

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

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

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 $O(s \cdot \text{nnz}(x))$, where $\text{nnz}(x)$ 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 $n \geq d$ 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 $m = O(d \text{ polylog}(d)/\varepsilon^2)$ and sparsity $s = O(\text{polylog}(d)/\varepsilon)$. 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 $n \geq d$ and $\varepsilon, \delta \in (0, 1)$ such that $1/\varepsilon\delta \leq \text{poly}(d)$, 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 $\tilde{O}(\text{nnz}(A)/\varepsilon)$ and considering the column subspace of A . However, given direct access to A , one may hope to get true input sparsity time $O(\text{nnz}(A))$ 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 i th leverage score of A is defined as the squared norm of the i th row of the matrix obtained by orthonormalizing the columns of A [14]. We show that if the i th leverage score of A is bounded by $l_i \in [0, 1]$, then the i th 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 $O(\text{nnz}(A) + d^\omega)$ required for approximating the leverage scores (here, ω is the matrix multiplication exponent). However, our approach significantly improves on these prior works in the $\text{poly}(d/\varepsilon)$ 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 $1/\varepsilon \leq \text{poly}(d)$, in*

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

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

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

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 $\tilde{O}(d^{2+\gamma}/\varepsilon^6)$ to $\tilde{O}(d^{2+\gamma}/\varepsilon)$, while achieving a matching $O(d/\varepsilon^2) \times d$ reduction.

Corollary 1.5 (Fast reduction for constrained least squares). *Given $A \in \mathbb{R}^{n \times d}$, $b \in \mathbb{R}^n$, $\varepsilon > 0$, function $g : \mathbb{R}^d \rightarrow \mathbb{R}_{\geq 0}$ and set $C \subseteq \mathbb{R}^d$ consider an $n \times d$ problem $\text{LS}_{C,g}(A, b, \varepsilon)$:*

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

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

$$(3.1) \quad \Pr(1 - \varepsilon \leq s_{\min}(\Pi U) \leq s_{\max}(\Pi U) \leq 1 + \varepsilon) \geq 1 - \delta,$$

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 $s = \tilde{O}(1/\varepsilon)$.

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

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

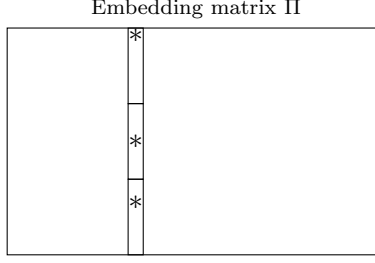


FIGURE 1. An example of a column divided into $s = 3$ subcolumns with each subcolumn having exactly one non-zero entry in a random position.

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*

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

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.

Remark 3.3. We note that if $1/\varepsilon$ is polynomial in d/δ , i.e., $\varepsilon \geq \frac{1}{(d/\delta)^K}$ for some absolute constant $K \geq 1$, then the $\log(1/\varepsilon)$ term in $\log(d/\varepsilon\delta) = \log(d/\delta) + \log(1/\varepsilon)$ is dominated by $\log(d/\delta)$. In this case, our requirement will become

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

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

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

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

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

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

$$(3.2) \quad \left| \|\Pi e_j\|_2^2 - 1 \right| = \left| \frac{N_j}{pm} - 1 \right| \leq \varepsilon, \quad N_j := \text{nnz}(S e_j) \sim \text{Binomial}(m, p).$$

Set

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

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

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

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

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

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

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 $pm = \Omega(\varepsilon^{-2})$.

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,

$$\mathbb{E}_{\Pi \sim \mathcal{D}} \|\Pi U\|^{\ell} < \varepsilon^{\ell} \delta.$$

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 $O(\text{nnz}(A)(\log^2(\frac{d}{\varepsilon\delta})/\varepsilon + \log^3(\frac{d}{\varepsilon\delta})))$. 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 $O(\text{nnz}(A) \log(d/\delta) + d^2 \log^4(d/\delta))$.

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 $O(\text{nnz}(A))$, plus an additional sublinear term. Given that the preprocessing step of approximating the leverage scores has to take at least $\text{nnz}(A)$ time, the linear term in the cost of applying LESS-IC is negligible.

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

Definition 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 $(l_1, \dots, l_n) \in [0, 1]^n$ of numbers are (β_1, β_2) -approximate

leverage scores for U if, for $1 \leq i \leq n$,

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

We say that the numbers $(l_1, \dots, l_n) \in [0, 1]^n$ 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, \dots, 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 $\beta_1 p m z_j$ many non-zero entries in column j in expectation. However, to achieve $1/\varepsilon$ dependency of the sparsity, we need to have *exactly* $\beta_1 p m z_j$ 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 $\beta_1 p m z_j$ many non-zero entries in column j , we would need $\beta_1 p m z_j$ 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 $m/(\beta_1 p m z_j) = 1/(\beta_1 p z_j)$. However, since $1/(\beta_1 p z_j)$ may not be an integer, we consider subcolumns of size $b_j := \max\{\lfloor 1/(\beta_1 p z_j) \rfloor, 1\}$.

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*

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

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

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

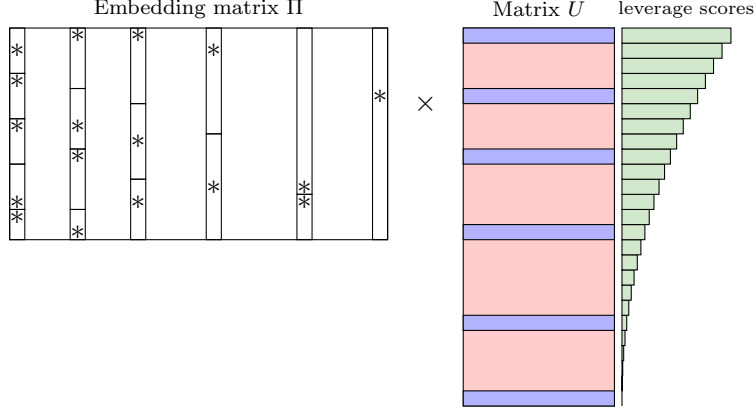


FIGURE 2. In the LESS-IC distribution, column j is filled with s_j many subcolumns, with the bottom-most subcolumn truncated to fit the size of Π . Each subcolumn has one non-zero entry. Notice that as the leverage scores decrease, the number of subcolumns decreases and the matrix becomes sparser. However, 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 $O(\text{nnz}(A) + \beta pmd^2)$ 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 $O(\text{nnz}(A)) + \tilde{O}(\beta d^2/\varepsilon)$. 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 $O(\gamma^{-1}(\text{nnz}(A) + d^2) + d^\omega)$. Applying LESS-IC (Theorem 3.10) with these leverage scores and parameters β_1, β_2 , computing ΠA takes $O(\text{nnz}(A) + n^\gamma d^2 \log^3(d/\varepsilon\delta)/\varepsilon)$, where $\text{nnz}(A)$ comes from the fact that every column of Π has at least one non-zero, while the second term accounts for the additional $O(\beta d \log^3(d/\varepsilon\delta)/\varepsilon)$ non-zeros.

