Optimal Oblivious Subspace Embeddings with Near-Optimal Sparsity

Shabarish Chenakkod
University of Michigan, Ann Arbor, MI, USA

Michał Dereziński
University of Michigan, Ann Arbor, MI, USA

Xiaoyu Dong
National University of Singapore, Singapore

Abstract

An oblivious subspace embedding is a random $m \times n$ matrix Π such that, for any d-dimensional subspace, with high probability Π preserves the norms of all vectors in that subspace within a $1 \pm \epsilon$ factor. In this work, we give an oblivious subspace embedding with the optimal dimension $m = \Theta(d/\epsilon^2)$ that has a near-optimal sparsity of $\tilde{O}(1/\epsilon)$ non-zero entries per column of Π. This is the first result to nearly match the conjecture of Nelson and Nguyen [FOCS 2013] in terms of the best sparsity attainable by an optimal oblivious subspace embedding, improving on a prior bound of $\tilde{O}(1/\epsilon^6)$ non-zeros per column [Chenakkod et al., STOC 2024]. We further extend our approach to the non-oblivious setting, proposing a new family of Leverage Score Sparsified embeddings with Independent Columns, which yield faster runtimes for matrix approximation and regression tasks.

In our analysis, we develop a new method which uses a decoupling argument together with the cumulant method for bounding the edge universality error of isotropic random matrices. To achieve near-optimal sparsity, we combine this general-purpose approach with new trace inequalities that leverage the specific structure of our subspace embedding construction.

2012 ACM Subject Classification Theory of computation → Sketching and sampling; Theory of computation → Random projections and metric embeddings

Keywords and phrases Randomized linear algebra, matrix sketching, subspace embeddings

Digital Object Identifier 10.4230/LIPIcs.ICALP.2025.55

Category Track A: Algorithms, Complexity and Games

Related Version *Full Version*: `https://arxiv.org/abs/2411.08773` [7]

Funding *Shabarish Chenakkod*: NSF grant DMS 20544.
Michał Dereziński: NSF grant CCF 2338655.

Acknowledgements The authors are grateful for the generous help and support of Mark Rudelson throughout the duration of this work.

1 Introduction

Subspace embeddings are one of the most fundamental techniques in dimensionality reduction, with applications in linear regression [28], low-rank approximation [10], clustering [12], and many more (see [32] for an overview). The key idea is to construct a random linear transformation $\Pi \in \mathbb{R}^{m \times n}$ which maps from a large dimension n to a small dimension m, while approximately preserving the geometry of all vectors in a low-dimensional subspace. In many applications, such embeddings must be constructed without the knowledge of the subspace they are supposed to preserve, in which case they are called *oblivious subspace embeddings*.

▶ **Definition 1.** *Random matrix $\Pi \in \mathbb{R}^{m\times n}$ is an (ε, δ, d)-oblivious subspace embedding (OSE) if for any d-dimensional subspace $T \subseteq \mathbb{R}^n$, it holds that*

$$\mathbb{P}\left(\forall x \in T, \quad (1-\varepsilon)\|x\| \le \|\Pi x\| \le (1+\varepsilon)\|x\|\right) \ge 1-\delta.$$

The two central concerns in constructing OSEs are: 1) how small can we make the embedding dimension m, and 2) how quickly can we apply Π to a vector or a matrix. A popular way to address the latter is to use a sparse embedding matrix: If Π has at most $s \ll m$ non-zero entries per column, then the cost of computing Πx equals $O(s \cdot \mathrm{nnz}(x))$, where $\mathrm{nnz}(x)$ denotes the number of non-zero coordinates in x. Designing oblivious subspace embeddings that simultaneously optimize the embedding dimension m and the sparsity s has been the subject of a long line of works [10, 25, 26, 3, 11, 6], aimed towards resolving the following conjecture of Nelson and Nguyen [26], which is supported by nearly-matching lower bounds [27, 23].

▶ **Conjecture 2** (Nelson and Nguyen, FOCS 2013 [26])**.** *For any $n \ge d$ and $\varepsilon, \delta \in (0,1)$, there is an (ε, δ, d)-oblivious subspace embedding $\Pi \in \mathbb{R}^{m\times n}$ with dimension $m = O((d+\log 1/\delta)/\varepsilon^2)$ having $s = O(\log(d/\delta)/\varepsilon)$ non-zeros per column.*

Nelson and Nguyen gave a simple construction that they conjectured would achieve these guarantees: For each column of Π, place scaled random signs $\pm 1/\sqrt{s}$ in s random locations. They showed that this construction achieves dimension $m = O(d\,\mathrm{polylog}(d)/\varepsilon^2)$ and sparsity $s = O(\mathrm{polylog}(d)/\varepsilon)$. A number of follow-up works [3, 11] improved on this; most notably, Cohen [11] showed that a sparse OSE can achieve $m = O(d\log(d)/\varepsilon^2)$ with $s = O(\log(d)/\varepsilon)$. However, none of these guarantees recover the optimal embedding dimension $m = \Theta(d/\varepsilon^2)$, with the extraneous $\log(d)$ factor arising due to a long-standing limitation in existing matrix concentration techniques [30].

This sub-optimality in dimension m was finally addressed in a recent work of Chenakkod, Dereziński, Dong and Rudelson [6], relying on a breakthrough in random matrix universality theory by Brailovskaya and van Handel [4]. They achieved $m = \Theta(d/\varepsilon^2)$, but only with a significantly sub-optimal sparsity $s = \tilde{O}(1/\varepsilon^6)$, which is a consequence of how the universality error is measured and analyzed in [4] (here, $\tilde{O}$ hides polylogarithmic factors in $d/\varepsilon\delta$). This raises the following natural question:

Can the optimal dimension $m = \Theta(d/\varepsilon^2)$ be achieved with the conjectured $\tilde{O}(1/\varepsilon)$ sparsity?

We give a positive answer to this question, thus matching Conjecture 2 in dimension m and nearly-matching it in sparsity s. To achieve this, we must substantially depart from the approach of Brailovskaya and van Handel, and as a by-product, develop a new set of tools for matrix universality which are likely of independent interest (see Section 4 for an overview). Remarkably, our result is attained by one of the simple constructions that were originally suggested by Nelson and Nguyen in their conjecture.

▶ **Theorem 3** (Oblivious Subspace Embedding)**.** *For any $n \ge d$ and $\varepsilon, \delta \in (0,1)$ such that $1/\epsilon\delta \le \mathrm{poly}(d)$, there is an (ε, δ, d)-oblivious subspace embedding $\Pi \in \mathbb{R}^{m\times n}$ with $m = O(d/\varepsilon^2)$ having $s = \tilde{O}(1/\varepsilon)$ non-zeros per column.*

Many applications of subspace embeddings arise in matrix approximation [32] where, given a large tall matrix $A \in \mathbb{R}^{n\times d}$, we seek a smaller $\tilde{A} \in \mathbb{R}^{m\times d}$ such that $\|\tilde{A}x\| = (1\pm\varepsilon)\|Ax\|$ for all $x \in \mathbb{R}^d$. Naturally, this can be accomplished with an (ε, δ, d)-OSE matrix $\Pi \in \mathbb{R}^{m\times n}$, by computing $\tilde{A} = \Pi A$ in time $\tilde{O}(\mathrm{nnz}(A)/\varepsilon)$ and considering the column subspace of A. However, given direct access to A, one may hope to get true input sparsity time $O(\mathrm{nnz}(A))$ by leveraging the fact that the embedding need not be oblivious.

To that end, we adapt our subspace embedding construction, so that it can be made even sparser given additional information about the leverage scores of matrix A. The ith leverage score of A is defined as the squared norm of the ith row of the matrix obtained by orthonormalizing the columns of A [21]. We show that if the ith leverage score of A is bounded by $l_i \in [0,1]$, then the ith column of Π needs only $\max\{1, \tilde{O}(l_i/\varepsilon)\}$ non-zero entries. Since the leverage scores of A can be approximated quickly [19], this leads to our new algorithm, Leverage Score Sparsified embedding with Independent Columns (LESS-IC), which is inspired by related constructions that use LESS with independent rows [17, 16, 15].

Just like recent prior works [8, 9, 6], our algorithm for constructing a subspace embedding from a matrix A incurs a preprocessing cost of $O(\mathrm{nnz}(A)+d^\omega)$ required for approximating the leverage scores (here, ω is the matrix multiplication exponent). However, our approach significantly improves on these prior works in the $\mathrm{poly}(d/\varepsilon)$ embedding cost, leading to matching speedups in downstream applications such as constrained/regularized least squares [6].

▶ **Theorem 4** (Fast Subspace Embedding). *Given $A \in \mathbb{R}^{n\times d}$, $\varepsilon, \gamma \in (0,1)$ and $1/\varepsilon \le \mathrm{poly}(d)$, in*

$$O\big(\gamma^{-1}\,\mathrm{nnz}(A) + d^\omega + \varepsilon^{-1}d^{2+\gamma}\,\mathrm{polylog}(d)\big) \quad \textit{time}$$

we can compute $\tilde{A} \in \mathbb{R}^{m\times d}$ such that $m = O(d/\varepsilon^2)$ and with probability ≥ 0.99

$$(1-\varepsilon)\|Ax\| \le \|\tilde{A}x\| \le (1+\varepsilon)\|Ax\| \qquad \forall x \in \mathbb{R}^d.$$

This is a direct improvement over the previous best known runtime for constructing an optimal subspace embedding [6], which suffers an additional $\tilde{O}(d^{2+\gamma}/\varepsilon^6)$ cost due to their sub-optimal sparsity. Remarkably, our result is also the first to achieve $\tilde{O}(d^{2+\gamma}/\varepsilon)$ dependence even if we allow a sub-optimal dimension, i.e., $m = O(d\log(d)/\varepsilon^2)$. Here, the previous best time [8, 9] has an additional $\tilde{O}(d^{2+\gamma}/\varepsilon^2)$ cost, due to using a two-stage leverage score sampling scheme in place of a sparse embedding matrix. Our new LESS-IC embedding is crucial in achieving the right dependence on ε, as neither of the previous constructions appear capable of overcoming the $\Omega(d^{2+\gamma}/\varepsilon^2)$ barrier.

