OPTIMAL OBLIVIOUS SUBSPACE EMBEDDINGS WITH NEAR-OPTIMAL SPARSITY

SHABARISH CHENAKKOD, MICHAŁ DEREZIŃSKI, AND XIAOYU DONG

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 traces inequalities that leverage the specific structure of our subspace embedding construction.

University of Michigan, Ann Arbor, MI, USA
National University of Singapore, Singapore
E-mail addresses: shabari@umich.edu, derezin@umich.edu, xdong@nus.edu.sg.

Partially supported by DMS 2054408 and CCF 2338655. The authors are very 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 [1], low-rank approximation [2], clustering [3], and many more (see [4] 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 1.1. 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

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

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 [2, 5–9], aimed towards resolving the following conjecture of Nelson and Nguyen [6], which is supported by nearly-matching lower bounds [10, 11].

Conjecture 1.2 (Nelson and Nguyen, FOCS 2013 [6])**.** *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 [7, 8] improved on this; most notably, Cohen [8] 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 [12].

This sub-optimality in dimension m was finally addressed in a recent work of Chenakkod, Dereziński, Dong and Rudelson [9], relying on a breakthrough in random matrix universality theory by Brailovskaya and van Handel [13]. 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 [13] (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 1.2 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 1.3 (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 [4] 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 [14]. 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 [15], 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 [16–18].

Just like recent prior works [9, 19, 20], 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 [9].

Theorem 1.4 (Fast Subspace Embedding). *Given $A\in\mathbb{R}^{n\times d}$, $\varepsilon, \gamma\in(0, 1)$ and $1/\varepsilon\leq\mathrm{poly}(d)$, 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

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

This is a direct improvement over the previous best known runtime for constructing an optimal subspace embedding [9], 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 [19, 20] 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 [7]. The following corollary follows immediately from Theorem 1.4, and is a direct improvement over Theorem 1.8 of [9] 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 1.5 (Fast reduction for constrained least squares). *Given $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, $\epsilon > 0$, function $g : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ and set $\mathcal{C} \subseteq \mathbb{R}^d$ consider an $n \times d$ problem $\mathrm{LS}_{\mathcal{C},g}(A, b, \epsilon)$:*

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

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 [1] (for an overview, see the following surveys and monographs [4, 21–

23]). Initially, these approaches focused on leveraging fast Hadamard transforms [24, 25] to achieve improved time complexity for linear algebraic tasks such as linear regression and low-rank approximation. Clarkson and Woodruff [2] 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 [26, 27], 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 [2, 5–9], some works have aimed to develop fast subspace embeddings that achieve optimal embedding dimension either without sparsity [19, 20], under additional assumptions [28], or with one-sided embedding bounds [29]. Our time complexity result, Theorem 1.4, 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 1.1, 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.,

$$\Pr(1-\varepsilon \le s_{\min}(\Pi U) \le s_{\max}(\Pi U) \le 1+\varepsilon) \ge 1-\delta, \tag{3.1}$$

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 (3.1) 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 [6]. As we explain later, our analysis techniques apply to other natural families of sparse embedding distributions, including those with i.i.d. entries [30], however the OSNAP-style construction is crucial for achieving the near-optimal sparsity $s = \widetilde{O}(1/\varepsilon)$.

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 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

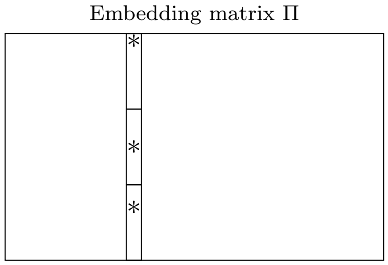

FIGURE 1. An example of a column divided into $s = 3$ subcolumns with each subcolumn having exactly one non-zero entry in a random position.

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 3.1 (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 3.2 (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 $c_{3.2.1}$ and $c_{3.2.2}$ 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 $m \ge c_{3.2.1}(d + \log(1/\delta\varepsilon))/\varepsilon^2$ and the sparsity $s = pm$ satisfies $s \ge \min\{c_{3.2.2}(\log^2(\frac{d}{\varepsilon\delta})/\varepsilon + \log^3(\frac{d}{\varepsilon\delta})), m\}$ non-zeros per column.

Remark 3.3. We note that if $1/\varepsilon$ is polynomial in d/δ, i.e., $\varepsilon \ge \frac{1}{(d/\delta)^K}$ for some absolute constant $K \ge 1$, 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

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

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:

$$m \geq 2c_{3.2.1}\frac{d+\log(1/\delta)}{\varepsilon^2}.$$

This is a direct improvement over Theorem 1.2 of [9], 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 [30], which was also considered by [9]. 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 for details; this is again a direct improvement over a result of [9]).

Remark 3.4. 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 $j \leq d$,

$$\big|\|\Pi e_j\|_2^2 - 1\big| = \left|\tfrac{N_j}{pm} - 1\right| \leq \varepsilon, \qquad N_j := \mathrm{nnz}(Se_j) \sim \mathrm{Binomial}(m,p). \tag{3.2}$$

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 3.2 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

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

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 3.2 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 [31] as an extension of the corresponding moment-based characterization of a Johnson-Lindenstrauss embedding [27].

Definition 3.5. 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,

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

Note that a simple application of Markov's inequality recovers the guarantee in Definition 1.1 from the $(\varepsilon, \delta, d, \ell)$-OSE moments property with any $\ell \geq 1$. Moreover, [31] 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 3.6. *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 $m \geq c_{3.6.1}(d + \log(1/\delta\varepsilon))/\varepsilon^2$ and $s \geq \min\{c_{3.6.2}(\log^2(\frac{d}{\varepsilon\delta})/\varepsilon + \log^3(\frac{d}{\varepsilon\delta})), m\}$.*

Remark 3.7. 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 [31, Remark 3], such runtimes can be further refined by chaining together several embeddings with an OSE moment property. For example, [8] 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 [9] has dealt with one such family of distributions termed LESS embeddings [16–18], 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 [9]. 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 $O(\mathrm{nnz}(A))$, plus an additional sublinear term. Given that the preprocessing step of approximating the leverage scores has to take at least $\mathrm{nnz}(A)$ 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 3.8 (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 $1 \le i \le n$,

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

We say that the numbers $(l_1, \dots, l_n) \in [0,1]^n$ 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 [9], based on a similar construction first proposed by [16]. Here, we once again start by defining an unscaled matrix S, which is then normalized to obtain the subspace embedding matrix Π.

Definition 3.9 (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). 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 5th 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 for the full definition).

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 3.10 (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 $c_{3.10.1}$ and $c_{3.10.2}$ 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 $m \ge c_{3.10.1}\left(\frac{d + \log^2(d/\delta) + \log(1/\varepsilon)}{\varepsilon^2} + \log^3(d/\delta)/\varepsilon\right)$ and

$$c_{3.10.2} \max\left\{\frac{(\log(d/\varepsilon\delta))^{2.5}}{\varepsilon}, (\log(d/\varepsilon\delta))^3\right\} \le pm \le m.$$

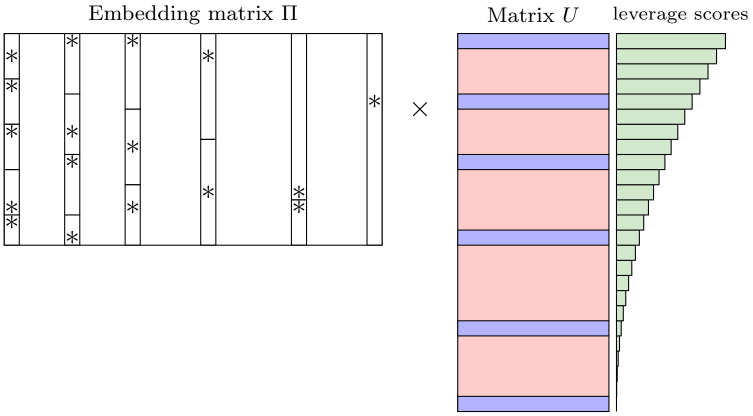

FIGURE 2. In the LESS-IC distribution, column j is filled with s_j many subcolumns, with the bottom-most subcolumn truncated to fit the size of Π. Each subcolumn has one non-zero entry. Notice that as the leverage scores decrease, the number of subcolumns decreases and the matrix becomes sparser. However,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

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 3.11. 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, [9] 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 1.4).** Here, we briefly outline how our LESS-IC embedding yields a fast subspace embedding construction to recover the time complexity claimed in Theorem 1.4. This follows analogously to the construction from Theorem 1.6 of [9], 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 [9] (adapted from Lemma 7.2 in [19]), we can construct coarse approximations of all leverage scores so that $\beta_1 = O(n^\gamma)$ and $\beta_2 = O(1)$ in time $O(\gamma^{-1}(\mathrm{nnz}(A)+d^2)+d^\omega)$. Applying LESS-IC (Theorem 3.10) with these leverage scores and parameters β_1,β_2, computing ΠA takes $O(\mathrm{nnz}(A)+n^\gamma d^2\log^3(d/\varepsilon\delta)/\varepsilon)$, where $\mathrm{nnz}(A)$ 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), $\mathrm{nnz}(A)\geq n$ (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 $n^\gamma d^2/\varepsilon\leq n^{0.5+\gamma+2c}\leq n^{0.8}\ll\mathrm{nnz}(A)$, and the time complexity is dominated by the $O(\gamma^{-1}\,\mathrm{nnz}(A))$ term.