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

Finally, we note that Corollary 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 $\{1, 2, \dots, n\}$ and the notation $\mathcal{P}([n])$ 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 l th coordinate vector is denoted by e_l . All matrices considered in this paper are real valued and the space of $m \times n$ matrices with real valued entries is denoted by $M_{m \times n}(\mathbb{R})$. Also, for a matrix $X \in M_{d \times d}(\mathbb{R})$, the notation $\text{Tr}(X)$ denotes the trace of the matrix X , and $\text{tr}(X) = \frac{1}{d} \text{Tr}(X)$ denotes the normalized trace. We write the operator norm of a matrix X as $\|X\|$, and it is also denoted by $\|X\|_{op}$ in some places where other norms appear for clarity. The spectrum of a matrix X is denoted by $\text{spec}(X)$. The standard probability measure is denoted by \mathbb{P} , and the symbol \mathbb{E} means taking the expectation with respect to this standard probability measure. To simplify the notation, we follow the convention from [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 $\text{Cov}(X, Y)$. The standard L_q norm of a random variable ξ is denoted by $\|\xi\|_q$, for $1 \leq q \leq \infty$. Throughout the paper, the symbols c_1, c_2, \dots , and $Const, Const', \dots$ 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

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

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):

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

In this case, standard calculations on Gaussian random matrices(see Lemma 6.8) show that $(\mathbb{E}[\text{tr}(G^T G - pm I_{d \times d})^{2q}])^{\frac{1}{2q}} \leq cpm\sqrt{\frac{d}{m}} = O(pm\varepsilon)$ when $m = \Omega(d/\varepsilon^2)$ and G has the covariance profile of SU . So it is enough to show that

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

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

Now, [13, Proposition 9.12] does take a similar approach of comparing $(SU)^T(SU) - pm I_d$ and $G^T G - pm I_d$, 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 $(SU)^T(SU) - pm I_d = \sum_{i=1}^m (U^T s_i s_i^T U - p I_d)$, 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 $\mathbb{E}[\text{tr}((SU)^T(SU) - pm \cdot I_d)^{2q}]^{\frac{1}{2q}}$ with $\mathbb{E}[\text{tr}(H)^{2q}]^{\frac{1}{2q}}$ where H is the Gaussian model for $(SU)^T(SU) - pm I_d$. 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 $(S_1 U)^T(S_2 U) + (S_2 U)^T(S_1 U)$ 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 $\mathbb{E}[\text{tr}((S_1 U)^T(S_2 U) + (S_2 U)^T(S_1 U))^{2q}]^{\frac{1}{2q}}$ into two parts, bounding $\mathbb{E}[\text{tr}(G_1^T G_2 + G_2^T G_1)^{2q}]^{\frac{1}{2q}}$ for the Gaussian model, and the difference

$$\mathbb{E}[\text{tr}((S_1 U)^T(S_2 U) + (S_2 U)^T(S_1 U))^{2q}]^{\frac{1}{2q}} - \mathbb{E}[\text{tr}(G_1^T G_2 + G_2^T G_1)^{2q}]^{\frac{1}{2q}},$$

which is called the universality error.

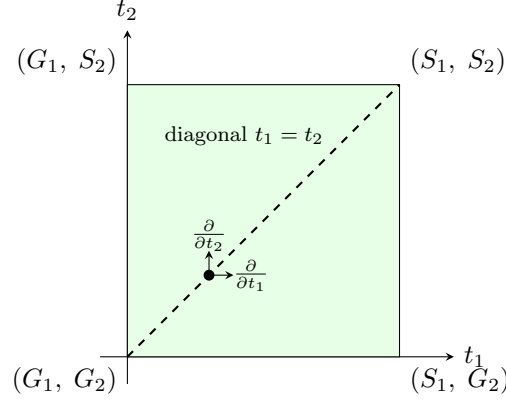


FIGURE 3. Two-dimensional interpolation in $(t_1, t_2) \in [0, 1]^2$, decomposed using the chain rule.

By standard calculations, we have $\mathbb{E}[\text{tr}(G_1^T G_2 + G_2^T G_1)^{2q}]^{\frac{1}{2q}} \leq c\sqrt{pm}\sqrt{pd} = O(pm\varepsilon)$ and the main task is still to bound the universality error. The advantage of the decoupling idea is that, informally speaking, since S_1 and S_2 are independent, we can condition on one of them, e.g., S_1 . For fixed S_1 , the random matrix $(S_1 U)^T (S_2 U)$ (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:

$$(4.1) \quad \left| \mathbb{E}[\text{tr}((S_1 U)^T (S_2 U) + (S_2 U)^T (S_1 U))^{2q}]^{\frac{1}{2q}} - \mathbb{E}[\text{tr}(G_1^T G_2 + G_2^T G_1)^{2q}]^{\frac{1}{2q}} \right| \leq \text{polylog}\left(\frac{d}{\varepsilon\delta}\right).$$

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

Nevertheless, the conditioning argument cannot be done directly because

$$\mathbb{E}\left[\text{tr}((S_1 U)^T (S_2 U) + (S_2 U)^T (S_1 U))^{2q}\right]^{\frac{1}{2q}} \neq \mathbb{E}_{S_1}\left[\mathbb{E}_{S_2}\left[\text{tr}((S_1 U)^T (S_2 U) + (S_2 U)^T (S_1 U))^{2q}\right]^{\frac{1}{2q}}\right].$$

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

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

By independence of S_1 and S_2 , we can condition on S_2 when we bound the partial derivative $\frac{\partial}{\partial t_1} \mathbb{E}[f(S_1(t_1), S_2(t_2))] \Big|_{t_1=t, t_2=t}$, 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 7.5) 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 sketch of Theorem 3.2. Let $X := \frac{1}{\sqrt{pm}}SU$. We first assume that the collection of all the random variables $\{\xi_{(l,\gamma)}, \mu_{(l,\gamma)}\}_{l \in [n], \gamma \in [s]}$ in the unscaled OSNAP construction are fully independent, and later we will check what is the minimum independence needed.

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

$$\mathbb{P}(\|X^T X - I_d\| \leq \varepsilon) \geq 1 - \delta.$$

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

$$\mathbb{P}\left(\|X^T X - I_d\| \geq \delta^{-\frac{1}{2q}} \mathbb{E}[d \operatorname{tr}(X^T X - I_d)^{2q}]^{\frac{1}{2q}}\right) \leq \delta$$

which, after simplification, becomes

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

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

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

Therefore, we have

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

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

Lemma 7.3 (Trace Moments of Embedding Error for OSNAP). *For X as above, there exist constants $c_{7.3.1}, c_{7.3.2}, c_{7.3.3} > 0$ such that for $q \in \mathbb{N}$ satisfying $2 \leq q \leq m$, we have*

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

$$(5.1) \quad \text{when} \quad m \geq c_{7.3.1} \frac{d+q}{\varepsilon^2}$$

$$(5.2) \quad \text{and} \quad pm \geq \left(\max \left\{ \frac{c_{7.3.2} q^2}{\varepsilon}, c_{7.3.3} q^3 \right\} \right)^{1 + \frac{2}{q-2}}.$$

Applying Lemma 7.3 (with appropriately adjusted ε) implies

$$\mathbb{P}(\|X^T X - I_d\| \geq \varepsilon) \leq \delta$$

when combined with the previous calculations.

It remains to check that conditions (5.1) and (5.2) are satisfied for $q = \lceil 2 \log(\frac{d}{\varepsilon\delta}) \rceil + 2$ by requiring $m \geq c_{3.2.1}(d + \log(1/\delta\varepsilon))/\varepsilon^2$ and $s = pm \geq c_{3.2.2}(\log^2(\frac{d}{\varepsilon\delta})/\varepsilon + \log^3(\frac{d}{\varepsilon\delta}))$, and this is done in the full version of the proof in Section 7.1.