As an example application of our results, we show how our fast subspace embedding construction can be used to speed up reductions for a wide class of optimization problems based on constrained or regularized least squares regression, including Lasso regression [3]. The following corollary follows immediately from Theorem 4, and is a direct improvement over Theorem 1.8 of [6] in terms of the runtime dependence on ϵ from $\tilde{O}(d^{2+\gamma}/\epsilon^6)$ to $\tilde{O}(d^{2+\gamma}/\epsilon)$, while achieving a matching $O(d/\epsilon^2)\times d$ reduction.

▶ **Corollary 5** (Fast reduction for constrained least squares). *Given $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, $\epsilon > 0$, function $g : \mathbb{R}^d \to \mathbb{R}_{\ge 0}$ and set $\mathcal{C} \subseteq \mathbb{R}^d$ consider an $n \times d$ problem $\mathrm{LS}_{\mathcal{C},g}(A,b,\epsilon)$:*

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

There is an algorithm that reduces this problem to an $O(d/\epsilon^2)\times d$ instance $\mathrm{LS}_{\mathcal{C},g}(\tilde{A},\tilde{b},0.1\epsilon)$ in $O(\gamma^{-1}\,\mathrm{nnz}(A) + d^\omega + \epsilon^{-1}d^{2+\gamma}\,\mathrm{polylog}(d))$ time.

2 Related Work

Subspace embeddings have played a central role in the area of randomized linear algebra ever since the work of Sarlos [28] (for an overview, see the following surveys and monographs [32, 20, 24, 18]). Initially, these approaches focused on leveraging fast Hadamard transforms [2, 29] to achieve improved time complexity for linear algebraic tasks such as linear regression

and low-rank approximation. Clarkson and Woodruff [10] were the first to propose a sparse subspace embedding matrix, the CountSketch, which has exactly one non-zero entry per column but does not recover the optimal embedding dimension guarantee. Before this, the idea of using a sparse random matrix for dimensionality reduction was successfully employed in the context of Johnson-Lindenstrauss embeddings [14, 22], which seek to preserve the geometry of a finite set, as opposed to an entire subspace.

In addition to the aforementioned efforts in improving sparse subspace embeddings [10, 25, 26, 3, 11, 6], some works have aimed to develop fast subspace embeddings that achieve optimal embedding dimension either without sparsity [8, 9], under additional assumptions [5], or with one-sided embedding bounds [31]. Our time complexity result, Theorem 4, improves on all of these in terms of the dependence on ε, thanks to a combination of our new analysis techniques and the new LESS-IC construction.

3 Main Results

In this section, we define the subspace embedding constructions used in our results, and provide detailed statements of our theorems.

As is customary in the literature, we shall work with an equivalent form of the subspace embedding guarantee from Definition 1, which frames this problem as a characterization of the extreme singular values of a class of random matrices. Namely, consider a deterministic $n \times d$ matrix U with orthonormal columns that form the basis of a d-dimensional subspace T. Then, a random matrix $\Pi \in \mathbb{R}^{m \times n}$ is an (ε, δ, d)-subspace embedding for T if and only if all of the singular values of the matrix ΠU lie in $[1-\varepsilon, 1+\varepsilon]$ with probability $1-\delta$, i.e.,

$$\Pr(1-\varepsilon \leq s_{\min}(\Pi U) \leq s_{\max}(\Pi U) \leq 1+\varepsilon) \geq 1-\delta, \tag{1}$$

where $s_{\min}$ and $s_{\max}$ denote the smallest and largest singular values. To ensure that Π is an oblivious subspace embedding, we must therefore ensure (1) for the family of all random matrices of the form ΠU, where U is any $n \times d$ matrix with orthonormal columns.

3.1 Oblivious Subspace Embeddings

Our subspace embedding guarantees are achieved by a family of OSEs which have a fixed number of non-zero entries in each column, a key property that was also required of sparse OSE distributions called OSNAP described by Nelson and Nguyen [26]. As we explain later, our analysis techniques apply to other natural families of sparse embedding distributions, including those with i.i.d. entries [1], however the OSNAP-style construction is crucial for achieving the near-optimal sparsity $s = \tilde{O}(1/\varepsilon)$.

In our construction of the $m \times n$ OSE matrix Π, we start by defining an unscaled version of the matrix, called S, which has entries in $\{-1, 0, 1\}$. We then scale S to appropriately normalize the entry-wise variances, obtaining Π. Concretely, we wish to obtain an $m \times n$ sparse random matrix S which has exactly s non-zero ± 1 entries in each column. Assume s exactly divides m. Then we can divide each column of S into s subcolumns and randomly populate one entry in each subcolumn by a Rademacher random variable (see Figure 1). We call this family of distributions (unscaled) OSNAP, carrying over Nelson and Nguyen's terminology (technically, their definition is somewhat broader than ours).

Each non-zero entry in the matrix S can be identified by a tuple $(l, \gamma) \in [n] \times [s]$ where l identifies the column of the non-zero entry and γ is the index of the entry in that column. Thus the $(l, \gamma)^{\text{th}}$ non-zero entry in S is located in column l and row $\mu_{(l,\gamma)}$, where $\mu_{(l,\gamma)}$ is a

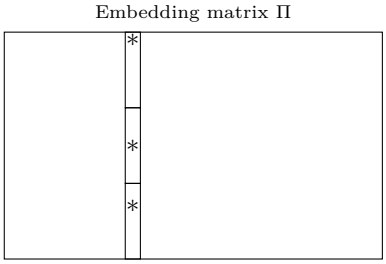

Figure 1 An example of a column divided into $s = 3$ subcolumns with each subcolumn having exactly one non-zero entry in a random position.

uniformly chosen integer from the interval $[(m/s)(\gamma-1)+1 : (m/s)\gamma]$. For example, the $(1,1)^{\text{th}}$ non-zero entry in S is located in column 1 and some row in the interval $[1 : m/s]$. An $m \times n$ matrix with a non-zero entry in column l and row $\mu_{(l,\gamma)}$ is given by $e_{\mu_{(l,\gamma)}} e_l^T$, where $e_{\mu_{(l,\gamma)}}$ and e_l represent standard basis vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, and for S we wish to place a random sign $\xi_{(l,\gamma)}$ at this position. This motivates our formal definition for OSNAP,

▶ **Definition 6** (OSNAP). *An $m \times n$ random matrix S is called an unscaled oblivious sparse norm-approximating projection with K-wise independent subcolumns (K-wise independent unscaled OSNAP) with parameters $p, \varepsilon, \delta \in (0,1]$ such that $s = pm$ divides m if,*

$$S = \sum_{l=1}^{n} \sum_{\gamma=1}^{s} \xi_{(l,\gamma)} e_{\mu_{(l,\gamma)}} e_l^\top$$

where,

- *$\{\xi_{(l,\gamma)}\}_{l\in[n],\gamma\in[s]}$ is a collection of K-wise independent Rademacher random variables.*
- *$\{\mu_{(l,\gamma)}\}_{l\in[n],\gamma\in[s]}$ is a collection of K-wise independent random variables such that each $\mu_{(l,\gamma)}$ is uniformly distributed in $[(m/s)(\gamma-1)+1 : (m/s)\gamma]$.*
- *The collection $\{\xi_{(l,\gamma)}\}_{l\in[n],\gamma\in[s]}$ is independent from the collection $\{\mu_{(l,\gamma)}\}_{l\in[n],\gamma\in[s]}$.*

In this case, $\Pi = (1/\sqrt{pm})S$ is called a K-wise independent OSNAP with parameters p, ε, δ. In addition, if all the random variables in the collections $\{\xi_{(l,\gamma)}\}_{l\in[n],\gamma\in[s]}$ and $\{\mu_{(l,\gamma)}\}_{l\in[n],\gamma\in[s]}$ are fully independent, then S is called a fully independent unscaled OSNAP and Π is called a fully independent OSNAP.

Thus, each column of the OSNAP matrix Π has $s = pm$ many non-zero entries, and the sparsity level can be varied by setting the parameter $p \in [0,1]$ appropriately. With the distribution formally defined, we now provide the full statement of our subspace embedding guarantee for OSNAP,

▶ **Theorem 7** (Subspace Embedding Guarantee for OSNAP). *Let $\Pi = (1/\sqrt{pm})S$ be an $m \times n$ matrix distributed according to the $8\lceil \log(\frac{d}{\varepsilon\delta})\rceil$-wise independent OSNAP distribution with parameter p. Let U be an arbitrary $n \times d$ deterministic matrix such that $U^\top U = I$. Then, there exist positive constants $c_{7.1}$ and $c_{7.2}$ such that for any $0 < \delta, \varepsilon < 1$ and $d > 10$, we have*

$$\mathbb{P}\left(1-\varepsilon \le s_{\min}(\Pi U) \le s_{\max}(\Pi U) \le 1+\varepsilon\right) \ge 1-\delta$$

if the embedding dimension satisfies $m \ge c_{7.1}(d + \log(1/\delta\varepsilon))/\varepsilon^2$ and the sparsity $s = pm$ satisfies $s \ge \min\{c_{7.2}(\log^2(\frac{d}{\varepsilon\delta})/\varepsilon + \log^3(\frac{d}{\varepsilon\delta})), m\}$ non-zeros per column.

▶ Remark 8. We note that if $1/\varepsilon$ is polynomial in d/δ, i.e., $\varepsilon \geq \frac{1}{(d/\delta)^K}$ for some absolute constant $K \geq 1$, then the $\log(1/\varepsilon)$ term in $\log(d/\varepsilon\delta) = \log(d/\delta) + \log(1/\varepsilon)$ is dominated by $\log(d/\delta)$. In this case, our requirement will become

$$pm \geq \min\left\{C(K)\left(\frac{(\log(d/\delta))^2}{\varepsilon} + (\log(d/\delta))^3\right), m\right\}$$

for some constant $C(K)$ depending only on K. A weaker lower bound on ε, $\varepsilon > 1/e^d$ is sufficient to reduce the requirement on m to:

$$m \geq 2c_{7.1}\frac{d + \log(1/\delta)}{\varepsilon^2}.$$

This is a direct improvement over Theorem 1.2 of [6], which requires sparsity $s \geq c\log^4(d/\delta)/\varepsilon^6$ where c is an absolute constant, with the same condition on m. The primary gain lies in the polynomial dependence on $1/\varepsilon$, but we note that our result also achieves a better logarithmic dependence on d, which means that an improvement is obtained even for $\varepsilon = \Theta(1)$.