Finally, we note that Corollary 1.5 follows simply by constructing a subspace embedding Π via Theorem 1.4 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 [9]. 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 3.2 and Theorem 3.10. Section 5 provides a sketch of the proof of Theorem 3.2, listing the main technical steps, leaving the full proof with all technical details to Section 7. The proof of Theorem 3.10 follows similarly and is covered in Section 8. The subspace embedding guarantee for a sparse matrix with independent entries is proved in Section 9. Section 6 contains some basic facts from the existing literature that are used throughout the paper.

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 $\mathrm{P}([n])$ 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 $\mathrm{Tr}(X)$ denotes the trace of the matrix X, and $\mathrm{tr}(X) = \frac{1}{d}\mathrm{Tr}(X)$ 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 $\mathrm{spec}(X)$. 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 [13] 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 $\mathrm{Cov}(X, Y)$. 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 3.2 and 3.10. 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 3.1).

Note that due to the equivalent characterization of the OSE property in (3.1), 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 [13] 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, [9] achieved optimal embedding dimension $m = O(d/\varepsilon^2)$ for OSEs by using the bound in [13, 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, [9] 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 $(\sqrt{pm})^{1/2} \leq \sqrt{pm}\varepsilon$ 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 $\|(\Pi U)^T \Pi U - I_d\| \leq \varepsilon$ or $\|(SU)^T SU - pm \cdot I_d\| \leq pm\varepsilon$ (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 3.2 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 Lemma 6.8) 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 $\mathbb{E}\left[\mathrm{tr}\left(X\right)^{2q}\right]^{\frac{1}{2q}} = \left(\mathbb{E}\,\mathrm{tr}\left(X\right)^{2q}\right)^{1/(2q)}$.

Now, [13, 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 $(SU)^T(SU) - pmI_d = \sum_{i=1}^m (U^T s_i s_i^T U - pI_d)$, 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 [13, 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 [13]. 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 7.3 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.

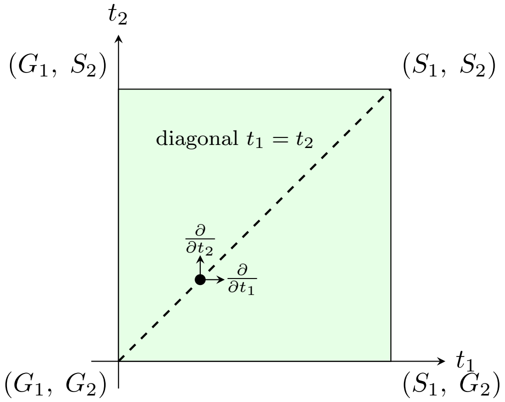

FIGURE 3. Two-dimensional interpolation in $(t_1, t_2) \in [0,1]^2$, decomposed using the chain rule.

By standard calculations, we have $\mathbb{E}[\mathrm{tr}(G_1^T G_2 + G_2^T G_1)^{2q}]^{\frac{1}{2q}} \le c\sqrt{pm}\sqrt{pd} = O(pm\varepsilon)$ 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:

$$\left|\mathbb{E}\big[\,\mathrm{tr}((S_1U)^T(S_2U) + (S_2U)^T(S_1U))^{2q}\big]^{\frac{1}{2q}} - \mathbb{E}\big[\,\mathrm{tr}(G_1^T G_2 + G_2^T G_1)^{2q}\big]^{\frac{1}{2q}}\right| \le \mathrm{polylog}(\tfrac{d}{\varepsilon\delta}). \tag{4.1}$$

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 $pm \ge \mathrm{polylog}(\frac{d}{\varepsilon\delta})/\varepsilon$, achieving the conjectured $1/\varepsilon$ dependence.

Nevertheless, the conditioning argument cannot be done directly because

$$\mathbb{E}\Big[\,\mathrm{tr}\left((S_1U)^T(S_2U) + (S_2U)^T(S_1U)\right)^{2q}\Big]^{\frac{1}{2q}} \ne \mathbb{E}_{S_1}\Big[\mathbb{E}_{S_2}\Big[\,\mathrm{tr}\left((S_1U)^T(S_2U) + (S_2U)^T(S_1U)\right)^{2q}\Big]^{\frac{1}{2q}}\Big].$$

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 $f(M_1, M_2) = \mathrm{tr}((M_1U)^T(M_2U) + (M_2U)^T(M_1U))^{2q}$, 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 [13] to get a differential inequality (Lemma 7.4) that leads to inequality (4.1). 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 3.2 for OSNAP. The full proof can be found in Section 7. The proof of the subspace embedding guarantee for LESS-IC, Theorem 3.10 is similar and can be found in Section 8.

Proof. TOPROVE 0 □

Finally, we show how the above arguments also imply the OSE moment property (Definition 3.5).

Proof. TOPROVE 1 □

5.1. **Controlling Trace Moments of the Embedding Error.** We now sketch the proof of Lemma 7.3, which obtains the moment bound for $X^TX - I$ used in the previous proof. The full proof can be found in Section 7.3.

Proof. TOPROVE 2 □

5.2. **Obtaining the differential inequality in Lemma 7.4.** 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 7.4.

Proof. TOPROVE 3 □

6. Preliminaries

6.1. **Oblivious Subspace Embeddings.** Here, we prove some important properties of the OSNAP distribution that we shall use later.

Lemma 6.1 (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 as in the unscaled OSNAP distribution. 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. TOPROVE 4 □

Lemma 6.2 (Norm of a Random Row in Interpolated OSNAP)**.** *Let $S(t) := \sqrt{t}S + \sqrt{1-t}G$, where S is as in the fully independent unscaled OSNAP distribution and G is an $m \times n$ matrix with i.i.d. Gausian entries with variance p. Let U be an $n \times d$ matrix such that $U^TU = I$. Let μ be a random variable uniformly distributed in $J \subset [m]$ and independent of S and G. Then, there exists $c_{6.2} > 0$ such that for any positive integer $q > 0$, we have*

$$\mathbb{E}_{\mu,S(t)}[\|e_\mu^T S(t)U\|^q]^{\frac{1}{q}} \leq c_{6.2}\sqrt{\max\{pd, q\}}$$

Proof. TOPROVE 5 □

6.2. **Basic Facts of the $L_q(S_q^d)$ Space.** To derive the trace inequalities that will be used in the interpolation argument, we need the following tools from [13]. For a $d \times d$ matrix M, following [13], we define the absolute value $|M| = \sqrt{M^*M}$ and normalized trace $\mathrm{tr}(M) = \frac{1}{d}\mathrm{Tr}(M)$. Let $L_q(S_q^d)$ be the normed vector space of $d \times d$ random matrices M with norm

$$\|M\|_q = \begin{cases} \left(\mathbb{E}[\mathrm{tr}\,|M|^q]\right)^{\frac{1}{q}} & \text{if } 1 \leq q < \infty \\ \|\|M\|_{op}\|_\infty & \text{if } q = \infty \end{cases}$$

More precisely, the space $L_q(S_q^d)$ consists of random matrices M with $\|M\|_q$ well defined (which means $\left(\mathbb{E}[\mathrm{tr}\,|M|^q]\right) < \infty$ for $1 \leq q < \infty$ and $\|\|M\|_{op}\|_\infty$ for $q = \infty$).

Next, we state a Hölder inequality for Schatten classes proved in [13].

Lemma 6.3 (Lemma 5.3. in [13]). *Let $1 \le \beta_1, \dots, \beta_k \le \infty$ satisfy $\sum_{i=1}^k \frac{1}{\beta_i} = 1$. Then*

$$|\,\mathbb{E}[\operatorname{tr} Y_1 \cdots Y_k]| \le \|Y_1\|_{\beta_1} \cdots \|Y_k\|_{\beta_k}$$

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

The proof of Lemma 6.3 (which we do not include here) relies on convexity and interpolation in $L_q(S_q^d)$. More precisely, one can first prove the result for the case when $\beta_1, ..., \beta_k$ are extreme exponents and then extend the result to general $\beta_1, ..., \beta_k$ by the following convexity result. Later we will also use this method to prove our trace inequalities (Lemma 7.6).

Lemma 6.4 (Lemma 5.2. in [13]). *Let $F : (L_\infty(S_\infty^d))^k \to \mathbb{C}$ be a multilinear functional. Then the map*

$$\left(\frac{1}{\beta_1}, \dots, \frac{1}{\beta_k}\right) \mapsto \log \sup_{M_1, \dots, M_k} \frac{|F(M_1, \dots, M_k)|}{\|M_1\|_{\beta_1} \cdots \|M_k\|_{\beta_k}}$$

is convex on $[0,1]^k$.

More generally, we have the following complex interpolation result from [34].