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

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

Proof of Corollary 3.6. By the proof of Theorem 3.2, we see that when $m \geq c_{3.2.1}(d + \log(1/\delta\varepsilon))/\varepsilon^2$ and $s \geq \min\{c_{3.2.2}(\log^2(\frac{d}{\varepsilon\delta})/\varepsilon + \log^3(\frac{d}{\varepsilon\delta})), m\}$, then $\mathbb{E}[\text{tr}(X^T X - I_d)^{2q}] \leq \varepsilon^{2q}$, for $q = 8 \log(\frac{d}{\varepsilon\delta})$ (The proof originally has $q = \lceil 2 \log(\frac{d}{\varepsilon\delta}) \rceil + 2$, but upon going through the proof we see that $q = 8 \log(\frac{d}{\varepsilon\delta})$ also works). To get $\mathbb{E}[\text{tr}(X^T X - I_d)^{2q}] \leq \varepsilon^{2q} \delta/d$, it suffices for m and s to satisfy the same lower bounds, but with ε replaced by $\varepsilon(\delta/d)^{\frac{1}{2q}} \geq c\varepsilon$ for some $c > 0$ since $q \geq \log(d/\delta)$. These new lower bounds can be achieved by lower bounds of the same form as Theorem 3.2, but with different constants. The claim follows, since $\|X^T X - I_d\|^{2q} \leq d \text{tr}(X^T X - I_d)^{2q}$. \square

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

Proof sketch of Lemma 7.3. Our first step is to observe that due to the property of S having a fixed number of non-zero entries in a column, all quadratic terms in $(SU)^T(SU) - pm \cdot I_d$ are square-free, and this allows us to use the decoupling technique to reduce the problem of controlling the moments of $(SU)^T(SU) - pm \cdot I_d$ to controlling the moments of $(S_1 U)^T(S_2 U) + (S_2 U)^T(S_1 U)$ where S_1 and S_2 are independent copies. This is shown in the following claim, with the proof deferred to Section 7.2.

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

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

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

$$\mathbb{E} \left[\text{tr} (U^T S^T S U - pm \cdot I_d)^{2q} \right] \leq \mathbb{E}_{S_1, S_2} \left[\text{tr} (2((S_1 U)^T S_2 U + (S_2 U)^T S_1 U))^{2q} \right]$$

where S_2 is an independent copy of S_1 .

To estimate the moments of $(S_1 U)^T S_2 U$, we compare them to moments from the Gaussian case, i.e. the moments of $(G_1 U)^T G_2 U$ where the entries of G_1 and G_2 are independent normal random variables with variance p (since the entries of S_1 and S_2 are also uncorrelated with mean 0 and variance p , see Lemma 6.1). In this case, due to orthogonal invariance of the Gaussian distribution, the matrices $G_1 U$ and $G_2 U$ are distributed as $\sqrt{p} H_1$ and $\sqrt{p} H_2$ where H_1 and H_2 are $m \times d$ matrices with independent standard normal entries. Thus, we can rely on the following bound, which uses standard results about the norms of Gaussian random matrices with independent entries.

Lemma (Trace Moment of Embedding Error for Decoupled Gaussian Model, Lemma 6.9). *Let H_1 and H_2 be independent $m \times d$ random matrices with i.i.d. Gaussian entries. Then for any positive integer q , there exists $c_{6.9} > 0$ such that*

$$\mathbb{E} \left[\text{tr} (H_1^T H_2 + H_2^T H_1)^{2q} \right]^{\frac{1}{2q}} \leq c_{6.9} \sqrt{\max\{d, q\}} \sqrt{\max\{m, q\}}.$$

To formally compare the moments of $(S_1 U)^T S_2 U$ and $(G_1 U)^T G_2 U$, we define the interpolating matrices $S_1(t), S_2(t)$ for $t \in [0, 1]$ as described in Section 4:

$$(5.3) \quad \begin{aligned} S_1(t) &= \sqrt{t} S_1 + \sqrt{1-t} G_1, \\ S_2(t) &= \sqrt{t} S_2 + \sqrt{1-t} G_2. \end{aligned}$$

Let $\Gamma(M_1, M_2) = (M_1 U)^T (M_2 U) + (M_2 U)^T (M_1 U)$ and $\Gamma(t) = \Gamma(S_1(t), S_2(t))$. Then, due to the decoupling lemma (Lemma 7.2), to prove Lemma 7.3 it is enough to show that $\mathbb{E}[\text{tr}(\Gamma(1))^{2q}]^{\frac{1}{2q}} \leq pm\varepsilon/2$. Now, by Lemma 6.9, we know that:

$$\mathbb{E}[\text{tr}(\Gamma(0))^{2q}]^{\frac{1}{2q}} = \mathbb{E}[\text{tr}(\Gamma(G_1, G_2))^{2q}]^{\frac{1}{2q}} \leq c_{6.9} p \sqrt{\max\{d, q\}} \sqrt{\max\{m, q\}}.$$

Since we want to find the conditions for which $\mathbb{E}[\text{tr}(\Gamma(0))^{2q}]^{\frac{1}{2q}} \leq pm\varepsilon/4$, it is enough to ensure that $c_{6.9} p \sqrt{\max\{d, q\}} \sqrt{\max\{m, q\}} \leq pm\varepsilon/4$. Clearly, this can only happen when $q \leq m$, and in this case the inequality holds when $m \geq \frac{c_{6.9} p}{\varepsilon^2}$. Thus, it suffices to show

$$(5.4) \quad \mathbb{E}[\text{tr} \Gamma(1)^{2q}]^{\frac{1}{2q}} - \mathbb{E}[\text{tr} \Gamma(0)^{2q}]^{\frac{1}{2q}} \leq \frac{1}{4} pm\varepsilon.$$

For this, we look to estimate the derivative $\frac{d}{dt} \mathbb{E}[\text{tr} \Gamma(t)^{2q}]$, and we obtain the following estimate in Lemma 7.5 using the 2D interpolation idea mentioned in Section 4.

Lemma 7.5 (Differential Inequality). *For $\Gamma(t)$ as defined above, there exists a constant $c_{7.5}$ such that, for any $q \geq 2$, we have*

$$\frac{d}{dt} \mathbb{E}[\text{tr} \Gamma(t)^{2q}] \leq \max_{4 \leq k \leq 2q} (c_{7.5} q)^k \left((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}} \right)^{\frac{qk-2q}{q-1}} \mathbb{E}[\text{tr} \Gamma(t)^{2q}]^{1-\frac{k-2}{2q-2}}.$$

This differential inequality can be separated into two distinct cases: $pd \leq q$ and $pd > q$. When $pd \leq q$, we can simplify the expression on the right using convexity arguments, and use Lemma 6.6 from [13] to solve the differential inequality and obtain the following bound:

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

for some $c_7 > 0$ (this is done in the full proof of Lemma 7.3). Thus, inequality (5.4) is satisfied when $c_7 (pm)^{\frac{1}{q}} q^2 < pm\varepsilon/4$, or

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

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

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

for some $c_{13} > 0$. In this case, inequality (5.4) is satisfied when

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

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

$$c_{13} q^3 (pm)^{\frac{2}{q}} \left(\frac{d}{m} \right)^{\frac{1}{2}} \leq \frac{1}{4} pm \sqrt{\frac{c_{14} d}{m}}$$

or, $pm \geq c_{15} (pm)^{\frac{2}{q}} q^3$

for some $c_{15} > 0$.