Our techniques can be used to obtain a similar result for a simple OSE model with i.i.d. sparse Rademacher entries [1], which was also considered by [6]. However, in this case, we need an additional requirement of $s = pm \geq c\log(\frac{d}{\varepsilon\delta})/\varepsilon^2$ for the sparsity (see Section 9 of the technical report for details; this is again a direct improvement over a result of [6]).

▶ Remark 9. The $1/\varepsilon^2$ factor in the column–sparsity of an OSE model with i.i.d. entries is unavoidable. To see why, let

$$U = \begin{bmatrix} I_d \\ 0 \end{bmatrix}, \qquad \Pi = \frac{1}{\sqrt{pm}}\, S,$$

and note that $\sigma_{\min}(\Pi U), \sigma_{\max}(\Pi U) \in [1-\varepsilon,\, 1+\varepsilon]$ forces, for every $j \leq d$,

$$\left|\|\Pi e_j\|_2^2 - 1\right| = \left|\tfrac{N_j}{pm} - 1\right| \leq \varepsilon, \qquad N_j := \mathrm{nnz}(Se_j) \sim \mathrm{Binomial}(m, p). \tag{2}$$

Set

$$Z := \frac{N_j - pm}{\sqrt{mp(1-p)}}, \qquad a := \varepsilon\sqrt{\frac{mp}{1-p}} \leq \sqrt{2}\,\varepsilon\sqrt{mp} \quad (\text{for } p \leq \tfrac{1}{2}).$$

Condition 2 is equivalent to $|Z| \leq a$. With $F_Z(x) = \Pr[Z \leq x]$ and Φ the standard normal cumulative distribution function, the Berry Esseen theorem gives

$$\sup_{x\in\mathbb{R}} |F_Z(x) - \Phi(x)| \leq \frac{6}{\sqrt{mp}}.$$

Hence

$$\Pr\big[|Z| \leq a\big] \;=\; F_Z(a) - F_Z(-a) \;\leq\; \big(\Phi(a) - \Phi(-a)\big) + \frac{12}{\sqrt{mp}}.$$

Using $\Phi(a) - \Phi(-a) = 2\int_0^a \phi(t)\,dt \leq a/\sqrt{\pi}$ and the bound on a, we have

$$\Pr\left(\left|\|\Pi e_j\|_2^2 - 1\right| \leq \varepsilon\right) \;\leq\; \frac{a}{\sqrt{\pi}} + \frac{12}{\sqrt{mp}} \;\leq\; \frac{\sqrt{2}}{\sqrt{\pi}}\varepsilon\sqrt{mp} + \frac{12}{\sqrt{mp}}$$

By general lower bounds for OSE, we know that, when $\varepsilon \to 0$, we need $pm \to \infty$ and therefore so $\frac{12}{\sqrt{mp}} \to 0$.

Therefore, for small enough ε, if $pm < c/\varepsilon^2$ with $c := \frac{1}{81}$, the right–hand side is $< \frac{1}{3}$. Thus any OSE-IE that succeeds with constant probability must satisfy $pm = \Omega\big(\varepsilon^{-2}\big)$.

3.2 Characterization via a Moment Property

Our proof techniques for Theorem 7 are based on the moment method, and thus, they naturally imply the following slightly stronger moment-based characterization of an oblivious subspace embedding, which was proposed by [13] as an extension of the corresponding moment-based characterization of a Johnson-Lindenstrauss embedding [22].

▶ **Definition 10.** *A distribution $\mathcal{D}$ over $\mathbb{R}^{m\times n}$ has $(\varepsilon,\delta,d,\ell)$-OSE moments if, for all matrices $U\in\mathbb{R}^{n\times d}$ with orthonormal columns,*

$$\mathbb{E}_{\Pi\sim\mathcal{D}}\left\|(\Pi U)^T(\Pi U)-I\right\|^\ell<\varepsilon^\ell\delta.$$

Note that a simple application of Markov's inequality recovers the guarantee in Definition 1 from the $(\varepsilon,\delta,d,\ell)$-OSE moments property with any $\ell\geq 1$. Moreover, [13] showed that this moment-based OSE characterization implies several other desirable guarantees of embedding matrices in the context of approximate matrix multiplication, generalized regression and low-rank approximation.

As an immediate consequence of our analysis, we obtain the following OSE moment guarantee for the OSNAP distribution.

▶ **Corollary 11.** *Let Π be an $m\times n$ matrix with an OSNAP distribution having sparsity s. Let $0<\delta,\varepsilon<1$ and $d>10$. Then Π has $(\varepsilon,\delta,d,\ell)$-OSE moments with $\ell=16\log(\frac{d}{\varepsilon\delta})$ when $m\geq c_{11.1}(d+\log(1/\delta\varepsilon))/\varepsilon^2$ and $s\geq\min\{c_{11.2}(\log^2(\frac{d}{\varepsilon\delta})/\varepsilon+\log^3(\frac{d}{\varepsilon\delta})),m\}$.*

▶ Remark 12. Π can be applied to a matrix A in time $O(\mathrm{nnz}(A)(\log^2(\frac{d}{\varepsilon\delta})/\varepsilon+\log^3(\frac{d}{\varepsilon\delta})))$. As noted by [13, Remark 3], such runtimes can be further refined by chaining together several embeddings with an OSE moment property. For example, [11] showed that OSNAP with $m=O(d\log(d/\delta)/\varepsilon^2)$ and $s=O(\log(d/\delta)/\varepsilon)$ has $(\varepsilon,\delta,d,\log(d/\delta))$-OSE moments. Thus, letting $\varepsilon=\Theta(1)$ for simplicity, we can combine a $O(d\log(d/\delta))\times n$ OSNAP matrix Π_1 having sparsity $s=O(\log(d/\delta))$ together with a $O(d+\log(1/\delta))\times O(d\log(d/\delta))$ OSNAP matrix having sparsity $s=O(\log^3(d/\delta))$ to obtain $\Pi=\Pi_2\Pi_1$ with $(\Theta(1),\delta,d,\log(d/\delta))$-OSE moments which can be applied to a matrix $A\in\mathbb{R}^{n\times d}$ in time $O(\mathrm{nnz}(A)\log(d/\delta)+d^2\log^4(d/\delta))$.

3.3 Leverage Score Sparsified Embedding with Independent Columns

In a related problem, we seek to embed a subspace given by a fixed $U\in\mathbb{R}^{n\times d}$, with information about the squared row norms of U being used to define the distribution of non-zero entries in Π. Such distributions for Π are called non-oblivious (a.k.a. data-aware) subspace embeddings. Previous work [6] has dealt with one such family of distributions termed LESS embeddings [17, 16, 15], showing that they require $\tilde{O}(1/\varepsilon^4)$ non-zero entries *per row* of Π to obtain an ε-embedding guarantee. Since the embedding matrix is very wide, this leads to a much sparser embedding (sparser than any OSE) that can be applied in time sublinear in the input size, leading to fast subspace embedding algorithms.

In this work, we show that our new techniques also extend to LESS embeddings and enable us to prove sharper sparsity estimates than [6]. To fully leverage our approach, we define a new type of sparse embedding (LESS-IC), which can be viewed as a cross between CountSketch and LESS. Here, IC stands for independent columns. At a high level, the CountSketch part ensures that we can use our decoupling method to achieve optimal dependence on $1/\varepsilon$, while the LESS part enables adaptivity to a fixed subspace.

Specifically, a LESS-IC embedding matrix Π has a fixed number of non-zero entries in each column, chosen so that it is proportional to the leverage score (i.e. the squared row norm) of the corresponding row of U. This is achieved by modifying the OSNAP distribution

such that the number of subcolumns is no longer the same in each column. For columns corresponding to very small leverage scores, we only have one "subcolumn". Thus, each column has at least one non-zero entry. This means that the cost of applying LESS-IC to an $n \times d$ matrix A can no longer be sublinear (like it can in the existing LESS embedding constructions), but rather has a fixed linear term of $O(\text{nnz}(A))$, plus an additional sublinear term. Given that the preprocessing step of approximating the leverage scores has to take at least $\text{nnz}(A)$ time, the linear term in the cost of applying LESS-IC is negligible.

To generate an embedding matrix with the LESS-IC distribution, it suffices to have a good enough approximation for the leverage scores of the matrix U, in the following sense.

▶ **Definition 13** (Approximate Leverage Scores). *Given a matrix $U \in \mathbb{R}^{n\times d}$ with orthonormal columns and $\beta_1 \geq 1, \beta_2 \geq 1$, a tuple $(l_1, \ldots, l_n) \in [0,1]^n$ of numbers are (β_1, β_2)-approximate leverage scores for U if, for $1 \leq i \leq n$,*

$$\frac{\|e_i^\top U\|^2}{\beta_1} \leq l_i \qquad \text{and} \qquad \sum_{i=1}^n l_i \leq \beta_2 \sum_{i=1}^n \|e_i^\top U\|^2 = \beta_2 d.$$

We say that the numbers $(l_1, \ldots, l_n) \in [0,1]^n$ are β-approximations of the leverage scores (i.e. squared row norms) of U with $\beta = \beta_1\beta_2$.

To see how approximate leverage scores determine the distribution of entries in the LESS-IC distribution, let us first consider a simpler distribution, LESS-IE from [6], based on a similar construction first proposed by [17]. Here, we once again start by defining an unscaled matrix S, which is then normalized to obtain the subspace embedding matrix Π.