Theorem 6.5 (4.4.1 Theorem in [34]). *Let $\{(A_{(\nu,0)}, A_{(\nu,0)})\}_{\nu=1}^n$ and (B_0, B_1) be compatible Banach spaces. Assume that $T : (A_{(1,0)} \cap A_{(1,0)}) \oplus \cdots \oplus (A_{(n,0)} \cap A_{(n,0)}) \to (B_0 \cap B_1)$ is multilinear. Also, assume that for any $(a_1, ..., a_n) \in (A_{(1,0)} \cap A_{(1,0)}) \oplus \cdots \oplus (A_{(n,0)} \cap A_{(n,0)})$, we have*

$$\|T(a_1, ..., a_n)\|_{B_0} \le M_0 \prod_{\nu=1}^n \|a_\nu\|_{A_{\nu,0}}$$

and

$$\|T(a_1, ..., a_n)\|_{B_1} \le M_1 \prod_{\nu=1}^n \|a_\nu\|_{A_{\nu,1}}$$

Then for any $0 \le \theta \le 1$, the function T may be uniquely extended to a multilinear mapping from $(A_{(1,0)}, A_{(1,0)})_{[\theta]} \oplus \cdots \oplus (A_{(n,0)}, A_{(n,0)})_{[\theta]}$ to $(B_0, B_1)_{[\theta]}$ with norm at most $M_0^{1-\theta} M_1^\theta$, where $(A_{(\nu,0)}, A_{(\nu,0)})_{[\theta]}$ is the complex interpolation space of the pair $(A_{(\nu,0)}, A_{(\nu,0)})$ with exponent θ defined in [34, p.88].

Applying Theorem 6.5 to $L_q(S_q^d)$ spaces, we have a more general version of Lemma 6.4.

Corollary 6.6 (Interpolation in $L_q(S_q^d)$ Space). *Let*

$$(\frac{1}{\beta_1(0)}, ..., \frac{1}{\beta_k(0)}) \in [0,1]^k \text{ and } (\frac{1}{\beta_1(1)}, ..., \frac{1}{\beta_k(1)}) \in [0,1]^k$$

Assume that

$$F : (L_{\beta_1(0)}(S_{\beta_1(0)}) \cap L_{\beta_1(1)}(S_{\beta_1(1)})) \oplus \cdots \oplus (L_{\beta_k(0)}(S_{\beta_k(0)}) \cap L_{\beta_k(1)}(S_{\beta_k(1)})) \to \mathbb{R}$$

is multilinear with

$$F(M_1, ..., M_k) \le K \prod_{\nu=1}^n \|M_\nu\|_{\beta_\nu(0)}$$

and

$$F(M_1, ..., M_k) \le K \prod_{\nu=1}^n \|M_\nu\|_{\beta_\nu(1)}$$

for all $(M_1, ..., M_k) \in (L_{\beta_1(0)}(S_{\beta_1(0)}) \cap L_{\beta_1(1)}(S_{\beta_1(1)})) \oplus \cdots \oplus (L_{\beta_k(0)}(S_{\beta_1(0)}) \cap L_{\beta_k(1)}(S_{\beta_1(1)}))$*. Define* $\beta_\nu(\theta)$ *for* $\nu = 1, 2, ..., k$ *and* $0 \leq \theta \leq 1$ *such that*

$$\frac{1}{\beta_1(\theta)} = (1-\theta)\frac{1}{\beta_1(0)} + \theta\frac{1}{\beta_1(1)}$$

Then for any $0 \leq \theta \leq 1$*, the multilinear functional* F *can be uniquely extended to*

$$L_{\beta_1(\theta)}(S_{\beta_1(\theta)}) \oplus \cdots \oplus L_{\beta_k(\theta)}(S_{\beta_1(\theta)})$$

with

$$F(M_1, ..., M_k) \leq K \prod_{\nu=1}^{n} \|M_\nu\|_{\beta_\nu(\theta)}$$

for all $(M_1, ..., M_k) \in L_{\beta_1(\theta)}(S_{\beta_1(\theta)}) \oplus \cdots \oplus L_{\beta_k(\theta)}(S_{\beta_1(\theta)})$*.*

Proof. TOPROVE 6 □

6.3. **Spectrum of Gaussian Matrices.** Here we collect results about the spectrum of various Gaussian models that will be used later in conjunction with universality results.

Lemma 6.7 ((2.3), [36])**.** *For* $m > d$*, let* G *be an* $m \times d$ *matrix whose entries are independent standard normal variables. Then,*

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

Corollary 6.8 (Trace Moment of Embedding Error for Gaussian Model)**.** *Let* G *be an* $m \times d$ *matrix whose entries are independent normal random variables with variance* $\frac{1}{m}$*. Let* $\varepsilon < \frac{1}{6}$ *and* $q \in \mathbb{N} \leq m\varepsilon^2$*. Then, there exists* $c_{6.8} > 1$ *such that for* $m \geq \frac{c_{6.8}d}{\varepsilon^2}$*,*

$$\mathbb{E}[\mathrm{tr}(G^TG - I_d)^{2q}]^{\frac{1}{2q}} \leq \varepsilon$$

Proof. TOPROVE 7 □

Lemma 6.9 (Trace Moment of Embedding Error for Decoupled Gaussian Model)**.** *Let* G_1 *and* G_2 *be independent* $m \times d$ *random matrices with i.i.d. Gaussian entries. Then for any positive integer* q*, there exists* $c_{6.9} > 0$ *such that*

$$\mathbb{E}\left[\mathrm{tr}\left(G_1^TG_2 + G_2^TG_1\right)^{2q}\right]^{\frac{1}{2q}} \leq c_{6.9}\sqrt{\max\{d, q\}}\sqrt{\max\{m, q\}}$$

Proof. TOPROVE 8 □

7. Full proof of Theorem 3.2

This section contains the full proof of Theorem 3.2 along with the various lemmas that are used along the way.

- Section 7.1 proves the final subspace embedding guarantee using a bound on the trace moments of the embedding error.
- Section 7.2 has details about the decoupling step that reduces the problem of controlling moments of $(SU)^TSU - pm \cdot I_d$ to controlling moments of $(S_1U)^TS_2U + (S_2U)^TS_1U$, for independent S_1 and S_2.
- Section 7.3 proves the trace moment bound using 2D interpolation of moments of the form $(S_1(t)U)^TS_2(t)U + (S_2(t)U)^TS_1(t)U$.
- Section 7.4 obtains the differential inequality for the derivative of the interpolant.
- Section 7.5 proves the trace inequality required for obtaining the differential inequality.

7.1. **Proving the subspace embedding guarantee for OSNAP.** In this section we give the full proof of Theorem 3.2.

Theorem 3.2 (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 $c_{3.2.1}$ and $c_{3.2.2}$ such that for any $0 < \delta, \varepsilon < 1$ and $d > 10$, we have*

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

if the embedding dimension satisfies $m \geq c_{3.2.1}(d + \log(1/\delta\varepsilon))/\varepsilon^2$ and the sparsity $s = pm$ satisfies $s \geq \min\{c_{3.2.2}(\log^2(\frac{d}{\varepsilon\delta})/\varepsilon + \log^3(\frac{d}{\varepsilon\delta})), m\}$ non-zeros per column.

Proof. TOPROVE 9 □

7.2. **Proving the decoupling lemma for OSNAP.** Let S have the fully independent unscaled OSNAP distribution as described in Definition 3.1. Then, recall that,

$$\begin{aligned} S &= \sum_{l=1}^{n}\sum_{\gamma=1}^{pm} \xi_{l,\gamma} e_{\mu_{(l,\gamma)}} e_l^T \\ &=: \sum_{l=1}^{n}\sum_{\gamma=1}^{pm} Z_{l,\gamma} \end{aligned}$$

where $\{\xi_{l,\gamma}\}_{l\in[n],\gamma\in[s]}$ is a collection of independent Rademacher random variables, $\{\mu_{l,\gamma}\}_{l\in[n],\gamma\in[s]}$ is a collection of independent random variables such that each $\mu_{l,\gamma}$ is uniformly distributed in $[(m/s)(\gamma-1)+1 : (m/s)\gamma]$ and $e_{\mu_{(l,\gamma)}}$ and e_l represent basis vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively.

Recalling that our goal is to look at the moments of $(SU)^T(SU) - pmI_d$, we observe that,

$$U^T S^T SU - pm \cdot I_d = \left(\sum_{i=1}^{m} U^T s_i s_i^T U\right) - pm \cdot I_d$$

where s_i^T denotes the i^{th} row of S. Then, letting s_{ij} denote the entries of S,

$$\begin{aligned} U^T S^T SU - pm \cdot I_d &= \left(\sum_{i=1}^{m} U^T \left(\sum_{j=1}^{n} s_{ij} e_j\right)\left(\sum_{j'=1}^{n} s_{ij'} e_{j'}^T\right) U\right) - pm \cdot I_d \\ &= \sum_{i=1}^{m}\left(\sum_{j=1}^{n} s_{ij} u_j\right)\left(\sum_{j'=1}^{n} s_{ij'} u_{j'}^T\right) - pm \cdot I_d \end{aligned}$$

where u_j^T denotes the j^{th} row of U. Separating the cases where $j = j'$ and $j \neq j'$,

$$\begin{aligned} U^T S^T SU - pm \cdot I_d &= \sum_{i=1}^{m}\sum_{j=1}^{n} s_{ij}^2 u_j u_j^T - pm \cdot I_d + \sum_{i=1}^{m}\sum_{\substack{j,j'=1\\ j\neq j'}}^{n} s_{ij}s_{ij'} u_j u_{j'}^T \\ &= \sum_{j=1}^{n}\left(\sum_{i=1}^{m} s_{ij}^2\right) u_j u_j^T - pm \cdot I_d + \sum_{i=1}^{m}\sum_{\substack{j,j'=1\\ j\neq j'}}^{n} s_{ij}s_{ij'} u_j u_{j'}^T \end{aligned} \tag{7.1}$$

Remark 7.1. In Equation (7.1), we decomposed the the embedding error $U^TS^TSU - pm \cdot I_d$ into two parts, the diagonal term

$$\sum_{j=1}^{n}\left(\sum_{i=1}^{m} s_{ij}^2\right) u_j u_j^T - pm \cdot I_d$$

and the off-diagonal term

$$\sum_{i=1}^{m}\sum_{\substack{j,j'=1\\ j\neq j'}}^{n} s_{ij}s_{ij'}u_j u_{j'}^T$$

This decomposition is the key to understand why OSNAP model only needs $\tilde{O}(\frac{1}{\varepsilon})$ nonzero entries per column but the i.i.d. entries model might require $\tilde{O}(\frac{1}{\varepsilon^2})$ nonzero entries per column. In fact, as we will see very soon, the key difference between the OSNAP model and the i.i.d. entries model is that the diagonal term vanishes in OSNAP model but does not vanish in the i.i.d. entries model.