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

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

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

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

which concludes the proof of Lemma 7.3 (see remaining details in Section 7.3). \square

5.2. Obtaining the differential inequality in Lemma 7.5. 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.

Sketch of proof of Lemma 7.5. There are two main ideas for obtaining this differential inequality. First, we use the cumulant method as in [13] to transform the derivative in t to matrix directional derivatives. Then, we bound the resulting terms in the expression by delicately using the matrix Hölder's inequality.

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

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

Now, recall that S_1 can be written in the form $\sum_{(l,\gamma) \in \Xi} Z_{(l,\gamma)}$ where $\Xi = [n] \times [pm]$ and $Z_{(l,\gamma)} = \xi_{(l,\gamma)} e_{\mu_{(l,\gamma)}} e_l^\top$ (see Definition 3.1). We then have the following lemma.

Lemma (Based on Corollary 6.1, [13]). *For any polynomial $\phi : M_{m \times d}(\mathbb{R}) \rightarrow \mathbb{R}$, we have*

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

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

Here, $P([k])$ denotes the set of all partitions of $[k]$, and $Z_{(l,\gamma),1|\pi}, \dots, Z_{(l,\gamma),k|\pi}$ are random matrices distributed as $Z_{(l,\gamma)}$. Crucially, those are independent of S_1, G_1, S_2 and G_2 (but not necessarily from each other). Further details are given in the full proof of Lemma 7.5 in Section 7.4.

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

$$\begin{aligned} \partial_{B_1} \cdots \partial_{B_k} f_{1,M_2}(M_1) \\ = \sum_{\sigma \in \text{sym}(k)} \sum_{\substack{r_1, \dots, r_{k+1} \geq 0 \\ r_1 + \dots + r_{k+1} = 2q - k}} \text{tr} \left(\Gamma(M_1, M_2)^{r_1} ((B_{\sigma(1)} U)^T M_2 U + (M_2 U)^T B_{\sigma(1)} U) \Gamma(M_1, M_2)^{r_2} \right. \\ \quad \left. ((B_{\sigma(2)} U)^T M_2 U + (M_2 U)^T B_{\sigma(2)} U) \cdots \Gamma(M_1, M_2)^{r_k} \right. \\ \quad \left. ((B_{\sigma(k)} U)^T M_2 U + (M_2 U)^T B_{\sigma(k)} U) \Gamma(M_1, M_2)^{r_{k+1}} \right). \end{aligned}$$

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

Then, estimating

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

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

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

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

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

Lemma (Matrix Hölder's inequality, Lemma 5.3 in [13]). *Let $1 \leq \beta_1, \dots, \beta_k \leq \infty$ satisfy $\sum_{i=1}^k \frac{1}{\beta_i} = 1$. Then*

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

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

This analysis based on matrix Hölder's inequality is done in Lemma Lemma 7.8. One important observation we use in this lemma (among many others) is that $\Theta_{(l,\gamma),\lambda,1}^T \Theta_{(l,\gamma),\lambda,2}$ are rank one matrices, which allows us to bound $\|\Theta_{(l,\gamma),\lambda,1}^T \Theta_{(l,\gamma),\lambda,2}\|_q$ with \sqrt{pd} instead of \sqrt{pm} . For further details, please refer to the full proof of Lemma 7.8 in Section 7.5. \square

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 $\text{Var}(s_{ij}) = p$ for all $i \in [m], j \in [n]$, and $\text{Cov}(s_{i_1 j_1}, s_{i_2 j_2}) = 0$ 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. Recall that in the unscaled OSNAP distribution, each subcolumn

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

of S for $\gamma \in [s], j \in [n]$ has the one hot distribution, and all these subcolumns are jointly $\log(mn)$ -independent. Therefore, the same argument as above, we directly calculate the variances of the entries

$$\mathbb{E}(s_{ij}^2) = \mathbb{P}(s_{ij} \neq 0) = \frac{1}{\frac{m}{pm}} = p$$

For the covariances, we first observe that $\text{Cov}(s_{i_1 j_1}, s_{i_2 j_2}) = 0$ if (i_1, j_1) and (i_2, j_2) belong to two different subcolumns by log-independence. More precisely, $\log(d/\varepsilon\delta)$ -independence implies 2-independence which implies zero covariance. If (i_1, j_1) and (i_2, j_2) belong to the same subcolumn, we have

$$\text{Cov}(s_{i_1 j_1}, s_{i_2 j_2}) = \mathbb{E}(s_{i_1 j_2} s_{i_2 j_1}) = \mathbb{E}(0) = 0$$

because each subcolumn has one hot distribution which means at most one of $s_{i_1 j_1}$ and $s_{i_2 j_2}$ can be nonzero. \square

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. Gaussian entries with variance p . Let U be an $n \times d$ matrix such that $U^T U = I$. 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. By Hölder's inequality, it suffices to prove these bounds for moments of the order of the smallest even integer bigger than q , so without loss of generality, we may assume that q is an even integer.

$$\begin{aligned} \mathbb{E}_{\mu, S(t)}[\|e_\mu^T S(t)U\|^q] &= \frac{1}{|J|} \sum_{j \in J} \mathbb{E}_{S(t)}[\|e_j^T S(t)U\|^q] \\ &= \mathbb{E}_{S(t)}[\|e_1^T S(t)U\|^q] \end{aligned}$$

since rows of $S(t)U$ are identically distributed.

Note that $S_{1,1}$ (which is the $(1,1)$ th entry of S) is non-zero when the one hot distribution on the column submatrix $S_{[1:m/s] \times 1}$ has its non zero entry on the first row. The probability that this happens is $1/(m/s) = s/m = p$. Moreover, the columns of S are independent. Thus, we conclude that for looking at the distribution of $e_1^T S(t)U$ we may assume that S has independent ± 1 entries sparsified by independent Bernoulli p random variables, i.e., $s_{i,j} = \delta_{i,j} \xi_{i,j}$ where $\delta_{i,j}$ are Bernoulli random variables taking value 1 with probability $p \in (0, 1]$, $\xi_{i,j}$ are random variables with $\mathbb{P}(\xi_{i,j} = 1) = \mathbb{P}(\xi_{i,j} = -1) = 1/2$ and the collection $\{\delta_{i,j}, \xi_{i,j}\}_{i \in [m], j \in [n]}$ is independent.

Now, conditioned on G , $\|e_1^T S(t)U\|$ is a convex function of the entries of the first row of S , and

$$\begin{aligned} E_S[\|e_1^T S(t)U\|] &\leq E_S[\|e_1^T S(t)U\|^2]^{\frac{1}{2}} \\ &= \sqrt{t \cdot pd + (1-t)\|e_1^T GU\|^2} \end{aligned}$$

By [32, Theorem 6.10], conditioned on G , we have, for $\lambda > 0$,

$$\mathbb{P}_S(\|e_1^T S(t)U\| \geq \sqrt{t \cdot pd + (1-t)\|e_1^T GU\|^2} + \lambda) \leq \exp(-c_1 \lambda^2)$$

for some $c_1 > 0$. Next, let $G = \sqrt{p}H$, where H is an $m \times d$ matrix with independent standard normal entries. By orthogonal invariance of Gaussians, the random vector $e_1^T HU$ is a d dimensional vector with independent standard normal entries. By concentration of norm for Gaussian vectors, [33, Theorem 3.1.1], we have

$$\mathbb{P}(\|e_1^T HU\| \geq \sqrt{d} + \lambda) \leq \exp(-c_2 \lambda^2)$$

which after scaling becomes

$$\mathbb{P}(\|e_1^T GU\| \geq \sqrt{pd} + \lambda\sqrt{p})$$

Therefore, we have

$$\mathbb{P}(\|e_1^T GU\| \geq \sqrt{pd} + \lambda) \leq \exp(-c_2 \lambda^2)$$

which after taking square becomes

$$\mathbb{P}(\|e_1^T GU\|^2 \geq 2pd + 2\lambda^2) \leq \exp(-c_2 \lambda^2)$$

for some $c_2 > 0$.