▶ **Definition 14** (LESS-IE). *An $m \times n$ random matrix S is called an unscaled leverage score sparsified embedding with independent entries (unscaled LESS-IE), and also $\Pi = (1/\sqrt{pm})S$ is called a LESS-IE, corresponding to (β_1, β_2)-approximate leverage scores $(z_1, ..., z_n)$ with parameter p, if S has entries $s_{i,j} = \frac{1}{\sqrt{\beta_1 z_j}}\delta_{i,j}\xi_{i,j}$ where $\delta_{i,j}$ are independent Bernoulli random variables taking value 1 with probability $p_{ij} = \beta_1 z_j p$, whereas $\xi_{i,j}$ are i.i.d. Rademacher random variables.*

In the LESS-IE model, we have $\beta_1 pmz_j$ many non-zero entries in column j in expectation. However, to achieve $1/\varepsilon$ dependency of the sparsity, we need to have *exactly* $\beta_1 pmz_j$ many non-zero entries in the column in the LESS-IC model to fully take advantage of the error cancellation that occurs in our decoupling argument (See Section 7.2 and Section 9.1 of the technical report). Though these sections deal with oblivious subspace embeddings, the same arguments still apply in the LESS case). This is done by modifying the OSNAP construction so that the size (and consequently, the number) of subcolumns is different across columns.

Notice that to have $\beta_1 pmz_j$ many non-zero entries in column j, we would need $\beta_1 pmz_j$ many subcolumns in column j each with one non-zero entry in a random position. This means that the size of each subcolumn needs to be $m/(\beta_1 pmz_j) = 1/(\beta_1 pz_j)$. However, since $1/(\beta_1 pz_j)$ may not be an integer, we consider subcolumns of size $b_j := \max\{\lfloor 1/(\beta_1 pz_j)\rfloor, 1\}$.

In column j, we stack subcolumns of size b_j until we fill up all the rows up to m. Let s_j be the smallest number of subcolumns to do this. Then, it may happen that the row indices of the bottom-most subcolumn exceed m. For example, consider the distribution on the first column of Π when $m = 70$, and $b_1 = 15$. In this case $s_1 = 5$, so we can stack four subcolumns of size 15 and the 5^{th} subcolumn only spans row indices $[61 : 70]$. In each subcolumn, we randomly choose a row to place a non-zero entry, which would be a Rademacher random variable. (See Figure 2). The non-zero entries are appropriately scaled so that all entries of the matrix have the same variance (See Section 8 of the technical report for the full definition).

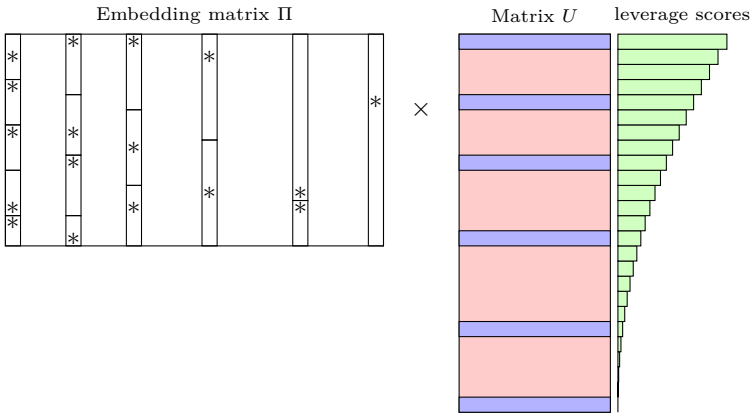

Figure 2 In the LESS-IC distribution, column j is filled with s_j many subcolumns, with the bottom-most subcolumn truncated to fit the size of Π. Each subcolumn has one non-zero entry. Notice that as the leverage scores decrease, the number of subcolumns decreases and the matrix becomes sparser. However, each column always has at least one non-zero entry.

For the LESS-IC distribution, we show the following subspace embedding guarantee. The structure of the proof is similar to the case of OSNAP, and only the specific expressions change due to the different distribution.

▶ **Theorem 15** (Subspace Embedding Guarantee for LESS-IC). *Let $\Pi = (1/\sqrt{pm})S$ be an $m \times n$ matrix distributed according to the $8\lceil \log(\frac{d}{\varepsilon\delta})\rceil$-wise independent LESS-IC distribution with parameter p for some fixed $n \times d$ matrix U satisfying $U^\top U = I$ with given (β_1, β_2)-approximate leverage scores. Then, there exist positive constants $c_{15.1}$ and $c_{15.2}$ such that for any $0 < \varepsilon, \delta < 1$, and $d > 10$, we have*

$$\mathbb{P}\left(1 - \varepsilon \le s_{\min}(\Pi U) \le s_{\max}(\Pi U) \le 1 + \varepsilon\right) \ge 1 - \delta$$

when $m \ge c_{15.1}\left(\frac{d+\log^2(d/\delta)+\log(1/\varepsilon)}{\varepsilon^2} + \log^3(d/\delta)/\varepsilon\right)$ *and*

$$c_{15.2}\max\left\{\frac{(\log(d/\varepsilon\delta))^{2.5}}{\varepsilon}, (\log(d/\varepsilon\delta))^3\right\} \le pm \le m.$$

The matrix Π has $O(n + \beta pmd)$ many non-zero entries and can be applied to an $n \times d$ matrix A in $O(\mathrm{nnz}(A) + \beta pmd^2)$ time, where $\beta = \beta_1\beta_2$ is the leverage score approximation factor.

▶ Remark 16. When $\delta = d^{-O(1)}$, we recover the optimal dimension $m = \Theta(d/\varepsilon^2)$ while showing that one can apply the LESS-IC embedding in time $O(\mathrm{nnz}(A)) + \tilde{O}(\beta d^2/\varepsilon)$. In comparison, [6] showed that a corresponding LESS-IE embedding can be applied in $\tilde{O}(\beta d^2/\varepsilon^6)$ time. Using our techniques, one could improve the runtime of LESS-IE to $\tilde{O}(\beta d^2/\varepsilon^2)$, but our new LESS-IC construction appears necessary to recover the best dependence on $1/\varepsilon$.

3.4 Fast Subspace Embedding (Proof of Theorem 4)

Here, we briefly outline how our LESS-IC embedding yields a fast subspace embedding construction to recover the time complexity claimed in Theorem 4. This follows analogously to the construction from Theorem 1.6 of [6], and our improvement in the dependence on $1/\varepsilon$ compared to their result (from $1/\varepsilon^6$ to $1/\varepsilon$) stems from the improved sparsity of our LESS-IC embedding.

The key preprocessing step for applying the LESS-IC embedding is approximating the leverage scores of the matrix A. Using Lemma 5.1 in [6] (adapted from Lemma 7.2 in [8]), we can construct coarse approximations of all leverage scores so that $\beta_1 = O(n^\gamma)$ and $\beta_2 = O(1)$ in time $O(\gamma^{-1}(\text{nnz}(A) + d^2) + d^\omega)$. Applying LESS-IC (Theorem 15) with these leverage scores and parameters β_1, β_2, computing ΠA takes $O(\text{nnz}(A) + n^\gamma d^2 \log^3(d/\varepsilon\delta)/\varepsilon)$, where $\text{nnz}(A)$ comes from the fact that every column of Π has at least one non-zero, while the second term accounts for the additional $O(\beta d \log^3(d/\varepsilon\delta)/\varepsilon)$ non-zeros.

Thus, if $d \geq n^c$ for, say, $c = 0.1$, then we conclude the claim by appropriately scaling γ by a constant factor. Now, suppose otherwise. First, note that without loss of generality we can assume that $\gamma < 0.1$ (through scaling the time complexity by a constant factor), $\text{nnz}(A) \geq n$ (by removing empty rows) and $\varepsilon \geq \sqrt{d/n}$ (because otherwise $m \geq n$ and we could use $\tilde{A} = A$). Thus, under our assumption that $d < n^c$, we have $n^\gamma d^2/\varepsilon \leq n^{0.5+\gamma+2c} \leq n^{0.8} \ll \text{nnz}(A)$, and the time complexity is dominated by the $O(\gamma^{-1}\,\text{nnz}(A))$ term.

Finally, we note that Corollary 5 follows simply by constructing a subspace embedding Π via Theorem 4 with respect to matrix $[A \mid b]$, and computing $\tilde{A} = \Pi A$, $\tilde{b} = \Pi b$. The proof of the claim is identical to the proof of Theorem 1.8 in [6]. Our improvement comes directly from the faster runtime of our subspace embedding construction.

3.5 Outline of the Paper

Section 4 provides a high level overview of the ideas used in the proofs of our main results, Theorem 7 and Theorem 15. Section 5 provides a sketch of the proof of Theorem 7, listing the main technical steps, leaving the full proof with all technical details to Section 7 of the technical report. The proof of Theorem 15 follows similarly and is covered in Section 8 of the technical report. The subspace embedding guarantee for a sparse matrix with independent entries is proved in Section 9 of the technical report.

3.6 Notation

The following notation and terminology will be used in the paper. The notation $[n]$ is used for the set $\{1, 2, ..., n\}$ and the notation $\text{P}([n])$ denotes the set of all partitions of $[n]$. Also, for two integers a and b with $a \leq b$, we use the notation $[a : b]$ for the set $\{k \in \mathbb{Z} : a \leq k \leq b\}$. For $x \in \mathbb{R}$, we use the notation $\lfloor x \rfloor$ to denote the greatest integer less than or equal to x and $\lceil x \rceil$ to denote the least integer greater than or equal to x. In $\mathbb{R}^n$ (or $\mathbb{R}^m$ or $\mathbb{R}^d$), the lth coordinate vector is denoted by e_l. All matrices considered in this paper are real valued and the space of $m \times n$ matrices with real valued entries is denoted by $M_{m\times n}(\mathbb{R})$. Also, for a matrix $X \in M_{d\times d}(\mathbb{R})$, the notation $\text{Tr}(X)$ denotes the trace of the matrix X, and $\text{tr}(X) = \frac{1}{d}\,\text{Tr}(X)$ denotes the normalized trace. We write the operator norm of a matrix X as $\|X\|$, and it is also denoted by $\|X\|_{op}$ in some places where other norms appear for clarity. The spectrum of a matrix X is denoted by $\text{spec}(X)$. The standard probability measure is denoted by $\mathbb{P}$, and the symbol $\mathbb{E}$ means taking the expectation with respect to this standard probability measure. To simplify the notation, we follow the convention from [4] and use the notation $\mathbb{E}[X]^\alpha$ for $(\mathbb{E}(X))^\alpha$, i.e., when a functional is followed by square brackets, it is applied before any other operations. The covariance of two random variables X and Y is denoted by $\text{Cov}(X, Y)$. The standard L_q norm of a random variable ξ is denoted by $\|\xi\|_q$, for $1 \leq q \leq \infty$. Throughout the paper, the symbols $c_1, c_2, ...$, and $Const, Const', ...$ denote absolute constants.