By construction, $\sum_{i=1}^{m} s_{ij}^2 = pm$, so the diagonal term becomes

$$\sum_{j=1}^{n}\left(\sum_{i=1}^{m} s_{ij}^2\right) u_j u_j^T - pm \cdot I_d = pm\left(\sum_{j=1}^{n} u_j u_j^T - I_d\right) = 0$$

and therefore we have

$$\begin{aligned} U^TS^TSU - pm \cdot I_d &= pm\left(\sum_{j=1}^{n} u_j u_j^T - I_d\right) + \sum_{i=1}^{m}\sum_{\substack{j,j'=1\\ j\neq j'}}^{n} s_{ij}s_{ij'}u_j u_{j'}^T \\ &= \sum_{i=1}^{m}\sum_{\substack{j,j'=1\\ j\neq j'}}^{n} s_{ij}s_{ij'}u_j u_{j'}^T \end{aligned}$$

To analyze the off-diagonal term, we use the standard technique of decoupling,

Lemma 7.2 (Decoupling)**.** *When S has the fully independent unscaled OSNAP distribution, we have*

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

Consequently, we have

$$\mathbb{E}[\operatorname{tr}(U^TS^TSU - pm \cdot I_d)^{2q}] \leq \mathbb{E}_{S,S'}\left[\operatorname{tr}\left(2\big((SU)^TS'U + (SU)^TS'U\big)\right)^{2q}\right]$$

where S' is an independent copy of S.

Proof. TOPROVE 10 □

7.3. **Controlling the trace moments of the embedding error for OSNAP.** In this section we give the full proof of Lemma 7.3.

Lemma 7.3 (Trace Moments of Embedding Error for OSNAP)**.** *Let S be an $m\times n$ matrix distributed according to the fully independent unscaled OSNAP distribution with parameter $p \leq 1$. Let U be an arbitrary $n \times d$ deterministic matrix such that $U^TU = I$. Define $X = \frac{1}{\sqrt{pm}}SU$. Then, there exist* *constants $c_{7.3.1}, c_{7.3.2}, c_{7.3.3} > 0$ such that for $q \in \mathbb{N}$ satisfying $2 \le q \le m$, when $m \ge c_{7.3.1}\frac{d+q}{\varepsilon^2}$ and $pm \ge (\max\{\frac{c_{7.3.2}q^2}{\varepsilon}, c_{7.3.3}q^3\})^{1+\frac{2}{q-2}}$, we have*

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

Proof. TOPROVE 11 □

7.4. **Differential inequality for the derivative of the interpolant.** In this section we give the full proof of Lemma 7.4.

Lemma 7.4 (Differential Inequality)**.** *Let S_1 and S_2 be independent random matrices such that either both S_1 and S_2 have the fully independent unscaled OSNAP distribution with parameter p or both S_1 and S_2 have the unscaled OSE-IE distribution with parameter p. Let G_1 and G_2 be independent random matrices with i.i.d. Guassian entries each with variance p, and define the interpolated random matrices,*

$$\begin{aligned} S_1(t) &= \sqrt{t}S_1 + \sqrt{1-t}G_1 \\ S_2(t) &= \sqrt{t}S_2 + \sqrt{1-t}G_2 \end{aligned} \tag{7.2}$$

Let $f(M_1, M_2) = \mathrm{tr}(((M_1U)^T(M_2U) + (M_2U)^T(M_1U))^{2q})$. Then, there exists a constant $c_{7.4}$ such that, for any $q \ge 2$, we have

$$\frac{d}{dt}\mathbb{E}[f(S_1(t), S_2(t))] \le \max_{4\le k\le 2q}(c_{7.4}q)^k((pm)^{\frac{1}{q}}\sqrt{\max\{pd, q\}})^{\frac{qk-2q}{q-1}}(\mathbb{E}[f(S_1(t), S_2(t))])^{1-\frac{k-2}{2q-2}}$$

Remark 7.5. The proof of Lemma 7.4 relies on a technical trace inequality, Lemma 7.6. To illustrate the main idea, we present the proof of Lemma 7.4 using Lemma 7.6 here first, and then prove Lemma 7.6 in the next section.

Proof. TOPROVE 12 □

7.5. **Proving the trace inequality needed to obtain the differential inequality for the derivative of the interpolant.** In this section, we explain the following trace inequality result which is the key to prove the bound (**??**) in the proof of Lemma 7.4.

Lemma 7.6 (Trace Inequalities for OSNAP)**.** *Let $S_1(t)$ and $S_2(t)$ be as in Lemma 7.4 with both having the fully independent unscaled OSNAP distribution. Let*

$$\Gamma(t) = (S_1(t)U)^T(S_2(t)U) + (S_2(t)U)^T(S_1(t)U)$$

Let $\mathcal{Z} = \{\xi_{(l,\gamma)} : (l,\gamma) \in [n] \times [pm]\} \cup \{\mu_{(l,\gamma)} : (l,\gamma) \in [n] \times [pm]\}$ be the family of mutually independent random variables generating an instance of S_1 with the fully independent unscaled OSNAP distribution. Let $q \ge 2$ and $3 \le k \le 2q$. Let $\{\mathcal{Z}_\lambda\}_{\lambda\in[k]}$ be a family of (possibly dependent) random elements, where for each $\lambda \in [k]$, the random element

$$\mathcal{Z}_\lambda = \{\xi_{(l,\gamma),\lambda} : (l,\gamma) \in [n] \times [pm]\} \cup \{\mu_{(l,\gamma),\lambda} : (l,\gamma) \in [n] \times [pm]\}$$

has the same distribution as

$$\mathcal{Z} = \{\xi_{(l,\gamma)} : (l,\gamma) \in [n] \times [pm]\} \cup \{\mu_{(l,\gamma)} : (l,\gamma) \in [n] \times [pm]\}$$

Let $Z_{(l,\gamma)} = \xi_{(l,\gamma)}e_{\mu_{(l,\gamma)}}e_l^T$ and $Z_{(l,\gamma),\lambda} = \xi_{(l,\gamma),\lambda}e_{\mu_{(l,\gamma),\lambda}}e_l^T$. Let $\{\Upsilon_1, ..., \Upsilon_k\}$ be a family of $L_\infty(S_\infty^d)$ random matrices. Assume further that the collection $\{\mathcal{Z}_\lambda\}_{\lambda\in[k]}$ is independent of S_1, S_2, G_1, G_2, and $\{\Upsilon_1, ..., \Upsilon_k\}$. (In other words, $\{\Upsilon_1, ..., \Upsilon_k\}$ can possibly be dependent with S_1, S_2, G_1, G_2.) For each $(l,\gamma) \in [n] \times [pm]$ and $\lambda \in k$, we define random vectors $\Theta_{(l,\gamma),\lambda,1}, \Theta_{(l,\gamma),\lambda,2} \in \mathbb{R}^d$ such that

$$\Theta_{(l,\gamma),\lambda,1} = \xi_{(l,\gamma),\lambda}u_l^T \text{ and } \Theta_{(l,\gamma),\lambda,2} = e^T_{\mu_{(l,\gamma),\lambda}}S_2(t)U$$

where $e_{\mu_{(l,\gamma),\lambda}}$ *represents the* $\mu_{(l,\gamma),\lambda}$*th coordinate vector. Then, given* $0 \leq \beta_1, ..., \beta_k \leq +\infty$ *such that* $\sum_{\lambda=1}^{k} \frac{1}{\beta_\lambda} = 1 - \frac{k}{2q}$, $\tau_1, \dots, \tau_k \in \mathrm{sym}(\{1,2\})$, *there exists* $c_{7.6} > 0$ *such that*

$$\begin{aligned}\sum_{(l,\gamma)\in[n]\times[pm]} & \mathbb{E}[\operatorname{tr} \Theta^T_{(l,\gamma),1,\tau_1(1)}\Theta_{(l,\gamma),1,\tau_1(2)}\Upsilon_1\Theta^T_{(l,\gamma),2,\tau_2(1)}\Theta_{(l,\gamma),2,\tau_2(2)} \\ & \cdots \Upsilon_2\Theta^T_{(l,\gamma),k,\tau_k(1)}\Theta_{(l,\gamma),k,\tau_k(2)}\Upsilon_k] \\ \leq & (c_{7.6}(pm)^{\frac{1}{q}}\sqrt{\max\{pd,q\}})^{\frac{2qk-4q}{2q-2}}(\mathbb{E}\operatorname{tr}((\Gamma(t))^{2q}))^{\frac{1}{q}\cdot\frac{2q-k}{2q-2}}\prod_{\lambda=1}^{k}\|\Upsilon_\lambda\|_{\beta_\lambda}\end{aligned} \tag{7.3}$$