Let \mathcal{E} be the event that $\|e_1^T GU\|^2 \geq 2pd + 2\lambda^2$. Then, $\mathbb{P}(\mathcal{E}) \leq \exp(-c_2 \lambda^2)$ and we have,

$$\begin{aligned} &\mathbb{P}(\|e_1^T S(t)U\| \geq \sqrt{2pd} + \sqrt{2}\lambda) \\ &\leq \mathbb{P}(\|e_1^T S(t)U\| \geq \sqrt{2pd + 2\lambda^2}) \\ &\leq \mathbb{P}(\mathcal{E}) + \mathbb{P}(\|e_1^T S(t)U\| \geq \sqrt{2pd + 2\lambda^2} | \mathcal{E}^C) \end{aligned}$$

When $\|e_1^T GU\|^2 \leq 2pd + 2\lambda^2$, $\sqrt{t \cdot pd + (1-t)\|e_1^T GU\|^2} \leq \sqrt{2pd + 2\lambda^2}$, so

$$\begin{aligned} & \mathbb{P}(\|e_1^T S(t)U\| \geq \sqrt{2pd} + \sqrt{2\lambda}) \\ & \leq \mathbb{P}(\mathcal{E}) + \mathbb{P}(\|e_1^T S(t)U\| \geq \sqrt{2pd + 2\lambda^2} | \mathcal{E}^C) \\ & \leq \mathbb{P}(\mathcal{E}) + \mathbb{P}(\|e_1^T S(t)U\| \geq \sqrt{t \cdot pd + (1-t)\|e_1^T GU\|^2} | \mathcal{E}^C) \\ & \leq \exp(-c_2\lambda^2) + \exp(-c_1\lambda^2) \\ & \leq \exp(-c_3\lambda^2) \end{aligned}$$

for $\lambda \geq \sqrt{pd} \geq 1$ and some constant $c_3 > 0$, i.e. $\|e_1^T S(t)U\|$ has a subgaussian tail. By standard moment computation of subgaussian random variables, we conclude that $\mathbb{E}[\|e_1^T S(t)U\|^q] \leq (\sqrt{2pd})^q + (c_4q)^{\frac{q}{2}}$ for some $c_4 \geq 0$. Thus,

$$\mathbb{E}[\|e_1^T S(t)U\|^q] \leq (c_5 \sqrt{\max\{pd, q\}})^q$$

for some constant c_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 $\text{tr}(M) = \frac{1}{d} \text{Tr}(M)$. Let $L_q(S_q^d)$ be the normed vector space of $d \times d$ random matrices M with norm

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

More precisely, the space $L_q(S_q^d)$ consists of random matrices M with $\|M\|_q$ well defined (which means $(\mathbb{E}[\text{tr}|M|^q]) < \infty$ 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 $1 \leq \beta_1, \dots, \beta_k \leq \infty$ satisfy $\sum_{i=1}^k \frac{1}{\beta_i} = 1$. Then*

$$|\mathbb{E}[\text{tr} Y_1 \cdots Y_k]| \leq \|Y_1\|_{\beta_1} \cdots \|Y_k\|_{\beta_k}$$

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 β_1, \dots, β_k are extreme exponents and then extend the result to general β_1, \dots, β_k by the following convexity result. Later we will also use this method to prove our trace inequalities (Lemma 7.8).

Lemma 6.4 (Lemma 5.2. in [13]). *Let $F : (L_{\infty}(S_{\infty}^d))^k \rightarrow \mathbb{C}$ 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 $T : (A_{(1,0)} \cap A_{(1,0)}) \oplus \cdots \oplus (A_{(n,0)} \cap A_{(n,0)}) \rightarrow (B_0 \cap B_1)$ is multilinear. Also,*

assume that for any $(a_1, \dots, a_n) \in (A_{(1,0)} \cap A_{(1,0)}) \oplus \dots \oplus (A_{(n,0)} \cap A_{(n,0)})$, we have

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

and

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

Then for any $0 \leq \theta \leq 1$, the function T may be uniquely extended to a multilinear mapping from $(A_{(1,0)}, A_{(1,0)})_{[\theta]} \oplus \dots \oplus (A_{(n,0)}, A_{(n,0)})_{[\theta]}$ 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*

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

Assume that

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

is multilinear with

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

and

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

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

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

with

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

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

Proof. By [35], the spaces $L_{\beta_\nu(\theta)}(S_{\beta_\nu(\theta)})$ are complex interpolation spaces of the compatible Banach spaces $(L_{\beta_\nu(0)}(S_{\beta_\nu(0)}), L_{\beta_\nu(1)}(S_{\beta_\nu(1)}))$. The result follows from Theorem 6.5. \square

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 $m \geq \frac{c_{6.8}d}{\varepsilon^2}$,*

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

Proof. By lemma 6.7, we have

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

which after scaling becomes

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

And therefore, we have

$$\mathbb{P}(\|G^T G - I_d\| \leq (1 + \sqrt{\frac{d}{m}} + \frac{t}{\sqrt{m}})^2 - 1) \geq 1 - 2e^{-t^2/2}$$

Let $s = \sqrt{\frac{d}{m}} + \frac{t}{\sqrt{m}}$. We have

$$\begin{aligned} \mathbb{P}(\|G^T G - I_d\| \geq 2s + s^2) &\leq 2\exp(-\frac{1}{2}(\sqrt{m}s - \sqrt{d})^2) \\ &\leq \exp(-\frac{m}{2}(s - \sqrt{d/m})^2) \end{aligned}$$

When $m \geq \frac{d}{\varepsilon^2}$, $\sqrt{d/m} \leq \varepsilon$, the above result becomes

$$\mathbb{P}(\|G^T G - I_d\| \geq 2s + s^2) \leq \exp(-\frac{m}{2}(s - \varepsilon)^2)$$

when $s \geq \varepsilon$.