4 Main Ideas

We next outline our new techniques which are needed to establish the main results, Theorems 7 and 15. Here, for notational convenience, we will refer to the unscaled random matrix S, as opposed to the subspace embedding matrix $\Pi = (1/\sqrt{pm})S$ (see Definition 6).

Note that due to the equivalent characterization of the OSE property in (1), all we need to show is that singular values of SU are clustered around $\sqrt{pm}$ at distance $O(\sqrt{pm}\varepsilon)$. In other words, we need to show that the difference between the spectrum of SU and the spectrum of $\sqrt{pm}I_d$ is small, of the order $O(\sqrt{pm}\varepsilon)$.

In all our models, the entries of S are uncorrelated with mean 0 and variance p, and therefore the entries of SU are uncorrelated with uniform variance. If we consider a random matrix G with Gaussian entries which keeps the covariance profile of the entries of SU, then this Gaussian random matrix G has independent Gaussian entries with variance p. Using classical results about singular values of Gaussian random matrices, it can be shown that the singular values of G are sufficiently clustered around $\sqrt{pm}$ with high probability for $m = \Omega(d/\varepsilon^2)$. Thus, it suffices to find conditions under which the singular values of SU are sufficiently close to the singular values of G. This is the phenomenon of universality whereby random systems show predictable (in this case Gaussian) behavior under certain limits.

Failure of black-box universality. Recent work by Brailovskaya-van Handel [4] on universality for certain random matrix models developed tools to bound the distance between the spectrum of a random matrix model obtained as a sum of independent random matrices and the spectrum of a Gaussian random matrix with the same covariance profile. Using these tools, [6] achieved optimal embedding dimension $m = O(d/\varepsilon^2)$ for OSEs by using the bound in [4, Theorem 2.6] to estimate the Hausdorff distance (a concept of distance between two subsets of $\mathbb{R}$; $A, B \subset \mathbb{R}$ are said to be ε-close in Hausdorff distance if A is in the ε-neighborhood of B and B is in the ε neighborhood of A) between the spectra of

$$\mathrm{sym}(SU) = \begin{bmatrix} & (SU)^T \\ SU & \end{bmatrix} \quad \text{and} \quad \mathrm{sym}(G) = \begin{bmatrix} & G^T \\ G & \end{bmatrix}.$$

This distance is shown to be $(O(\sqrt{pm}))^{2/3}$, which is of order $\sqrt{pm}\varepsilon$ only when pm has $1/\varepsilon^6$ dependence. Thus, [6] did not obtain the conjectured dependency of the sparsity on ε, which requires pm to only have $1/\varepsilon$ dependency. To get better ε dependency, we would either need a sharper bound on the Hausdorff distance, or have the distance decrease with ε. For example, if the $(O(\sqrt{pm}))^{2/3}$ bound was improved to $(O(\sqrt{pm}))^{1/2}$, we would only need $(\sqrt{pm})^{1/2} \le \sqrt{pm}\varepsilon$ which can be achieved when pm has $1/\varepsilon^4$ dependence. On the other hand, if the $(O(\sqrt{pm}))^{2/3}$ bound was improved to $(O(\sqrt{pm}))^{2/3}\varepsilon^{1/2}$, we would only need pm to have $1/\varepsilon^3$ dependence.

Key idea: Universality of centered moments. One can instead look at a different approach to characterize the clustering of singular values. To show that the singular values of ΠU are between $1 \pm \varepsilon$, it is enough to show that $\|(\Pi U)^T \Pi U - I_d\| \le \varepsilon$ or $\|(SU)^T SU - pm \cdot I_d\| \le pm\varepsilon$ (Note that $S = \sqrt{pm}\Pi$). One way to achieve this bound with high probability is to use the moment method, i.e., to show that (see proof of Theorem 7 in Section 5):

$$\mathbb{E}\left[\mathrm{tr}\left((SU)^T(SU) - pm \cdot I_d\right)^{2q}\right]^{\frac{1}{2q}} = O(pm\varepsilon).$$

In this case, standard calculations on Gaussian random matrices (see Section 6 of the technical report) show that $(\mathbb{E}[\mathrm{tr}(G^TG - pmI_{d\times d})^{2q}])^{\frac{1}{2q}} \leq cpm\sqrt{\frac{d}{m}} = O(pm\varepsilon)$ when $m = \Omega(d/\varepsilon^2)$ and G has the covariance profile of SU. So it is enough to show that

$$\mathbb{E}\left[\mathrm{tr}\left((SU)^T(SU) - pmI_d\right)^{2q}\right]^{\frac{1}{2q}} - \mathbb{E}\left[\mathrm{tr}\left(G^TG - pmI_d\right)^{2q}\right]^{\frac{1}{2q}} = O(pm\varepsilon).$$

where we recall the notation $\mathbb{E}\left[\mathrm{tr}\left(X\right)^{2q}\right]^{\frac{1}{2q}} = \left(\mathbb{E}\,\mathrm{tr}\,(X)^{2q}\right)^{1/(2q)}$.

Now, [4, Proposition 9.12] does take a similar approach of comparing $(SU)^T(SU) - pmI_d$ and $G^TG - pmI_d$, by relying on an interpolation argument, where one defines a mixture $S(t) = \sqrt{t}S + \sqrt{1-t}G$ and controls the change in the moments along the trajectory specified by $t \in [0,1]$. Unfortunately, using that result gives a larger power of pm in the bound than desired, resulting again in a worse ε dependence.

One can also, by viewing $(SU)^T(SU) - pmI_d = \sum_{i=1}^m (U^Ts_is_i^TU - pI_d)$, get a random matrix model which is a sum of independent random matrices (this is not true for OSNAP, but some other models of OSEs), and then compare $\mathbb{E}[\mathrm{tr}((SU)^T(SU) - pm\cdot I_d)^{2q}]^{\frac{1}{2q}}$ with $\mathbb{E}[\mathrm{tr}(H)^{2q}]^{\frac{1}{2q}}$ where H is the Gaussian model for $(SU)^T(SU) - pmI_d$. This is the approach of [4, Proposition 9.15], but it fails in obtaining the optimal embedding dimension $m = d/\varepsilon^2$.

Key technique: Decoupling. To overcome these obstacles, we develop a fresh analysis while still using the ideas of [4]. Our first step is to observe that due to the property of S having a fixed number of non-zero entries in a column for the OSNAP distribution, all quadratic terms in $(SU)^T(SU) - pm\cdot I_d$ are square-free, and this allows us to use the decoupling technique to reduce the problem of controlling the moments of $(SU)^T(SU) - pm\cdot I_d$ to controlling the moments of $(S_1U)^T(S_2U) + (S_2U)^T(S_1U)$ where S_1 and S_2 are independent copies of S (See proof of Lemma 17 in Section 5).

We still have to separate bounding $\mathbb{E}[\mathrm{tr}((S_1U)^T(S_2U) + (S_2U)^T(S_1U))^{2q}]^{\frac{1}{2q}}$ into two parts, bounding $\mathbb{E}[\mathrm{tr}(G_1^TG_2 + G_2^TG_1)^{2q}]^{\frac{1}{2q}}$ for the Gaussian model, and the difference

$$\mathbb{E}[\mathrm{tr}((S_1U)^T(S_2U) + (S_2U)^T(S_1U))^{2q}]^{\frac{1}{2q}} - \mathbb{E}[\mathrm{tr}(G_1^TG_2 + G_2^TG_1)^{2q}]^{\frac{1}{2q}},$$

which is called the universality error.

By standard calculations, we have $\mathbb{E}[\mathrm{tr}(G_1^TG_2 + G_2^TG_1)^{2q}]^{\frac{1}{2q}} \leq c\sqrt{pm}\sqrt{pd} = O(pm\varepsilon)$ and the main task is still to bound the universality error. The advantage of the decoupling idea is that, informally speaking, since S_1 and S_2 are independent, we can condition on one of them, e.g., S_1. For fixed S_1, the random matrix $(S_1U)^T(S_2U)$ (where all randomness comes from S_2) can be viewed as a sum of independent random matrices, with the individual summands having moments of smaller order than the previous approach. We can then use an interpolation argument to bound the trace universality error for $q = \log(\frac{d}{\varepsilon\delta})$ as follows:

$$\begin{aligned}&\left|\mathbb{E}\big[\,\mathrm{tr}((S_1U)^T(S_2U) + (S_2U)^T(S_1U))^{2q}\big]^{\frac{1}{2q}} - \mathbb{E}\big[\,\mathrm{tr}(G_1^TG_2 + G_2^TG_1)^{2q}\big]^{\frac{1}{2q}}\right| \\ &\leq \mathrm{polylog}(\tfrac{d}{\varepsilon\delta}).\end{aligned} \tag{3}$$

Notice that there is no pm dependence on the right hand side. So our requirement that this quantity be bounded by $pm\varepsilon$ is satisfied when $pm \geq \mathrm{polylog}(\frac{d}{\varepsilon\delta})/\varepsilon$, achieving the conjectured $1/\varepsilon$ dependence.

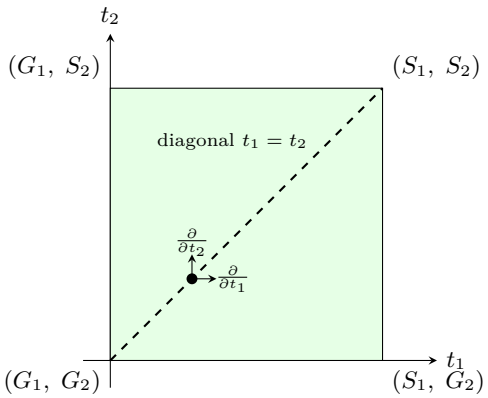

Figure 3 Two-dimensional interpolation in $(t_1, t_2) \in [0,1]^2$, decomposed using the chain rule.