Remark 7.7. The main idea of this lemma is simple. We first use matrix Holder inequality to transform the left hand side into smaller factors, and then we bound those small factors separately. Following this idea, there could be different variants of this lemma. However, not all of them will eventually lead to the optimal dependency of the sparsity on ε. We will explain the idea on how to choose the correct bound. As explained in Proposition 7.3, the sparsity requirement comes from the condition

$$(\mathbb{E}[\operatorname{tr}\Gamma(1)^{2q}])^{\frac{1}{2q}} - (\mathbb{E}[\operatorname{tr}\Gamma(0)^{2q}])^{\frac{1}{2q}} \leq \frac{1}{4}pm\varepsilon$$

If we want this condition to be implied by a requirement of the type

$$pm > \frac{C(\log(d))^{(\text{some power})}}{\varepsilon}$$

then a natural attempt would be to try to bound $(\mathbb{E}[\operatorname{tr}\Gamma(1)^{2q}])^{\frac{1}{2q}} - (\mathbb{E}[\operatorname{tr}\Gamma(0)^{2q}])^{\frac{1}{2q}}$ by just a constant times some power of $\log(d)$.

To bound $(\mathbb{E}[\operatorname{tr}\Gamma(1)^{2q}])^{\frac{1}{2q}} - (\mathbb{E}[\operatorname{tr}\Gamma(0)^{2q}])^{\frac{1}{2q}}$, we combine our differential inequality with Lemma 6.6 in [13].

By Lemma 6.6 in [13], we can get the bound of the type

$$(\mathbb{E}[\operatorname{tr}\Gamma(1)^{2q}])^{\alpha} - (\mathbb{E}[\operatorname{tr}\Gamma(0)^{2q}])^{\alpha} \leq C\alpha + K^{1-\alpha}$$

if we have a differential inequality of the form

$$|\frac{d}{dt}(\mathbb{E}[\operatorname{tr}\Gamma(t)^{2q}])| \leq C\max\{(\mathbb{E}[\operatorname{tr}\Gamma(t)^{2q}])^{1-\alpha}, K^{1-\alpha}\}$$

So we mainly want to obtain a differential inequality where C and K only contain factors of $\log(d)$ but do not contain any positive powers of pm or any negative power of $\varepsilon \approx \sqrt{d/m}$. To this end, we need to choose a variant of the bound for the left hand side of (7.3) which does not contain any positive powers of pm or any negative power of $\varepsilon \approx \sqrt{d/m}$.

Therefore, when choosing the different variants of the bound in Lemma 7.6, we seek to remove those factors that we do not want. One natural attempt would be to use the method similar to the proof of the second inequality in Theorem 2.9 in [13], where they first bound the small factors that arise after using Holder by some matrix parameters, and then replace those matrix parameters by the trace moments by Jensen inequality. In our case, this means to replace $\sqrt{pmpd}$ by $(\mathbb{E}[\operatorname{tr}\Gamma(t)^{2q}])^{\frac{1}{2q}}$, thereby removing the factors pm and pd.

However, in our case, after bounding the small factors obtained by directly separating the left hand side of (7.3) using Holder, we can only obtain factors of the form $\sqrt{\max\{pd,q\}}$ and $\sqrt{\max\{pm,q\}}$. (Recall that $q \sim \log(d/\varepsilon\delta)$). The product $\sqrt{\max\{pd,q\}\cdot\max\{pm,q\}}$ of these two factors can be interchanged by the factor $\sqrt{pmpd}$ only when $pd \geq \log(d/\varepsilon\delta)$. But this already means that $pm \geq O(\frac{\log(d/\varepsilon\delta)}{\varepsilon^2})$, since $\varepsilon = O(\sqrt{\frac{d}{m}})$.

To get rid of the unbalanced factor $\sqrt{\max\{pd,q\}\cdot\max\{pm,q\}}$, we develop a completely different method. We combine several factors at early stage and then use a key observation, Lemma **??**, to replace this combined factor by $(\mathbb{E}[\operatorname{tr}\Gamma(t)^{2q}])^{\frac{1}{2q}}$ directly. Eventually this method allowed us to get a more precise upper estimation, namely the right hand side of (7.3). In this upper estimation, although there are still some factors of $\max\{pd,q\}$, they are not harmful, because $\max\{pd,q\}$ is of smaller order than $\sqrt{pdpm}$ (see the proof of Proposition 7.3 for details).

Proof. TOPROVE 13 □

8. Leverage Score Sparsified Embeddings

In this section, we prove our subspace embedding guarantee for the LESS-IC distribution, Theorem 3.10. The proof is similar to the OSNAP case, and is accomplished via the following results,

- Theorem 3.10 establishes the subspace embedding guarantee from a bound on the trace moments of the embedding error analogous to Theorem 3.2 in the OSNAP case.
- Lemma 8.3 shows that it is sufficient to bound the moments of $(S_1U)^TS_2U+(S_2U)^TS_1U$ to control the trace moments of the embedding error, analogous to Lemma 7.2. This is the decoupling step.
- Lemma 8.4 establishes that the entries of the LESS-IC distribution are uncorrelated and have variance p, which means the Gaussian model that we interpolate with should have independent entries with variance p, just as in the case of OSNAP.
- Lemma 8.5 bounds the trace moments of the embedding error via interpolation after decoupling, analogous to Lemma 7.3.
- Lemma 8.6 establishes the differential inequality for the derivative of the interpolant, analogous to Lemma 7.4 in the OSNAP and OSE-IE case.
- Lemma 8.7 establishes the trace inequality required to obtain the differential inequality in Lemma 8.6, analogous to Lemma 7.6 in the case of OSNAP.
- The trace inequality in Lemma 8.7 in turn requires a bound for the row norm moments of $S(t)U$. This is provided by Lemma 8.8, analogous to Lemma 6.2.

Before we proceed, we state the formal definition of the LESS-IC distribution. The construction is similar to OSNAP, with some changes to reflect the different number and size of subcolumns and the different scaling for non-zero entries across columns.

Definition 8.1 (LESS-IC)**.** Given (β_1,β_2) leverage scores $z_1,...,z_n$, and $0<p<1$, define

$$b_j := \max\left\{\left\lfloor\frac{1}{\beta_1 p z_j}\right\rfloor,1\right\}\quad\text{and}\quad s_j := \left\lceil\frac{m}{b_j}\right\rceil.$$

An $m\times n$ random matrix S is called a K-wise independent unscaled leverage score sparsified embedding with independent columns (K-wise independent unscaled LESS-IC), and also $\Pi=(1/\sqrt{pm})S$ is called a K-wise independent LESS-IC, corresponding to (β_1,β_2)-approximate leverage scores $(z_1,...,z_n)$ with parameter p if it is distributed as

$$S=\sum_{l=1}^{n}\sum_{\gamma_l=1}^{s_l}\alpha_{(l,\gamma_l)}\xi_{(l,\gamma_l)}e_{\mu_{(l,\gamma_l)}}e_l^\top$$

where in this expression

- the collections $\{\xi_{(l,\gamma_l)}\}_{l\in[n],\gamma_l\in[s_l]}$ and $\{\mu_{(l,\gamma_l)}\}_{l\in[n],\gamma_l\in[s_l]}$ are mutually independent;
- $\{\xi_{l,\gamma_l}\}_{l\in[n],\gamma_l\in[s_l]}$ is a collection of K-wise independent Rademacher random variables;
- $\{\mu_{(l,\gamma_l)}\}_{l\in[n],\gamma_l\in[s_l]}$ is a collection of K-wise independent random variables such that each $\mu_{(l,\gamma_l)}$ is uniformly distributed in $[b_l(\gamma_l-1)+1:\min\{b_l\gamma_l,m\}]$;
- $\alpha_{(l,\gamma_l)} := \sqrt{p(\min\{b_l\gamma_l,m\}-b_l(\gamma_l-1))}$;
- $e_{\mu_{(l,\gamma)}}$ and e_l represent basis vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively.

In addition, if all the random variables in the collections $\{\xi_{(l,\gamma_l)}\}_{l\in[n],\gamma\in[s]}$ and $\{\mu_{(l,\gamma_l)}\}_{l\in[n],\gamma\in[s]}$ are fully independent, then S is called a fully independent unscaled LESS-IC and Π is called a fully independent LESS-IC.