In conclusion, we have

$$\mathbb{P}(\|G^T G - I_d\| \geq 3s) \leq \exp(-\frac{m}{8}s^2) \quad \text{when } 2\varepsilon \leq s < 1$$

and

$$\mathbb{P}(\|G^T G - I_d\| \geq 3s^2) \leq \exp(-\frac{m}{8}s^2) \quad \text{when } s \geq 1$$

By change of variables, we have

$$\mathbb{P}(\|G^T G - I_d\| \geq s') \leq \exp(-\frac{m}{72}s'^2) \quad \text{when } 6\varepsilon \leq s' < 3$$

and

$$\mathbb{P}(\|G^T G - I_d\| \geq s') \leq \exp(-\frac{m}{24}s') \quad \text{when } s' \geq 3$$

Denote the random variable $\|G^T G - I_d\|$ by R . Then,

$$\begin{aligned}\mathbb{E}(R^{2q}) &= 2q \int_0^\infty t^{2q-1} \mathbb{P}(R \geq t) dt \\ &\leq 2q(6\varepsilon)^{2q} + 2q \int_{6\varepsilon}^3 t^{2q-1} \exp\left(-\frac{m}{72}t^2\right) dt + 2q \int_3^\infty t^{2q-1} \exp\left(-\frac{m}{24}t\right) dt\end{aligned}$$

Performing change of variables $x = mt^2/72$ and $y = mt/24$ in the second and third integrals respectively,

$$\begin{aligned}\mathbb{E}(R^{2q}) &\leq 2q(6\varepsilon)^{2q} + 2q \int_0^\infty \left(\frac{72x}{m}\right)^{\frac{2q-1}{2}} \exp(-x) \frac{72}{2m} \left(\frac{72x}{m}\right)^{-1/2} dx \\ &\quad + 2q \int_0^\infty \left(\frac{24y}{m}\right)^{2q-1} \exp(-y) \frac{24}{m} dy \\ &\leq 2q(6\varepsilon)^{2q} + q \left(\frac{72}{m}\right)^q \int_0^\infty x^{q-1} \exp(-x) dx + 2q \left(\frac{24}{m}\right)^{2q} \int_0^\infty y^{2q-1} \exp(-y) dy \\ &\leq 2q(6\varepsilon)^{2q} + q \left(\frac{72}{m}\right)^q (q-2)! + 2q \left(\frac{24}{m}\right)^{2q} (2q-2)! \\ &\leq 2q(6\varepsilon)^{2q} + \left(\frac{72q}{m}\right)^q + \left(\frac{48q}{m}\right)^{2q}\end{aligned}$$

Since $q \leq m\varepsilon^2$, $\left(\frac{72q}{m}\right)^q \leq (72\varepsilon^2)^q$ and $\left(\frac{48q}{m}\right)^{2q} \leq (48\varepsilon^2)^{2q} \leq (48\varepsilon)^{2q}$. So,

$$\begin{aligned}\mathbb{E}(R^{2q}) &\leq (2q+2)(48\varepsilon)^{2q} \\ \mathbb{E}(R^{2q})^{\frac{1}{2q}} &\leq (2q+2)^{\frac{1}{2q}} 48\varepsilon\end{aligned}$$

Now, $48(2q+2)^{\frac{1}{2q}} < C$, for some $C > 0$. Thus,

$$\mathbb{E}(R^{2q})^{\frac{1}{2q}} \leq C\varepsilon$$

for some $C > 1$ independent of ε . When $m \geq C^2 \frac{d}{\varepsilon^2}$, we can run the same argument with ε/C in place of ε to get,

$$\mathbb{E}(R^{2q})^{\frac{1}{2q}} \leq \varepsilon$$

□

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*

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

Proof. We wish to bound,

$$\begin{aligned}\mathbb{E} \left[\text{tr} (G_1^T G_2 + G_2^T G_1)^{2q} \right]^{\frac{1}{2q}} &\leq \mathbb{E} \left[\|G_1^T G_2 + G_2^T G_1\|^{2q} \right]^{\frac{1}{2q}} \\ &\leq \mathbb{E} \left[(\|G_1^T G_2\| + \|G_2^T G_1\|)^{2q} \right]^{\frac{1}{2q}} \\ &\leq \mathbb{E} \left[\|G_2^T G_1\|^{2q} \right]^{\frac{1}{2q}} + \mathbb{E} \left[\|G_1^T G_2\|^{2q} \right]^{\frac{1}{2q}}\end{aligned}$$

Since $G_2^T G_1$ and $G_1^T G_2$ are identically distributed, it is enough to bound $\mathbb{E} [\|G_1^T G_2\|^{2q}]^{\frac{1}{2q}}$. We shall first condition on G_2 and compute $\mathbb{E}_{G_1} [\|G_1^T G_2\|^{2q}]$. Let $G_2 = U|G_2|V^T$ be the SVD of G_2 where U is an $m \times d$ matrix with orthogonal columns. Then,

$$\begin{aligned} \mathbb{E}_{G_1} [\|G_1^T G_2\|^{2q}] &\leq \mathbb{E}_{G_1} [\|G_1^T U |G_2| V^T\|^{2q}] \\ &\leq \|G_2\|^{2q} \mathbb{E}_{G_1} [\|G_1^T U\|^{2q}] \end{aligned}$$

By orthogonal invariance of Gaussians, $G_1^T U$ is a $d \times d$ Gaussian matrix. Thus,

$$\mathbb{E}_{G_1} [\|G_1^T U\|^{2q}] \leq (c_1 \sqrt{\max\{d, q\}})^{2q}$$

for some constant $c_1 > 0$, by applying Proposition 2.1 and Lemma 2.2 from [37] to

$$\text{sym}(G_1^T U) := \begin{bmatrix} 0 & G_1^T U \\ G_1^T U^T & 0 \end{bmatrix}.$$

Similarly, we can get

$$\mathbb{E}[\|G_2\|^{2q}] \leq (c_2 \sqrt{\max\{m, q\}})^{2q}$$

□

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 $(SU)^T S U - pm \cdot I_d$ to controlling moments of $(S_1 U)^T S_2 U + (S_2 U)^T S_1 U$, for independent S_1 and S_2 .
- Section 7.3 proves the trace moment bound using 2D interpolation of moments of the form $(S_1(t)U)^T S_2(t)U + (S_2(t)U)^T S_1(t)U$.
- 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*

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

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

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

$$\mathbb{P}(\|X^T X - I_d\| \leq \varepsilon) \geq 1 - \delta$$

because

$$\|X^T X - I_d\| < \varepsilon$$

will imply

$$1 - \varepsilon \leq v^T X^T X v \leq 1 + \varepsilon$$

and

$$(1 - \varepsilon)^2 \leq \|Xv\|^2 \leq (1 + \varepsilon)^2$$

for any $v \in \mathbb{R}^d$, and then the claim about singular values follows from the min-max principle. To show

$$\mathbb{P}(\|X^T X - I_d\| \leq \varepsilon) \geq 1 - \delta$$

we first use Markov's inequality and obtain

$$\mathbb{P}(\|X^T X - I_d\|^{2q} \geq \delta^{-1} \mathbb{E}[\|X^T X - I_d\|^{2q}]) \leq \delta$$

Taking the $(2q)$ th root, we have

$$\mathbb{P}\left(\|X^T X - I_d\| \geq \delta^{-\frac{1}{2q}} \mathbb{E}[\|X^T X - I_d\|^{2q}]^{\frac{1}{2q}}\right) \leq \delta$$

We observe that

$$\|X^T X - I_d\|^{2q} = (s_{\max}(X^T X - I_d))^{2q} \leq \sum_{i=1}^d (s_i(X^T X - I_d))^{2q} = d \operatorname{tr}(X^T X - I_d)^{2q}$$

Therefore, we have

$$\mathbb{P}\left(\|X^T X - I_d\| \geq \delta^{-\frac{1}{2q}} \mathbb{E}[d \operatorname{tr}(X^T X - I_d)^{2q}]^{\frac{1}{2q}}\right) \leq \delta$$

which, after simplification, becomes

$$\mathbb{P}\left(\|X^T X - I_d\| \geq (d/\delta)^{\frac{1}{2q}} \mathbb{E}[\operatorname{tr}(X^T X - I_d)^{2q}]^{\frac{1}{2q}}\right) \leq \delta$$

For $q > \log(\frac{d}{\delta})$, we have

$$(d/\delta)^{\frac{1}{2q}} = \exp(\log(d/\delta) \frac{1}{2q}) \leq \exp(\log(d/\delta) \frac{1}{2 \log(\frac{d}{\delta})}) \leq \sqrt{e}$$