Nevertheless, the conditioning argument cannot be done directly because

$$\mathbb{E}\Big[\operatorname{tr}\big((S_1U)^T(S_2U) + (S_2U)^T(S_1U)\big)^{2q}\Big]^{\frac{1}{2q}}$$
$$\neq \mathbb{E}_{S_1}\Big[\mathbb{E}_{S_2}\Big[\operatorname{tr}\big((S_1U)^T(S_2U) + (S_2U)^T(S_1U)\big)^{2q}\Big]^{\frac{1}{2q}}\Big].$$

Key technique: 2D interpolation via chain rule. So, instead we develop a new approach which incorporates the conditioning step directly into a two-dimensional interpolation argument, through the use of the chain rule (see Figure 3). Define

$$S_1(t_1) = \sqrt{t_1}\, S_1 + \sqrt{1-t_1}\, G_1, \quad S_2(t_2) = \sqrt{t_2}\, S_2 + \sqrt{1-t_2}\, G_2.$$

We start from (G_1, G_2) at $(t_1, t_2) = (0,0)$ and move to (S_1, S_2) at $(1,1)$, interpolating between the easier-to-analyze Gaussian matrices (G_1, G_2) and the true random matrices (S_1, S_2) of interest and controlling the changes in their moments (or the error terms) step by step.

Defining $f(M_1, M_2) = \operatorname{tr}((M_1U)^T(M_2U) + (M_2U)^T(M_1U))^{2q}$, and applying the chain rule on the diagonal $t_1 = t_2 = t$, we obtain:

$$\frac{d}{dt}\mathbb{E}\big[f\big(S_1(t), S_2(t)\big)\big]$$
$$= \frac{\partial}{\partial t_1}\,\mathbb{E}\big[f\big(S_1(t_1), S_2(t_2)\big)\big]\Big|_{t_1=t,\, t_2=t} + \frac{\partial}{\partial t_2}\,\mathbb{E}\big[f\big(S_1(t_1), S_2(t_2)\big)\big]\Big|_{t_1=t,\, t_2=t}.$$

By independence of S_1 and S_2, we can condition on S_2 when we bound the partial derivative $\frac{\partial}{\partial t_1}\,\mathbb{E}\big[f\big(S_1(t_1), S_2(t_2)\big)\big]\big|_{t_1,t_2=t}$, and do similar calculations for the other term. The benefit of doing this is that we can now fine tune the techniques of [4] to get a differential inequality (Lemma 20) that leads to inequality (3). In doing so, we are able to find the optimal bounds and exponents in the differential inequality.

5 Proof Sketch for the Oblivious Subspace Embedding

We now sketch the proof of our main subspace embedding guarantee, Theorem 7 for OSNAP. The full proof can be found in Section 7 of the technical report. The proof of the subspace embedding guarantee for LESS-IC, Theorem 15 is similar and can be found in Section 8 of the technical report.

Proof sketch of Theorem 7. Let $X := \frac{1}{\sqrt{pm}} SU$. We first assume that the collection of all the random variables $\{\xi_{(l,\gamma)}, \mu_{(l,\gamma)}\}_{l\in[n],\gamma\in[s]}$ in the unscaled OSNAP construction are fully independent, and later we will check what is the minimum independence needed.

We observe that to prove the theorem, it is enough to show that

$$\mathbb{P}\left(\|X^T X - I_d\| \le \varepsilon\right) \ge 1 - \delta.$$

We call the quantity $X^T X - I_d$ the embedding error. By Markov's inequality, we have

$$\mathbb{P}\left(\|X^T X - I_d\| \ge \delta^{-\frac{1}{2q}}\, \mathbb{E}[d\,\mathrm{tr}(X^T X - I_d)^{2q}]^{\frac{1}{2q}}\right) \le \delta$$

which, after simplification, becomes

$$\mathbb{P}\left(\|X^T X - I_d\| \ge (d/\delta)^{\frac{1}{2q}}\, \mathbb{E}[\mathrm{tr}(X^T X - I_d)^{2q}]^{\frac{1}{2q}}\right) \le \delta.$$

For $q > \log(d/\delta)$, we have

$$(d/\delta)^{\frac{1}{2q}} = \exp\left(\log(d/\delta)\frac{1}{2q}\right) \le \exp\left(\log(d/\delta)\frac{1}{2\log(\frac{d}{\delta})}\right) \le \sqrt{e}.$$

Therefore, we have

$$\mathbb{P}\left(\|X^T X - I_d\| \ge \sqrt{e}\,\mathbb{E}[\mathrm{tr}(X^T X - I_d)^{2q}]^{\frac{1}{2q}}\right) \le \delta.$$

Thus we need to control moments of order $2q$ of the embedding error for $q > \log(d/\delta)$, and this is done in the following lemma.

▶ **Lemma 17** (Trace Moments of Embedding Error for OSNAP, see Section 7 of the technical report)**.** *For X as above, there exist constants $c_{17.1}, c_{17.2}, c_{17.3} > 0$ such that for $q \in \mathbb{N}$ satisfying $2 \le q \le m$, we have*

$$\mathbb{E}[\mathrm{tr}(X^T X - I_d)^{2q}]^{\frac{1}{2q}} \le \varepsilon,$$

$$\text{when} \quad m \ge c_{17.1}\frac{d+q}{\varepsilon^2} \tag{4}$$

$$\text{and} \quad pm \ge \left(\max\left\{\frac{c_{17.2}q^2}{\varepsilon}, c_{17.3}q^3\right\}\right)^{1+\frac{2}{q-2}}. \tag{5}$$

Applying Lemma 17 (with appropriately adjusted ε) implies

$$\mathbb{P}\left(\|X^T X - I_d\| \ge \varepsilon\right) \le \delta$$

when combined with the previous calculations.

It remains to check that conditions (4) and (5) are satisfied for $q = \lceil 2\log(\frac{d}{\varepsilon\delta})\rceil + 2$ by requiring $m \ge c_{7.1}(d + \log(1/\delta\varepsilon))/\varepsilon^2$ and $s = pm \ge c_{7.2}(\log^2(\frac{d}{\varepsilon\delta})/\varepsilon + \log^3(\frac{d}{\varepsilon\delta}))$, and this is done in the full version of the proof in Section 7 of the technical report.

Note that the expression for $\mathbb{E}[\mathrm{tr}(X^T X - I_d)^{2q}]$ depends only on $2q$ fold products of the entries of X. So, the quantity $\mathbb{E}[\mathrm{tr}(X^T X - I_d)^{2q}]$ remains unchanged if we only assume that subsets of the entries of X of size $2q$ are independent instead of arbitrary subsets of the entries of X being independent. Since it suffices to choose $q = \lceil 2\log(\frac{d}{\varepsilon\delta})\rceil + 2$, we only need S to be an $O(\log(d/\varepsilon\delta))$-wise independent unscaled OSNAP. ◀

Finally, we show how the above arguments also imply the OSE moment property (Definition 10).

Proof of Corollary 11. By the proof of Theorem 7, we see that when $m \geq c_{7.1}(d + \log(1/\delta\varepsilon))/\varepsilon^2$ and $s \geq \min\{c_{7.2}(\log^2(\frac{d}{\varepsilon\delta})/\varepsilon + \log^3(\frac{d}{\varepsilon\delta})), m\}$, then $\mathbb{E}[\mathrm{tr}(X^TX - I_d)^{2q}] \leq \varepsilon^{2q}$, for $q = 8\log(\frac{d}{\varepsilon\delta})$ (The proof originally has $q = \lceil 2\log(\frac{d}{\varepsilon\delta})\rceil + 2$, but upon going through the proof we see that $q = 8\log(\frac{d}{\varepsilon\delta})$ also works). To get $\mathbb{E}[\mathrm{tr}(X^TX - I_d)^{2q}] \leq \varepsilon^{2q}\delta/d$, it suffices for m and s to satisfy the same lower bounds, but with ε replaced by $\varepsilon(\delta/d)^{\frac{1}{2q}} \geq c\varepsilon$ for some $c > 0$ since $q \geq \log(d/\delta)$. These new lower bounds can be achieved by lower bounds of the same form as Theorem 7, but with different constants. The claim follows, since $\|X^TX - I_d\|^{2q} \leq d\,\mathrm{tr}(X^TX - I_d)^{2q}$. ◀

5.1 Controlling Trace Moments of the Embedding Error

We now sketch the proof of Lemma 17, which obtains the moment bound for $X^TX - I$ used in the previous proof. The full proof can be found in Section 7 of the technical report.

Proof sketch of Lemma 17. Our first step is to observe that due to the property of S having a fixed number of non-zero entries in a column, all quadratic terms in $(SU)^T(SU) - pm \cdot I_d$ are square-free, and this allows us to use the decoupling technique to reduce the problem of controlling the moments of $(SU)^T(SU) - pm \cdot I_d$ to controlling the moments of $(S_1U)^T(S_2U) + (S_2U)^T(S_1U)$ where S_1 and S_2 are independent copies. This is shown in the following claim, with the proof deferred to Section 7 of the technical report.

▶ **Lemma 18** (Decoupling). *When S has the fully independent unscaled OSNAP distribution, we have*

$$\mathbb{E}\left[\mathrm{tr}\left(U^TS^TSU - pm \cdot I_d\right)^{2q}\right] = \mathbb{E}\left[\mathrm{tr}\left(\sum_{i=1}^{m}\sum_{j,j'=1,j\neq j'}^{n} s_{ij}s_{ij'}u_ju_{j'}^T\right)^{2q}\right]$$

where $\{u_j^T\}_{j\in[n]}$ denote the rows of U. Consequently, we have

$$\mathbb{E}\left[\mathrm{tr}\left(U^TS^TSU - pm \cdot I_d\right)^{2q}\right] \leq \mathbb{E}_{S_1,S_2}\left[\mathrm{tr}\left(2\big((S_1U)^TS_2U + (S_2U)^TS_1U\big)\right)^{2q}\right]$$

where S_2 is an independent copy of S_1.