Theorem 3.10 (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 $c_{3.10.1}$ and $c_{3.10.2}$ 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 $m \ge c_{3.10.1}\left(\frac{d+\log^2(d/\delta)+\log(1/\varepsilon)}{\varepsilon^2} + \log^3(d/\delta)/\varepsilon\right)$ and

$$c_{3.10.2}\max\left\{\frac{(\log(d/\varepsilon\delta))^{2.5}}{\varepsilon}, (\log(d/\varepsilon\delta))^3\right\} \le pm \le m.$$

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 8.2. We can obtain a subspace embedding guarantee for the LESS-IE distribution by following the proof of Theorem 9.2 suitably modified for the case of LESS. One can check that Theorem 9.3 holds even when S has the LESS-IE distribution with the same values of σ and R. Thus, a subspace embedding for the LESS-IE distribution holds under the same conditions as Theorem 3.10 with the additional requirement $pm \ge \frac{c\log(\frac{d}{\varepsilon\delta})}{\varepsilon^2}$.

Proof. TOPROVE 14 □

Lemma 8.3 (Decoupling)**.** *When S has the fully independent unscaled LESS-IC distribution,*

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

Consequently,

$$\mathbb{E}[\mathrm{tr}(U^TS^TSU - pm\cdot I_d)^{2q}] \le \mathbb{E}_{S,S'}\left[\mathrm{tr}\left(2\left((SU)^TS'U + (SU)^TS'U\right)\right)^{2q}\right]$$

where S' is an independent copy of S.

Proof. TOPROVE 15 □

Lemma 8.4 (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 fully independent unscaled LESS-IC 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. TOPROVE 16 □

Using the decoupling result, we bound the trace moments of the embedding error by interpolating between LESS and its Gaussian model exactly as in the proof of Lemma 7.3 in Section 7.3.

Lemma 8.5 (Trace Moments of Embedding Error for LESS)**.** *Let S be an $m\times n$ matrix distributed according to the fully independent unscaled LESS-IC distribution with parameter p for some fixed matrix U satisfying $U^TU = I$ with given (β_1,β_2)-approximate leverage scores. Define $X = \frac{1}{\sqrt{pm}}SU$.*

Given $0 < \varepsilon < 1$, there exist constants $c_{8.5.1}, c_{8.5.2}, c_{8.5.3}$ such that for $m \geq c_{8.5.1}\frac{d+q}{\varepsilon^2}$ and $q \in \mathbb{N}$ satisfying $2 \leq q \leq m$ and $pm \geq \left(\max\left\{\frac{c_{8.5.2}q^{5/2}}{\varepsilon}, c_{8.5.3}q^3\right\}\right)^{1+\frac{2}{q-2}}$,

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

Proof. TOPROVE 17 □

Here we obtain the differential inequality that arises during interpolation in the proof of Lemma 8.5.

Lemma 8.6 (Differential Inequality for LESS)**.** *Let $p > 0$. Let S_1 and S_2 be independent random matrices with the fully independent unscaled LESS-IC distribution with parameter p for some fixed matrix U satisfying $U^TU = I$ with given (β_1, β_2)-approximate leverage scores. Let G_1 and G_2 be independent random matrices with i.i.d. Guassian entries each with variance p, and define the interpolated random matrices,*

$$\begin{aligned} S_1(t) &= \sqrt{t}S_1 + \sqrt{1-t}G_1 \\ S_2(t) &= \sqrt{t}S_2 + \sqrt{1-t}G_2 \end{aligned} \tag{8.1}$$

Let $f(M_1, M_2) = \mathrm{tr}(((M_1U)^T(M_2U) + (M_2U)^T(M_1U))^{2q})$. Then there exists $c_{8.6} > 0$ such that, for any $q \geq 2$,

$$\frac{d}{dt}\mathbb{E}[f(S_1(t), S_2(t))] \leq \max_{4 \leq k \leq 2q}(c_{8.6}q)^k\left((pm)^{\frac{1}{q}}\sqrt{\max\{pd, q^2\}}\right)^{\frac{2qk-4q}{2q-2}}(\mathbb{E}[f(S_1(t), S_2(t))])^{1-\frac{k-2}{2q-2}}$$

Proof. TOPROVE 18 □

As in the oblivious case, a trace inequality is the key step in the proof of Lemma 8.6.

Lemma 8.7 (Trace Inequalities for LESS)**.** *Let $S_1(t)$ and $S_2(t)$ be as in Lemma 8.6. Let $\Gamma(t) = (S_1(t)U)^T(S_2(t)U) + (S_2(t)U)^T(S_1(t)U)$. Let $\mathcal{Z} = \{\xi_{(l,\gamma)} : l \in [n], \gamma_l \in [s_l]\} \cup \{\mu_{(l,\gamma)} : l \in [n], \gamma_l \in [s_l]\}$ be a family of mutually independent random variables be the family of mutually independent random variables generating an instance of S_1 with the unscaled LESS-IC distribution corresponding to some (β_1, β_2)-approximate leverage scores for U. Let $q \geq 2$ and $3 \leq k \leq 2q$. Let $\{\mathcal{Z}_\lambda\}_{\lambda\in[k]}$ be a family of (possibly dependent) random elements, where for each $\lambda \in [k]$, the random element*

$$\mathcal{Z}_\lambda = \{\xi_{(l,\gamma),\lambda} : l \in [n], \gamma_l \in [s_l]\} \cup \{\mu_{(l,\gamma),\lambda} : l \in [n], \gamma_l \in [s_l]\}$$

has the same distribution as

$$\mathcal{Z} = \{\xi_{(l,\gamma)} : l \in [n], \gamma_l \in [s_l]\} \cup \{\mu_{(l,\gamma)} : l \in [n], \gamma_l \in [s_l]\}$$

Let $Z_{(l,\gamma)} = \xi_{(l,\gamma)}e_{\mu_{(l,\gamma)}}e_l^T$ and $Z_{(l,\gamma),\lambda} = \xi_{(l,\gamma),\lambda}e_{\mu_{(l,\gamma),\lambda}}e_l^T$. Let $\{\Upsilon_1, ..., \Upsilon_k\}$ be a family of $L_\infty(S_\infty^d)$ random matrices. Assume further that the collection $\{\mathcal{Z}_\lambda\}_{\lambda\in[k]}$ is independent of S_1, S_2, G_1, G_2, and $\{\Upsilon_1, ..., \Upsilon_k\}$. (In other words, $\{\Upsilon_1, ..., \Upsilon_k\}$ can possibly be dependent with S_1, S_2, G_1, G_2.) For each $l \in [n], \gamma_l \in [s_l]$ and $\lambda \in k$, we define random vectors $\Theta_{(l,\gamma),\lambda,1}, \Theta_{(l,\gamma),\lambda,2} \in \mathbb{R}^d$ such that

$$\Theta_{(l,\gamma),\lambda,1} = \xi_{(l,\gamma),\lambda}\alpha_{(l,\gamma),\lambda}u_l^T \text{ and } \Theta_{(l,\gamma),\lambda,2} = e^T_{\mu_{(l,\gamma),\lambda}}S_2(t)U$$

where $e_{\mu_{(l,\gamma),\lambda}}$ represents the $\mu_{(l,\gamma),\lambda}$th coordinate vector. Then, given $0 \le \rho_1, ..., \rho_k \le +\infty$ such that $\sum_{\lambda=1}^{k} \frac{1}{\rho_\lambda} = 1 - \frac{k}{2q}$, $\tau_1, \ldots, \tau_k \in \mathrm{sym}(\{1,2\})$, there exists $c_{8.7} > 0$ such that

$$\begin{aligned}&\sum_{l\in[n],\gamma_l\in[s_l]} \mathbb{E}[\mathrm{tr}\, \Theta^T_{(l,\gamma),1,\tau_1(1)}\Theta_{(l,\gamma),1,\tau_1(2)}\Upsilon_1\Theta^T_{(l,\gamma),2,\tau_2(1)}\Theta_{(l,\gamma),2,\tau_2(2)}\\ &\cdots\Upsilon_2\Theta^T_{(l,\gamma),k,\tau_k(1)}\Theta_{(l,\gamma),k,\tau_k(2)}\Upsilon_k]\\ \le&(c_{8.7}(pm)^{\frac{1}{q}}\sqrt{\max\{\beta pd, q^2\}})^{\frac{2qk-4q}{2q-2}}(\mathbb{E}\,\mathrm{tr}((\Gamma(t))^{2q}))^{\frac{1}{q}\cdot\frac{2q-k}{2q-2}}\prod_{\lambda=1}^{k}\|\Upsilon_\lambda\|_{\rho_\lambda}\end{aligned}$$

Proof. TOPROVE 19 □

Lemma 8.8 (Row Norm for LESS-IC)**.** *Let $S(t) := \sqrt{t}S + \sqrt{1-t}G$, where S has the fully independent unscaled LESS-IC distribution as in Lemma 8.6 and G is an $m \times n$ matrix with i.i.d. Gausian entries with variance p. Let μ be a random variable uniformly distributed in $\phi \ne I \subset [m]$ and independent of S and G. Then, there exists $c_{8.8} > 0$, such that*

$$\mathbb{E}_{\mu,S(t)}[\|e_\mu^T S(t)U\|^q]^{\frac{1}{q}} \le c_{8.8}\sqrt{\max\{pd, q^2\}}$$

Proof. TOPROVE 20 □

9. Oblivious Subspace Embedding with Independent Entries

In this section, we consider the subspace embedding property for the following classical model with independent entries in the matrix S.