Therefore, we have

$$\mathbb{P}\left(\|X^T X - I_d\| \geq \sqrt{e} \mathbb{E}[\operatorname{tr}(X^T X - I_d)^{2q}]^{\frac{1}{2q}}\right) \leq \delta$$

By Theorem 7.3, when $m \geq c_{7.3.1} \frac{d+q}{(\varepsilon/\sqrt{e})^2}$ and $q \in \mathbb{N}$ satisfying $2 \leq q \leq m$ and

$$pm \geq (\max\{\frac{c_{7.3.2} q^2}{(\varepsilon/\sqrt{e})}, c_{7.3.3} q^3\})^{1+\frac{2}{q-2}}$$

we have

$$\mathbb{E}[\operatorname{tr}(X^T X - I_d)^{2q}]^{\frac{1}{2q}} \leq (\varepsilon/\sqrt{e})$$

which will imply

$$\mathbb{P}(\|X^T X - I_d\| \geq \varepsilon) \leq \delta$$

when combined with the previous calculations. To ensure that the assumptions of Theorem 7.3 are satisfied, we need to choose q such that

$$pm \geq (\max\left\{\frac{c_{7.3.2} \sqrt{e} q^2}{\varepsilon}, c_{7.3.3} q^3\right\})^{1+\frac{2}{q-2}}$$

Without loss of generality, we assume that $c_{7.3.2} > 1$ and $c_{7.3.3} > 1$. In addition, we recall that we also need the requirement $q > \log(\frac{d}{\delta})$ for one of the previous steps.

Now, we claim that it suffices to require $m \geq 8c_{7.3.1} \frac{d+\log(d/\varepsilon\delta)}{(\varepsilon/\sqrt{e})^2}$, and

$$(7.1) \quad pm \geq \max \{c_{7.3.2}e8^2e^3, c_{7.3.3}8^3e^3\} \left(\max \left\{ \frac{c_{7.3.2}e(8\log(\frac{d}{\varepsilon\delta}))^2}{\varepsilon}, c_{7.3.3}(8\log(\frac{d}{\varepsilon\delta}))^3 \right\} \right)$$

and choose

$$q = \lceil 2\log(\frac{d}{\varepsilon\delta}) \rceil + 2$$

to obtain the desired high probability bound. Since the quantity $\mathbb{E}[\text{tr}(X^T X - I_d)^{2q}]$ remains unchanged if we only assume that subsets of the entries of X of size $2q$ are independent instead of arbitrary subsets of the entries of X being independent and we choose $q = \lceil 2\log(\frac{d}{\varepsilon\delta}) \rceil + 2$, we only need S to be an $O(\log(d/\varepsilon\delta))$ -wise independent unscaled OSNAP.

Now we will check that this choice really works. First, by definition, we have $q > \log(d/\delta)$ and the lower bound on m implies $q < m$, so we still have

$$\mathbb{P} \left(\|X^T X - I_d\| \geq \sqrt{e} \mathbb{E}[\text{tr}(X^T X - I_d)^{2q}]^{\frac{1}{2q}} \right) \leq \delta$$

So it remains to check

$$pm \geq \left(\max \left\{ \frac{c_{7.3.2}eq^2}{\varepsilon}, c_{7.3.3}q^3 \right\} \right)^{\frac{q_1}{q_1-2}}$$

Note that $q = 2\lceil \log(\frac{d}{\varepsilon\delta}) \rceil + 2 < 8\log(\frac{d}{\varepsilon\delta})$ because $d > 10$, and therefore the above requirement is implied by

$$pm \geq \left(\max \left\{ \frac{c_{7.3.2}e(8\log(\frac{d}{\varepsilon\delta}))^2}{\varepsilon}, c_{7.3.3}(8\log(\frac{d}{\varepsilon\delta}))^3 \right\} \right)^{1+\frac{2}{q-2}}$$

Since $q - 2 \geq 2\log(\frac{d}{\varepsilon\delta})$, we also claim that

$$\begin{aligned} & \left(\max \left\{ \frac{c_{7.3.2}e(8\log(\frac{d}{\varepsilon\delta}))^2}{\varepsilon}, c_{7.3.3}(8\log(\frac{d}{\varepsilon\delta}))^3 \right\} \right)^{\frac{2}{q-2}} \\ & \leq \left(\max \left\{ \frac{c_{7.3.2}e(8\log(\frac{d}{\varepsilon\delta}))^2}{\varepsilon}, c_{7.3.3}(8\log(\frac{d}{\varepsilon\delta}))^3 \right\} \right)^{\frac{1}{\log(\frac{d}{\varepsilon\delta})}} \end{aligned}$$

will be bounded by an explicit constant, and this is where we need the extra factor $\frac{1}{\varepsilon}$ inside the log. We observe that

$$\begin{aligned} & \left(\frac{(\log(\frac{d}{\varepsilon\delta}))^2}{\varepsilon} \right)^{\frac{1}{\log(\frac{d}{\varepsilon\delta})}} \\ & = \exp(2\log\log(\frac{d}{\varepsilon\delta})/\log(\frac{d}{\varepsilon\delta})) \cdot \left(\frac{1}{\varepsilon} \right)^{\frac{1}{\log(\frac{d}{\varepsilon\delta})}} \\ & \leq e^2 \cdot \left(\frac{1}{\varepsilon} \right)^{\frac{1}{\log(\frac{1}{\varepsilon})}} \\ & \leq e^3 \end{aligned}$$

and

$$\begin{aligned} & \left(\log(\frac{d}{\varepsilon\delta}) \right)^{\frac{3}{\log(\frac{d}{\varepsilon\delta})}} \\ & = \exp(3\log\log(\frac{d}{\varepsilon\delta})/\log(\frac{d}{\varepsilon\delta})) \leq e^3 \end{aligned}$$

where we use the fact that $\log \log(\frac{d}{\varepsilon\delta}) > 0$ when $d > 10$. Therefore, we have

$$\begin{aligned} & \left(\max \left\{ \frac{c_{7.3.2}e(8 \log(\frac{d}{\varepsilon\delta}))^2}{\varepsilon}, c_{7.3.3}(8 \log(\frac{d}{\varepsilon\delta}))^3 \right\} \right)^{\frac{2}{q-2}} \\ & \leq \left(\max \left\{ \frac{c_{7.3.2}e(8 \log(\frac{d}{\varepsilon\delta}))^2}{\varepsilon}, c_{7.3.3}(8 \log(\frac{d}{\varepsilon\delta}))^3 \right\} \right)^{\frac{1}{\log(\frac{d}{\varepsilon\delta})}} \\ & \leq \max \{ c_{7.3.2}e8^2e^3, c_{7.3.3}8^3e^3 \} \end{aligned}$$

Therefore it is enough to require,

$$pm \geq \max \{ c_{7.3.2}e8^2e^3, c_{7.3.3}8^3e^3 \} \left(\max \left\{ \frac{c_{7.3.2}e(8 \log(\frac{d}{\varepsilon\delta}))^2}{\varepsilon}, c_{7.3.3}(8 \log(\frac{d}{\varepsilon\delta}))^3 \right\} \right)$$

which is guaranteed by our assumption (7.1) on pm .