To estimate the moments of $(S_1U)^TS_2U$, we compare them to moments from the Gaussian case, i.e. the moments of $(G_1U)^TG_2U$ where the entries of G_1 and G_2 are independent normal random variables with variance p (since the entries of S_1 and S_2 are also uncorrelated with mean 0 and variance p, see Section 6 of the technical report). In this case, due to orthogonal invariance of the Gaussian distribution, the matrices G_1U and G_2U are distributed as $\sqrt{p}H_1$ and $\sqrt{p}H_2$ where H_1 and H_2 are $m \times d$ matrices with independent standard normal entries. Thus, we can rely on the following bound, which uses standard results about the norms of Gaussian random matrices with independent entries.

▶ **Lemma 19** (Trace Moment of Embedding Error for Decoupled Gaussian Model, see Section 6 of the technical report). *Let H_1 and H_2 be independent $m \times d$ random matrices with i.i.d. Gaussian entries. Then for any positive integer q, there exists $c_{19} > 0$ such that*

$$\mathbb{E}\left[\mathrm{tr}\left(H_1^TH_2 + H_2^TH_1\right)^{2q}\right]^{\frac{1}{2q}} \leq c_{19}\sqrt{\max\{d,q\}}\sqrt{\max\{m,q\}}.$$

To formally compare the moments of $(S_1U)^TS_2U$ and $(G_1U)^TG_2U$, we define the interpolating matrices $S_1(t), S_2(t)$ for $t \in [0,1]$ as described in Section 4:

$$S_1(t) = \sqrt{t}S_1 + \sqrt{1-t}G_1,$$
$$S_2(t) = \sqrt{t}S_2 + \sqrt{1-t}G_2. \tag{6}$$

Let $\Gamma(M_1, M_2) = (M_1U)^T(M_2U) + (M_2U)^T(M_1U)$ and $\Gamma(t) = \Gamma(S_1(t), S_2(t))$. Then, due to the decoupling lemma (Lemma 18), to prove Lemma 17 it is enough to show that $\mathbb{E}[\mathrm{tr}(\Gamma(1))^{2q}]^{\frac{1}{2q}} \leq pm\varepsilon/2$. Now, by Lemma 19, we know that:

$$\mathbb{E}\left[\mathrm{tr}(\Gamma(0))^{2q}\right]^{\frac{1}{2q}} = \mathbb{E}\left[\mathrm{tr}(\Gamma(G_1, G_2))^{2q}\right]^{\frac{1}{2q}} \leq c_{19}p\sqrt{\max\{d,q\}}\sqrt{\max\{m,q\}}.$$

Since we want to find the conditions for which $\mathbb{E}[\mathrm{tr}(\Gamma(0))^{2q}]^{\frac{1}{2q}} \leq pm\varepsilon/4$, it is enough to ensure that $c_{19}p\sqrt{\max\{d,q\}}\sqrt{\max\{m,q\}} \leq pm\varepsilon/4$. Clearly, this can only happen when $q \leq m$, and in this case the inequality holds when $m \geq \frac{c(d+q)}{\varepsilon^2}$. Thus, it suffices to show

$$\mathbb{E}\left[\mathrm{tr}\,\Gamma(1)^{2q}\right]^{\frac{1}{2q}} - \mathbb{E}\left[\mathrm{tr}\,\Gamma(0)^{2q}\right]^{\frac{1}{2q}} \leq \frac{1}{4}pm\varepsilon. \tag{7}$$

For this, we look to estimate the derivative $\frac{d}{dt}\mathbb{E}[\mathrm{tr}\,\Gamma(t)^{2q}]$, and we obtain the following estimate in Lemma 20 using the 2D interpolation idea mentioned in Section 4.

▶ **Lemma 20** (Differential Inequality). *For $\Gamma(t)$ as defined above, there exists a constant c_{20} such that, for any $q \geq 2$, we have*

$$\frac{d}{dt}\mathbb{E}\left[\mathrm{tr}\,\Gamma(t)^{2q}\right] \leq \max_{4\leq k\leq 2q}(c_{20}q)^k\left((pm)^{\frac{1}{q}}\sqrt{\max\{pd,q\}}\right)^{\frac{qk-2q}{q-1}}\mathbb{E}\left[\mathrm{tr}\,\Gamma(t)^{2q}\right]^{1-\frac{k-2}{2q-2}}.$$

This differential inequality can be separated into two distinct cases: $pd \leq q$ and $pd > q$. When $pd \leq q$, we can simplify the expression on the right using convexity arguments, and use Lemma 6.6 from [4] to solve the differential inequality and obtain the following bound:

$$\mathbb{E}\left[\mathrm{tr}\,\Gamma(1)^{2q}\right]^{\frac{1}{2q}} - \mathbb{E}\left[\mathrm{tr}\,\Gamma(0)^{2q}\right]^{\frac{1}{2q}} \leq c_7(pm)^{\frac{1}{q}}q^2.$$

for some $c_7 > 0$ (this is done in the full proof of Lemma 17 in the technical report). Thus, inequality (7) is satisfied when $c_7(pm)^{\frac{1}{q}}q^2 < pm\varepsilon/4$, or

$$pm \geq \frac{4c_7(pm)^{\frac{1}{q}}q^2}{\varepsilon}.$$

When $pd > q$, the expression on the right of the above differential inequality has some pd factors. We replace these pd factors by terms involving only pm and $\mathbb{E}[\mathrm{tr}\,\Gamma(t)^{2q}]$ and similarly obtain:

$$\mathbb{E}\left[\mathrm{tr}\,\Gamma(1)^{2q}\right]^{\frac{1}{2q}} - \mathbb{E}\left[\mathrm{tr}\,\Gamma(0)^{2q}\right]^{\frac{1}{2q}} \leq c_{13}q^3(pm)^{\frac{2}{q}}\left(\frac{d}{m}\right)^{\frac{1}{2}}$$

for some $c_{13} > 0$. In this case, inequality (7) is satisfied when

$$c_{13}q^3(pm)^{\frac{2}{q}}\left(\frac{d}{m}\right)^{\frac{1}{2}} \leq \frac{1}{4}pm\varepsilon.$$

Since we have $m \geq \frac{c_{14}d}{\varepsilon^2}$ for some constant c_{14}, we have $\varepsilon \geq \sqrt{\frac{c_{14}d}{m}}$, so it suffices to require

$$c_{13}q^3(pm)^{\frac{2}{q}}\left(\frac{d}{m}\right)^{\frac{1}{2}} \leq \frac{1}{4}pm\sqrt{\frac{c_{14}d}{m}}$$
$$\text{or,} \quad pm \geq c_{15}(pm)^{\frac{2}{q}}q^3$$

for some $c_{15} > 0$.

Combining the analysis for the two cases, it suffices to require

$$pm \geq (pm)^{\frac{2}{q}} \max\left\{\frac{c_{16}q^2}{\varepsilon}, c_{17}q^3\right\}$$

for some constants $c_{16} > 0$ and $c_{17} > 0$. This requirement is equivalent to

$$pm \geq \left(\max\left\{\frac{c_{16}q^2}{\varepsilon}, c_{17}q^3\right\}\right)^{\frac{1}{1-2/q}},$$

which concludes the proof of Lemma 17 (see remaining details in Section 7 of the technical report). ◀

5.2 Obtaining the differential inequality in Lemma 20

We now discuss the proof of the technical part of our argument in the previous proof, which is to control the derivative of the interpolant. The full proof can be found in Section 7 of the technical report.

Sketch of proof of Lemma 20. There are two main ideas for obtaining this differential inequality. First, we use the cumulant method as in [4] to transform the derivative in t to matrix directional derivatives. Then, we bound the resulting terms in the expression by delicately using the matrix Hölder's inequality.

Fix M_2 and define $f_{1,M_2}(M_1) := \operatorname{tr}(\Gamma(M_1, M_2)^{2q})$ as a function of M_1. We shall first obtain an expression for $\frac{d}{dt}\mathbb{E}[f_{1,M_2}(S_1(t))]$. To see why this is sufficient, note that the derivative we are interested in is the directional derivative along the path $t \to (t,t)$ for the multivariate function $(t_1,t_2) \to \mathbb{E}[\operatorname{tr}(\Gamma(S_1(t_1), S_2(t_2))^{2q})]$ and by the chain rule (as mentioned in Section 4),

$$\frac{d}{dt}\mathbb{E}[\operatorname{tr}\Gamma(t)^{2q}] = \frac{d}{dt_1}\mathbb{E}[\operatorname{tr}(\Gamma(S_1(t_1), S_2(t_2))^{2q})]\Big|_{t_1,t_2=t} + \frac{d}{dt_2}\mathbb{E}[\operatorname{tr}(\Gamma(S_1(t_1), S_2(t_2))^{2q})]\Big|_{t_1,t_2=t}$$

Now, recall that S_1 can be written in the form $\sum_{(l,\gamma)\in\Xi} Z_{(l,\gamma)}$ where $\Xi = [n] \times [pm]$ and $Z_{(l,\gamma)} = \xi_{(l,\gamma)} e_{\mu_{(l,\gamma)}} e_l^\top$ (see Definition 6). We then have the following lemma.

▶ **Lemma** (Based on Corollary 6.1, [4]). *For any polynomial $\phi : M_{m\times d}(\mathbb{R}) \to \mathbb{R}$, we have*

$$\frac{d}{dt}\mathbb{E}[\phi(S_1(t))] = \frac{1}{2}\sum_{k=4}^{\infty}\frac{t^{\frac{k}{2}-1}}{(k-1)!}\sum_{\pi\in\mathrm{P}([k])}(-1)^{|\pi|-1}(|\pi|-1)!\,\mathbb{E}\left[\sum_{(l,\gamma)\in\Xi}\partial_{Z_{(l,\gamma),1|\pi}}\cdots\partial_{Z_{(l,\gamma),k|\pi}}\phi(S_1(t))\right],$$

where $\partial_Z\phi$ denotes the directional derivative of ϕ in the direction $Z \in M_{m\times d}(\mathbb{R})$.

Here, $\mathrm{P}([k])$ denotes the set of all partitions of $[k]$, and $Z_{(l,\gamma),1|\pi}, \ldots, Z_{(l,\gamma),k|\pi}$ are random matrices distributed as $Z_{(l,\gamma)}$. Crucially, those are independent of S_1, G_1, S_2 and G_2 (but not necessarily from each other). Further details are given in the full proof in Section 7 of the technical report.