Definition 9.1 (OSE-IE)**.** An $m \times n$ random matrix S is called an unscaled oblivious subspace embedding with independent entries (unscaled OSE-IE) with parameter p if S has i.i.d. entries $s_{i,j} = \delta_{(i,j)}\xi_{(i,j)}$ where $\delta_{(i,j)}$ are i.i.d. Bernoulli random variables taking value 1 with probability $p \in (0,1]$ and $\xi_{(i,j)}$ are i.i.d. random variables independent with $\delta_{(i,j)}$ and satisfy $\mathbb{P}(\xi_{(i,j)} = 1) = \mathbb{P}(\xi_{(i,j)} = -1) = 1/2$. And in this case, $\Pi = (1/\sqrt{pm})S$ is call an OSE-IE with parameter p.

For this model, we have the following subspace embedding guarantee,

Theorem 9.2 (High Probability Bounds for the Embedding Error for OSE-IE)**.** *Let S be an $m\times n$ matrix distributed according to the unscaled OSE-IE distribution with parameter p. Let U be an arbitrary $n \times d$ deterministic matrix such that $U^TU = I$. Then, there exist constants $c_{9.2.1} > 0$ and $c_{9.2.2} > 0$ such that for any $0 < \varepsilon, \delta < 1$ and $d > 10$, we have*

$$\mathbb{P}\left(1-\varepsilon \le s_{\min}((1/\sqrt{pm})SU) \le s_{\max}((1/\sqrt{pm})SU) \le 1+\varepsilon\right) \ge 1-\delta \tag{9.1}$$

when $m > c_{9.2.1}\frac{d+\log(1/\varepsilon\delta)}{\varepsilon^2}$ and

$$pm \ge \min\left\{c_{9.2.2}\left(\frac{(\log(d/\varepsilon\delta))^2}{\varepsilon} + \frac{(\log(d/\varepsilon\delta))}{\varepsilon^2} + (\log(d/\varepsilon\delta))^3\right), m\right\}$$

The overall structure of the proof of Theorem 9.2 is the same as the proof in the OSNAP case and we only highlight the differences from the proof of the OSNAP case discussed in Section 7. A key difference between the above result and the corresponding result for OSNAP is the $1/\varepsilon^2$ dependence in the lower bound for sparsity. This arises due to our approach of decomposing $U^TS^TSU - pm\cdot I_d$ as

$$U^TS^TSU - pm\cdot I_d = \sum_{j=1}^{n}\left(\sum_{i=1}^{m} s_{ij}^2\right)u_ju_j^T - pm\cdot I_d + \sum_{i=1}^{m}\sum_{\substack{j,j'=1\\ j\ne j'}}^{n} s_{ij}s_{ij'}u_ju_{j'}^T$$

where we label the former term as the diagonal term and the latter as the off diagonal term. In the OSNAP model, we have $\sum_{i=1}^m s_{ij}^2 = pm$ by construction and therefore the diagonal term vanishes, but when S has the OSE-IE distribution, we need not necessarily have $\sum_{i=1}^m s_{ij}^2 = pm$, which means we need to analyze diagonal term in addition to the off-diagonal term analyzed in Lemma 7.2. Observing that,

$$\sum_{j=1}^n \left(\sum_{i=1}^m s_{ij}^2\right) u_j u_j^T - pm \cdot I_d = \sum_{j=1}^n \left(\sum_{i=1}^m (s_{ij}^2 - p)\right) u_j u_j^T$$

and using Minkowski's inequality,

$$\mathbb{E}[\operatorname{tr}(U^T S^T S U - pm \cdot I_d)^{2q}]^{\frac{1}{2q}} \le \mathbb{E}\left[\operatorname{tr}\left(\sum_{i=1}^m \sum_{j=1}^n (s_{ij}^2 - p) u_j u_j^T\right)^{2q}\right]^{\frac{1}{2q}} + \mathbb{E}\left[\operatorname{tr}\left(\sum_{i=1}^m \sum_{j,j'=1, j\neq j'}^n s_{ij} s_{ij'} u_j u_{j'}^T\right)^{2q}\right]^{\frac{1}{2q}}$$

Proposition 9.3 controls the diagonal term (the former term), which gives rise to the $1/\varepsilon^2$ dependence, and is analyzed in Section 9.1. The analysis of the off-diagonal term (the latter term) is similar to the OSNAP case and is discussed in Section 9.2.

9.1. Controlling the diagonal term when S is the OSE-IE distribution.

Proposition 9.3 (Diagonal Term)**.** *Let S be an $m \times n$ matrix distributed according to the unscaled OSE-IE distribution with parameter p. Let U be an arbitrary $n \times d$ deterministic matrix such that $U^T U = I$. Given $0 < \varepsilon < 1$ and $q \ge \log(d)$, there exist constants $c_{9.3}$ such that for $m \ge d \ge 20$ and $pm \ge c_{9.3} \frac{q}{\varepsilon^2}$,*

$$\mathbb{E}\left[\operatorname{tr}\left(\frac{1}{pm}\sum_{i=1}^m \sum_{j=1}^n (s_{ij}^2 - p) u_j u_j^T\right)^{2q}\right]^{\frac{1}{2q}} \le \varepsilon$$

Proposition 9.3 shows that, to make the diagonal term small, it suffices to require the nonzero entries per column to be greater than or equal to $c\frac{\log(d)}{\varepsilon^2}$. Note that

$$\mathbb{E}\left[\operatorname{tr}\left(\frac{1}{pm}\sum_{i=1}^m \sum_{j=1}^n (s_{ij}^2 - p) u_j u_j^T\right)^{2q}\right]^{\frac{1}{2q}} \ge \mathbb{E}\left[\operatorname{tr}\left(\frac{1}{pm}\sum_{i=1}^m \sum_{j=1}^n (s_{ij}^2 - p) u_j u_j^T\right)^{2}\right]^{\frac{1}{2}}$$

and the latter can be of the order $1/\sqrt{pm}$ when U corresponds to a d-dimensional coordinate subspace of $\mathbb{R}^n$. Thus, the $1/\varepsilon^2$ dependence in the lower bound on pm is necessary when using this approach to prove the subspace embedding guarantee.

Proof. TOPROVE 21 □

9.2. Proving the subspace embedding guarantee for the OSE-IE distribution.

In this section, we prove 9.2.

Theorem 9.2 (High Probability Bounds for the Embedding Error for OSE-IE)**.** *Let S be an $m \times n$ matrix distributed according to the unscaled OSE-IE distribution with parameter p. Let U be an* *arbitrary $n \times d$ deterministic matrix such that $U^TU = I$. Then, there exist constants $c_{9.2.1} > 0$ and $c_{9.2.2} > 0$ such that for any $0 < \varepsilon, \delta < 1$ and $d > 10$, we have*

$$\mathbb{P}\left(1 - \varepsilon \le s_{\min}((1/\sqrt{pm})SU) \le s_{\max}((1/\sqrt{pm})SU) \le 1 + \varepsilon\right) \ge 1 - \delta \tag{9.1}$$

when $m > c_{9.2.1}\frac{d + \log(1/\varepsilon\delta)}{\varepsilon^2}$ and

$$pm \ge \min\left\{c_{9.2.2}\left(\frac{(\log(d/\varepsilon\delta))^2}{\varepsilon} + \frac{(\log(d/\varepsilon\delta))}{\varepsilon^2} + (\log(d/\varepsilon\delta))^3\right), m\right\}$$

Proof. TOPROVE 22 □

9.3. **Trace Inequality in the OSE-IE Case.** The trace inequality required to obtain the differential inequality for the interpolant between the moments of $(S_1U)^TS_2U$ and $(G_1U)^TG_2U$ has a slightly different proof in the OSE-IE case than in the OSNAP case, even though both bounds are the same.

Lemma 9.4. *Let $S_1(t)$ and $S_2(t)$ be as in Lemma 7.4 with both having the OSE-IE distribution. Let $\Gamma(t) = (S_1(t)U)^T(S_2(t)U) + (S_2(t)U)^T(S_1(t)U)$. Let $\mathcal{Z} = \{\xi_{(l,\gamma)} : (l,\gamma) \in [n] \times [m]\} \cup \{\delta_{(l,\gamma)} : (l,\gamma) \in [n] \times [m]\}$ be the family of mutually independent random variables generating an instance of S_1 with the OSE-IE distribution. Let $q \ge 2$ and $3 \le k \le 2q$. Let $\{\mathcal{Z}_\lambda\}_{\lambda\in[k]}$ be a family of (possibly dependent) random elements, where for each $\lambda \in [k]$, the random element*

$$\mathcal{Z}_\lambda = \{\xi_{(l,\gamma),\lambda} : (l,\gamma) \in [n] \times [m]\} \cup \{\delta_{(l,\gamma),\lambda} : (l,\gamma) \in [n] \times [m]\}$$

has the same distribution as

$$\mathcal{Z} = \{\xi_{(l,\gamma)} : (l,\gamma) \in [n] \times [m]\} \cup \{\delta_{(l,\gamma)} : (l,\gamma) \in [n] \times [m]\}$$

