint arXiv:2105.11815* (2021).

[6] Xue Chen and Michal Dereziński. 2021. Query complexity of least absolute deviation regression via robust uniform convergence. In *Conference on Learning Theory*. PMLR, 1144–1179.

[7] Nadiia Chepurko, Kenneth L Clarkson, Praneeth Kacham, and David P Woodruff. 2022. Near-optimal algorithms for linear algebra in the current matrix multiplication time. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 3043–3068.

[8] Yeshwanth Cherapanamjeri, Sandeep Silwal, David P Woodruff, and Samson Zhou. 2023. Optimal algorithms for linear algebra in the current matrix multiplication time. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. SIAM, 4026–4049.

[9] Kenneth L Clarkson, Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, Xiangrui Meng, and David P Woodruff. 2016. The fast cauchy transform and faster robust linear regression. *SIAM J. Comput.* 45, 3 (2016), 763–810.

[10] Kenneth L Clarkson and David P Woodruff. 2009. Numerical linear algebra in the streaming model. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*. 205–214.

[11] Kenneth L Clarkson and David P Woodruff. 2013. Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of Computing*. 81–90.

[12] Michael B Cohen. 2016. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proceedings of the twenty-seventh annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 278–287.

[13] Michael B Cohen, Yin Tat Lee, and Zhao Song. 2021. Solving linear programs in the current matrix multiplication time. *Journal of the ACM (JACM)* 68, 1 (2021), 1–39.

[14] Michael B Cohen, Cameron Musco, and Christopher Musco. 2017. Input sparsity time low-rank approximation via ridge leverage score sampling. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 1758–1777.

[15] Michael B Cohen and Richard Peng. 2015. Lp row sampling by lewis weights. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 183–192.

[16] Michał Dereziński. 2023. Algorithmic gaussianization through sketching: Converting data into sub-gaussian random designs. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 3137–3172.
[17] Michał Dereziński, Jonathan Lacotte, Mert Pilanci, and Michael W Mahoney. 2021. Newton-LESS: Sparsification without Trade-offs for the Sketched Newton Update. *Advances in Neural Information Processing Systems* 34 (2021), 2835–2847.
[18] Michał Dereziński, Zhenyu Liao, Edgar Dobriban, and Michael Mahoney. 2021. Sparse sketches with small inversion bias. In *Conference on Learning Theory*. PMLR, 1467–1510.
[19] Michał Derezinski and Michael W Mahoney. 2021. Determinantal point processes in randomized numerical linear algebra. *Notices of the American Mathematical Society* 68, 1 (2021), 34–45.
[20] Michał Dereziński and Elizaveta Rebrova. 2024. Sharp analysis of sketch-and-project methods via a connection to randomized singular value decomposition. *SIAM Journal on Mathematics of Data Science* 6, 1 (2024), 127–153.
[21] Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. 2012. Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research* 13, 1 (2012), 3475–3506.
[22] Petros Drineas and Michael W Mahoney. 2016. RandNLA: randomized numerical linear algebra. *Commun. ACM* 59, 6 (2016), 80–90.
[23] Petros Drineas, Michael W Mahoney, and S Muthukrishnan. 2006. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*. 1127–1136.
[24] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review* 53, 2 (2011), 217–288.
[25] Haotian Jiang, Tarun Kathuria, Yin Tat Lee, Swati Padmanabhan, and Zhao Song. 2020. A faster interior point method for semidefinite programming. In *2020 IEEE 61st annual symposium on foundations of computer science (FOCS)*. IEEE, 910–918.
[26] W B Johnson and J Lindenstrauss. 1984. Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.* 26 (1984), 189–206.
[27] Yi Li and David Woodruff. 2020. Input-sparsity low rank approximation in Schatten norm. In *International Conference on Machine Learning*. PMLR, 6001–6009.
[28] Per-Gunnar Martinsson and Joel A Tropp. 2020. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica* 29 (2020), 403–572.
[29] Xiangrui Meng and Michael W Mahoney. 2013. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. 91–100.
[30] R. Murray, J. Demmel, M. W. Mahoney, N. B. Erichson, M. Melnichenko, O. A. Malik, L. Grigori, M. Dereziński, M. E. Lopes, T. Liang, and H. Luo. 2023. *Randomized Numerical Linear Algebra – A Perspective on the Field with an Eye to Software*. Technical Report arXiv preprint arXiv:2302.11474.
[31] Jelani Nelson and Huy L Nguyên. 2013. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 ieee 54th annual symposium on foundations of computer science*. IEEE, 117–126.
[32] Jelani Nelson and Huy L Nguyên. 2014. Lower bounds for oblivious subspace embeddings. In *International Colloquium on Automata, Languages, and Programming*. Springer, 883–894.
[33] Vladimir Rokhlin and Mark Tygert. 2008. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences* 105, 36 (2008), 13212–13217.
[34] Mark Rudelson and Roman Vershynin. 2010. Non-asymptotic theory of random matrices: extreme singular values. In *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*. World Scientific, 1576–1602.
[35] Tamas Sarlos. 2006. Improved approximation algorithms for large matrices via random projections. In *2006 47th annual IEEE symposium on foundations of computer science (FOCS'06)*. IEEE, 143–152.
[36] Zhao Song, David P Woodruff, and Peilin Zhong. 2019. Relative error tensor low rank approximation. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 2772–2789.
[37] Joel A Tropp. 2011. Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis* 3, 01n02 (2011), 115–126.
[38] Ruosong Wang and David P Woodruff. 2022. Tight Bounds for l1 Oblivious Subspace Embeddings. *ACM Transactions on Algorithms (TALG)* 18, 1 (2022), 1–32.
[39] David P Woodruff. 2014. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science* 10, 1–2 (2014), 1–157.
[40] Jiyan Yang, Yin-Lam Chow, Christopher Ré, and Michael W Mahoney. 2017. Weighted SGD for lp regression with randomized preconditioning. *The Journal of Machine Learning Research* 18, 1 (2017), 7811–7853.

Received 10-NOV-2023; accepted 2024-02-11