Applying this lemma to $\frac{d}{dt_1}\mathbb{E}[\operatorname{tr}(\Gamma(S_1(t_1), S_2(t_2))^{2q})] = \frac{d}{dt_1}\mathbb{E}[f_{1,S_2(t_2)}(S_1(t_1))]$, we need to deal with the directional derivatives of $f_{1,S_2(t_2)}$ along $Z_{(l,\gamma),1|\pi}, \ldots, Z_{(l,\gamma),k|\pi}$. Using a general expression for derivatives of multinomials via the product rule, we have, for any deterministic $m \times d$ matrices $B_1, \ldots, B_k, M_1$ and M_2,

$$\begin{aligned}&\partial_{B_1}\cdots\partial_{B_k} f_{1,M_2}(M_1)\\ &= \sum_{\sigma\in\mathrm{sym}(k)} \sum_{\substack{r_1,\ldots,r_{k+1}\geq 0\\ r_1+\cdots+r_{k+1}=2q-k}} \operatorname{tr}\Big(\Gamma(M_1,M_2)^{r_1}((B_{\sigma(1)}U)^T M_2 U + (M_2U)^T B_{\sigma(1)}U)\Gamma(M_1,M_2)^{r_2}\\ &\qquad\qquad ((B_{\sigma(2)}U)^T M_2 U + (M_2U)^T B_{\sigma(2)}U)\cdots\Gamma(M_1,M_2)^{r_k}\\ &\qquad\qquad ((B_{\sigma(k)}U)^T M_2 U + (M_2U)^T B_{\sigma(k)}U)\Gamma(M_1,M_2)^{r_{k+1}}\Big).\end{aligned}$$

In our case, for each fixed (l,γ), we have to analyze $\partial_{Z_{(l,\gamma),1|\pi}}\cdots\partial_{Z_{(l,\gamma),k|\pi}} f_{1,S_2(t_2)}(S_1(t_1))$, which means that we have $B_\lambda = Z_{(l,\gamma),\lambda|\pi}$ for $\lambda\in[k]$, and $M_2 = S_2(t)$. So terms of the form $(B_\lambda U)^T M_2 U$ become $(Z_{(l,\gamma),\lambda|\pi}U)^T S_2(t)U$. Crucially, $(Z_{(l,\gamma),\lambda|\pi}U)^T S_2(t)U$ is a rank one matrix, so it can be written as an outer product of the form $\Theta^T_{(l,\gamma),\lambda,1}\Theta_{(l,\gamma),\lambda,2}$.

Then, estimating

$$\mathbb{E}\left[\sum_{(l,\gamma)\in\Xi} \partial_{Z_{(l,\gamma),1|\pi}}\cdots\partial_{Z_{(l,\gamma),k|\pi}} f_{1,S_2(t_2)}(S_1(t_1))\right]$$

for $t_1 = t_2 = t$ boils down to estimating terms of the form

$$\mathbb{E}\left[\operatorname{tr}\Gamma(t)^{r_1}\Theta^T_{(l,\gamma),\sigma(1),\tau_1(1)}\Theta_{(l,\gamma),\sigma(1),\tau_1(2)}\cdot\Gamma(t)^{r_2}\cdots\Gamma(t)^{r_k}\Theta^T_{(l,\gamma),\sigma(k),\tau_k(1)}\Theta_{(l,\gamma),\sigma(k),\tau_k(2)}\Gamma(t)^{r_{k+1}}\right]$$

where $\tau_i\in\mathrm{sym}(\{1,2\})$ are permutations of the set $\{1,2\}$.

For the remainder of the proof, we delicately analyze terms of this form using the matrix Hölder's inequality, and appropriately estimate the terms that arise.

▶ **Lemma** (Matrix Hölder's inequality, Lemma 5.3 in [4]). *Let $1\leq\beta_1,\ldots,\beta_k\leq\infty$ satisfy $\sum_{i=1}^k\frac{1}{\beta_i}=1$. Then*

$$\left|\mathbb{E}\big[\operatorname{tr} Y_1\cdots Y_k\big]\right| \leq \|Y_1\|_{\beta_1}\cdots\|Y_k\|_{\beta_k}$$

for any $d\times d$ random matrices $Y_1,\ldots,Y_k$.

This analysis based on matrix Hölder's inequality is done in Section 7 of the technical report. One important observation we use in this lemma (among many others) is that $\Theta^T_{(l,\gamma),\lambda,1}\Theta_{(l,\gamma),\lambda,2}$ are rank one matrices, which allows us to bound $\|\Theta^T_{(l,\gamma),\lambda,1}\Theta_{(l,\gamma),\lambda,2}\|_q$ with $\sqrt{pd}$ instead of $\sqrt{pm}$. For further details, please refer to the full proof in Section 7 of the technical report. ◀

6 Conclusions

We give an oblivious subspace embedding with optimal embedding dimension that achieves near-optimal sparsity, thus nearly matching a conjecture of Nelson and Nguyen in terms of the best sparsity attainable by an optimal oblivious subspace embedding. We also propose a fast algorithm for constructing low-distortion subspace embeddings, based on a new family of Leverage Score Sparsified embeddings with Independent Columns (LESS-IC). This new algorithm leads to speedups in downstream applications such as optimization problems based on constrained or regularized least squares. As a by-product of our analysis, we develop a new set of tools for matrix universality, combining a decoupling argument with a two-dimensional interpolation method, which are likely of independent interest.

References

1 Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4):671–687, 2003. doi:10.1016/S0022-0000(03)00025-4.

2 Nir Ailon and Bernard Chazelle. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 39(1):302–322, 2009. doi:10.1137/060673096.

3 Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. In *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing*, pages 499–508, 2015. doi:10.1145/2746539.2746541.

4 Tatiana Brailovskaya and Ramon van Handel. Universality and sharp matrix concentration inequalities. *Geometric and Functional Analysis*, pages 1–105, 2024.

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

6 Shabarish Chenakkod, Michał Dereziński, Xiaoyu Dong, and Mark Rudelson. Optimal embedding dimension for sparse subspace embeddings. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 1106–1117, 2024.

7 Shabarish Chenakkod, Michał Dereziński, and Xiaoyu Dong. Optimal oblivious subspace embeddings with near-optimal sparsity, 2024. doi:10.48550/arXiv.2411.08773.

8 Nadiia Chepurko, Kenneth L Clarkson, Praneeth Kacham, and David P Woodruff. 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)*, pages 3043–3068. SIAM, 2022. doi:10.1137/1.9781611977073.118.

9 Yeshwanth Cherapanamjeri, Sandeep Silwal, David P Woodruff, and Samson Zhou. Optimal algorithms for linear algebra in the current matrix multiplication time. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 4026–4049. SIAM, 2023. doi:10.1137/1.9781611977554.CH154.

10 Kenneth L Clarkson and David P Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the forty-fifth annual ACM symposium on Theory of Computing*, pages 81–90, 2013. doi:10.1145/2488608.2488620.

11 Michael B Cohen. Nearly tight oblivious subspace embeddings by trace inequalities. In *Proc. of the 27th annual ACM-SIAM Symposium on Discrete Algorithms*, pages 278–287. SIAM, 2016. doi:10.1137/1.9781611974331.CH21.

12 Michael B Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 163–172, 2015. doi:10.1145/2746539.2746569.

13 Michael B Cohen, Jelani Nelson, and David P Woodruff. Optimal approximate matrix product in terms of stable rank. In *International Colloquium on Automata, Languages, and Programming*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, 2016.

14 Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse johnson: Lindenstrauss transform. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 341–350, 2010. doi:10.1145/1806689.1806737.

15 Michał Dereziński. Algorithmic gaussianization through sketching: Converting data into sub-gaussian random designs. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 3137–3172. PMLR, 2023.

16 Michał Dereziński, Jonathan Lacotte, Mert Pilanci, and Michael W Mahoney. Newton-less: Sparsification without trade-offs for the sketched newton update. *Advances in Neural Information Processing Systems*, 34:2835–2847, 2021.

17 Michał Dereziński, Zhenyu Liao, Edgar Dobriban, and Michael Mahoney. Sparse sketches with small inversion bias. In *Conference on Learning Theory*, pages 1467–1510. PMLR, 2021.

18 Michał Dereziński and Michael W Mahoney. Recent and upcoming developments in randomized numerical linear algebra for machine learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6470–6479, 2024.

19 Petros Drineas, Malik Magdon-Ismail, Michael W Mahoney, and David P Woodruff. Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research*, 13(1):3475–3506, 2012. doi:10.5555/2503308.2503352.

20 Petros Drineas and Michael W Mahoney. Randnla: randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016. doi:10.1145/2842602.

21 Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136, 2006.

22 Daniel M Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):1–23, 2014. doi:10.1145/2559902.

23 Yi Li and Mingmou Liu. Lower bounds for sparse oblivious subspace embeddings. In *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 251–260, 2022. doi:10.1145/3517804.3526224.

24 Per-Gunnar Martinsson and Joel A Tropp. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020. doi:10.1017/S0962492920000021.

25 Xiangrui Meng and Michael W Mahoney. 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*, pages 91–100, 2013. doi:10.1145/2488608.2488621.

26 Jelani Nelson and Huy L Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 ieee 54th annual symposium on foundations of computer science*, pages 117–126. IEEE, 2013.

27 Jelani Nelson and Huy L Nguyên. Lower bounds for oblivious subspace embeddings. In *International Colloquium on Automata, Languages, and Programming*, pages 883–894. Springer, 2014. doi:10.1007/978-3-662-43948-7_73.

28 Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th annual IEEE symposium on foundations of computer science (FOCS'06)*, pages 143–152. IEEE, 2006.

29 Joel A Tropp. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, 3(01n02):115–126, 2011. doi:10.1142/S1793536911000787.

30 Joel A. Tropp. An introduction to matrix concentration inequalities. *Found. Trends Mach. Learn.*, 8(1–2):1–230, May 2015. doi:10.1561/2200000048.

31 Joel A Tropp. Comparison theorems for the minimum eigenvalue of a random positive-semidefinite matrix. *arXiv preprint arXiv:2501.16578*, 2025. doi:10.48550/arXiv.2501.16578.

32 David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014. doi:10.1561/0400000060.