Let $Z_{(l,\gamma)} = \xi_{(l,\gamma)}\delta_{(l,\gamma)}e_\gamma e_l^T$ and $Z_{(l,\gamma),\lambda} = \xi_{(l,\gamma),\lambda}\delta_{(l,\gamma),\lambda}e_\gamma e_l^T$. Let $\{\Upsilon_1, ..., \Upsilon_k\}$ be a family of $L_\infty(S_\infty^d)$ random matrices. Assume further that the collection $\{\mathcal{Z}_\lambda\}_{\lambda\in[k]}$ is independent of S_1, S_2, G_1, G_2, and $\{\Upsilon_1, ..., \Upsilon_k\}$. (In other words, $\{\Upsilon_1, ..., \Upsilon_k\}$ can possibly be dependent with S_1, S_2, G_1, G_2.) For each $(l,\gamma) \in [n] \times [m]$ and $\lambda \in k$, we define random vectors $\Theta_{(l,\gamma),\lambda,1}, \Theta_{(l,\gamma),\lambda,2} \in \mathbb{R}^d$ such that

$$\Theta_{(l,\gamma),\lambda,1} = \xi_{(l,\gamma),\lambda}\delta_{(l,\gamma),\lambda}\mathbf{u}_l^T \text{ and } \Theta_{(l,\gamma),\lambda,2} = e_\gamma^T S_2(t)U$$

where e_γ represents the γ-th coordinate vector. Then, given $0 \le \beta_1, ..., \beta_k \le +\infty$ such that $\sum_{\lambda=1}^k \frac{1}{\beta_\lambda} = 1 - \frac{k}{2q}$, $\tau_1, \ldots, \tau_k \in \mathrm{sym}(\{1,2\})$, there exists $c_{9.4} > 0$ such that

$$\begin{aligned}\sum_{(l,\gamma)\in[n]\times[m]} & \mathbb{E}[\mathrm{tr}\, \Theta^T_{(l,\gamma),1,\tau_1(1)}\Theta_{(l,\gamma),1,\tau_1(2)}\Upsilon_1\Theta^T_{(l,\gamma),2,\tau_2(1)}\Theta_{(l,\gamma),2,\tau_2(2)} \\ & \cdots \Upsilon_2\Theta^T_{(l,\gamma),k,\tau_k(1)}\Theta_{(l,\gamma),k,\tau_k(2)}\Upsilon_k] \\ \le & (c_{9.4}(pm)^{\frac{1}{q}}\sqrt{\max\{pd,q\}})^{\frac{2qk-4q}{2q-2}}(\mathbb{E}\,\mathrm{tr}((\Gamma(t))^{2q}))^{\frac{1}{q}\cdot\frac{2q-k}{2q-2}}\prod_{\lambda=1}^k \|\Upsilon_\lambda\|_{\beta_\lambda}\end{aligned}$$

Proof. TOPROVE 23 □

10. 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. Sarlos, T. *Improved approximation algorithms for large matrices via random projections* in *2006 47th annual IEEE symposium on foundations of computer science (FOCS'06)* (2006), 143–152.
2. Clarkson, K. L. & Woodruff, D. P. *Low rank approximation and regression in input sparsity time* in *Proceedings of the forty-fifth annual ACM symposium on Theory of Computing* (2013), 81–90.
3. Cohen, M. B., Elder, S., Musco, C., Musco, C. & Persu, M. *Dimensionality reduction for k-means clustering and low rank approximation* in *Proceedings of the forty-seventh annual ACM symposium on Theory of computing* (2015), 163–172.
4. Woodruff, D. P. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science* **10,** 1–157 (2014).
5. Meng, X. & Mahoney, M. W. *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* (2013), 91–100.
6. Nelson, J. & Nguyên, H. L. *OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings* in *2013 ieee 54th annual symposium on foundations of computer science* (2013), 117–126.
7. Bourgain, J., Dirksen, S. & Nelson, J. *Toward a unified theory of sparse dimensionality reduction in euclidean space* in *Proceedings of the forty-seventh annual ACM symposium on Theory of Computing* (2015), 499–508.
8. Cohen, M. B. *Nearly tight oblivious subspace embeddings by trace inequalities* in *Proc. of the 27th annual ACM-SIAM Symposium on Discrete Algorithms* (2016), 278–287.
9. Chenakkod, S., Dereziński, M., Dong, X. & Rudelson, M. *Optimal embedding dimension for sparse subspace embeddings* in *Proceedings of the 56th Annual ACM Symposium on Theory of Computing* (2024), 1106–1117.
10. Nelson, J. & Nguyên, H. L. *Lower bounds for oblivious subspace embeddings* in *International Colloquium on Automata, Languages, and Programming* (2014), 883–894.
11. Li, Y. & Liu, M. *Lower bounds for sparse oblivious subspace embeddings* in *Proceedings of the 41st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* (2022), 251–260.
12. Tropp, J. A. An Introduction to Matrix Concentration Inequalities. *Found. Trends Mach. Learn.* **8,** 1–230. ISSN: 1935-8237 (May 2015).
13. Brailovskaya, T. & van Handel, R. Universality and sharp matrix concentration inequalities. *Geometric and Functional Analysis,* 1–105 (2024).
14. Drineas, P., Mahoney, M. W. & Muthukrishnan, S. *Sampling algorithms for ℓ_2 regression and applications* in *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm* (2006), 1127–1136.
15. Drineas, P., Magdon-Ismail, M., Mahoney, M. W. & Woodruff, D. P. Fast approximation of matrix coherence and statistical leverage. *The Journal of Machine Learning Research* **13,** 3475–3506 (2012).
16. Dereziński, M., Liao, Z., Dobriban, E. & Mahoney, M. *Sparse sketches with small inversion bias* in *Conference on Learning Theory* (2021), 1467–1510.
17. Dereziński, M., Lacotte, J., Pilanci, M. & Mahoney, M. W. Newton-LESS: Sparsification without Trade-offs for the Sketched Newton Update. *Advances in Neural Information Processing Systems* **34,** 2835–2847 (2021).

18. Dereziński, M. *Algorithmic gaussianization through sketching: Converting data into sub-gaussian random designs* in *The Thirty Sixth Annual Conference on Learning Theory* (2023), 3137–3172.
19. Chepurko, N., Clarkson, K. L., Kacham, P. & Woodruff, D. P. *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)* (2022), 3043–3068.
20. Cherapanamjeri, Y., Silwal, S., Woodruff, D. P. & Zhou, S. *Optimal algorithms for linear algebra in the current matrix multiplication time* in *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)* (2023), 4026–4049.
21. Drineas, P. & Mahoney, M. W. RandNLA: randomized numerical linear algebra. *Communications of the ACM* **59,** 80–90 (2016).
22. Martinsson, P.-G. & Tropp, J. A. Randomized numerical linear algebra: Foundations and algorithms. *Acta Numerica* **29,** 403–572 (2020).
23. Dereziński, M. & Mahoney, M. W. *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* (2024), 6470–6479.
24. Ailon, N. & Chazelle, B. The fast Johnson–Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing* **39,** 302–322 (2009).
25. Tropp, J. A. Improved analysis of the subsampled randomized Hadamard transform. *Advances in Adaptive Data Analysis* **3,** 115–126 (2011).
26. Dasgupta, A., Kumar, R. & Sarlós, T. *A sparse johnson: Lindenstrauss transform* in *Proceedings of the forty-second ACM symposium on Theory of computing* (2010), 341–350.
27. Kane, D. M. & Nelson, J. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)* **61,** 1–23 (2014).
28. Cartis, C., Fiala, J. & Shao, Z. Hashing embeddings of optimal dimension, with applications to linear least squares. *arXiv preprint arXiv:2105.11815* (2021).
29. Tropp, J. A. Comparison theorems for the minimum eigenvalue of a random positive-semidefinite matrix. *arXiv preprint arXiv:2501.16578* (2025).
30. Achlioptas, D. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *Journal of computer and System Sciences* **66,** 671–687 (2003).
31. Cohen, M. B., Nelson, J. & Woodruff, D. P. *Optimal approximate matrix product in terms of stable rank* in *International Colloquium on Automata, Languages, and Programming* (2016).
32. Boucheron, S., Lugosi, G. & Massart, P. *Concentration inequalities: A nonasymptotic theory of independence* (Oxford university press, 2013).
33. Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science* ISBN: 9781108244541. https://books.google.com/books?id=TahxDwAAQBAJ (Cambridge University Press, 2018).
34. Bergh, J. & Löfström, J. *Interpolation spaces: an introduction* (Springer Science & Business Media, 2012).
35. Pisier, G. & Xu, Q. in *Handbook of the geometry of Banach spaces* 1459–1517 (Elsevier, 2003).
36. Rudelson, M. & Vershynin, R. *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* (2010), 1576–1602.
37. Bandeira, A. S. & van Handel, R. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *Annals of Probability* **44,** 2479–2506 (2016).
38. Vershynin, R. *High-dimensional probability: An introduction with applications in data science* (Cambridge university press, 2018).
39. Kallenberg, O. & Kallenberg, O. *Foundations of modern probability Third Edition* (Springer, 2021).

40. Folland, G. B. *Real analysis: modern techniques and their applications* (John Wiley & Sons, 2013).

41. Carlen, E. Trace inequalities and quantum entropy: an introductory course. *Entropy and the quantum* **529,** 73–140 (2010).