Moreover, recall that the definition of OSNAP requires $p \leq 1$. Therefore, if the requirement (7.1) already forces $p > 1$, then we cannot construct an OSNAP satisfying the requirement. This will generally not be an issue because lower bound in the requirement (7.1) grows much slower than m . However, when m is small, this might happen because the constants can be large. For the completeness of the result, we mention that we can always use $p = 1$ even when (7.1) requires $p > 1$. To prove this claim, it is enough to find values of m for which a dense $m \times n$ matrix populated with random signs satisfies the OSE property for a given $0 < \varepsilon, \delta < 1$, and this follows directly for $m \geq c_4 \cdot \frac{d + \log(1/\delta)}{\varepsilon^2}$ using [38, Theorem 4.6.1] for some constant c_4 . \square

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,

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

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

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

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

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

Remark 7.1. In Equation (7.2), we decomposed the the embedding error $U^T S^T S U - pm \cdot I_d$ 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

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

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*

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

Consequently, we have

$$\mathbb{E}[\text{tr}(U^T S^T S U - pm \cdot I_d)^{2q}] \leq \mathbb{E}_{S, S'} \left[\text{tr} \left(2((SU)^T S' U + (SU)^T S' U) \right)^{2q} \right]$$

where S' is an independent copy of S .

Proof. Let $\mathbf{w}^T = (w_1, \dots, w_n)$ be a vector of independent random variables such that $\mathbb{P}(w_i = 1) = \mathbb{P}(w_i = 0) = 1/2$. Then, whenever $j \neq j'$, $\mathbb{E}[\mathbf{1}_{w_j \neq w_{j'}}] = 1/2$. So we have,

$$\sum_{i=1}^m \sum_{j,j'=1, j \neq j'}^n s_{ij} s_{ij'} u_j u_{j'}^T = 2 \mathbb{E}_{\mathbf{w}} \left[\sum_{i=1}^m \sum_{j,j'=1}^n \mathbf{1}_{w_j \neq w_{j'}} s_{ij} s_{ij'} u_j u_{j'}^T \right]$$

By Jensen's inequality [39, Lemma 4.5, p.86]

$$\begin{aligned} \mathbb{E} \left[\text{tr} \left(\sum_{i=1}^m \sum_{j,j'=1, j \neq j'}^n s_{ij} s_{ij'} u_j u_{j'}^T \right)^{2q} \right] &= \mathbb{E}_S \left[\text{tr} \left(\mathbb{E}_{\mathbf{w}} \left[2 \sum_{i=1}^m \sum_{j,j'=1}^n \mathbf{1}_{w_j \neq w_{j'}} s_{ij} s_{ij'} u_j u_{j'}^T \right] \right)^{2q} \right] \\ &\leq \mathbb{E}_S \mathbb{E}_{\mathbf{w}} \left[\text{tr} \left(2 \sum_{i=1}^m \sum_{j,j'=1}^n \mathbf{1}_{w_j \neq w_{j'}} s_{ij} s_{ij'} u_j u_{j'}^T \right)^{2q} \right] \\ &\leq \mathbb{E}_{\mathbf{w}} \mathbb{E}_S \left[\text{tr} \left(2 \sum_{i=1}^m \sum_{j,j'=1}^n \mathbf{1}_{w_j \neq w_{j'}} s_{ij} s_{ij'} u_j u_{j'}^T \right)^{2q} \right] \end{aligned}$$

Fix \mathbf{w} , and let $J = \{j \in [n] | w_j = 1\}$. Then,

$$\begin{aligned} &\mathbb{E}_S \left[\text{tr} \left(2 \sum_{i=1}^m \sum_{j,j'=1}^n \mathbf{1}_{w_j \neq w_{j'}} s_{ij} s_{ij'} u_j u_{j'}^T \right)^{2q} \right] \\ &= \mathbb{E}_S \left[\text{tr} \left(2 \sum_{i=1}^m \left(\sum_{j \in J, j' \in J^C} s_{ij} s_{ij'} u_j u_{j'}^T + \sum_{j \in J^C, j' \in J} s_{ij} s_{ij'} u_j u_{j'}^T \right) \right)^{2q} \right] \end{aligned}$$

The key observation for decoupling is that the collection of random variables $\mathcal{S}_J := \{s_{ij}\}_{i \in [m], j \in J}$ is independent of the collection $\mathcal{S}_{J^C} := \{s_{ij}\}_{i \in [m], j \in J^C}$, so the above expectation does not change if the collection \mathcal{S}_J is replaced by an independent copy $\mathcal{S}'_J := \{s'_{ij}\}_{i \in [m], j \in J}$ thought of as coming

from the entries of an independent copy of S , say, S' , i.e.,

$$\begin{aligned}
& \mathbb{E}_{S_J, S_{J^C}} \left[\text{tr} \left(2 \sum_{i=1}^m \left(\sum_{j \in J, j' \in J^C} s_{ij} s_{ij'} u_j u_{j'}^T + \sum_{j \in J^C, j' \in J} s_{ij} s_{ij'} u_j u_{j'}^T \right) \right)^{2q} \right] \\
&= \mathbb{E}_{S'_J, S_{J^C}} \left[\text{tr} \left(2 \sum_{i=1}^m \left(\sum_{j \in J, j' \in J^C} s'_{ij} s_{ij'} u_j u_{j'}^T + \sum_{j \in J^C, j' \in J} s_{ij} s'_{ij'} u_j u_{j'}^T \right) \right)^{2q} \right] \\
&= \mathbb{E}_{S'_J, S_{J^C}} \left[\text{tr} \left(2 \sum_{i=1}^m \left(\sum_{j \in J, j' \in J^C} s'_{ij} s_{ij'} u_j u_{j'}^T + \sum_{j \in J^C, j' \in J} s_{ij} s'_{ij'} u_j u_{j'}^T \right. \right. \right. \\
&\quad \left. \left. + \mathbb{E}_{S_J} \left[\sum_{j \in J, j' \in J} s'_{ij} s_{ij'} u_j u_{j'}^T \right] + \mathbb{E}_{S_{J^C}, S_J} \left[\sum_{j \in J^C, j' \in [n]} s'_{ij} s_{ij'} u_j u_{j'}^T \right] \right. \right. \\
&\quad \left. \left. + \mathbb{E}_{S_{J^C}} \left[\sum_{j \in J^C, j' \in J^C} s_{ij} s'_{ij'} u_j u_{j'}^T \right] + \mathbb{E}_{S_J, S_{J^C}} \left[\sum_{j \in J, j' \in [n]} s_{ij} s'_{ij'} u_j u_{j'}^T \right] \right) \right)^{2q} \right] \\
&\leq \mathbb{E}_{S, S'} \left[\text{tr} \left(2 \sum_{i=1}^m \sum_{j, j'=1}^n \left(s_{ij} s'_{ij'} u_j u_{j'}^T + s'_{ij} s_{ij'} u_j u_{j'}^T \right) \right)^{2q} \right] \\
&= \mathbb{E}_{S, S'} \left[\text{tr} \left(2((SU)^T S'U + (S'U)^T SU) \right)^{2q} \right]
\end{aligned}$$

where in the second equality the expectations we add are all 0, and the inequality step follows by taking the expectations outside using Jensen's inequality. \square

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 $U^T U = I$. 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 $2 \leq q \leq m$, when $m \geq c_{7.3.1} \frac{d+q}{\varepsilon^2}$ and $pm \geq (\max\{\frac{c_{7.3.2} q^2}{\varepsilon}, c_{7.3.3} q^3\})^{1+\frac{2}{q-2}}$, we have*

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

Proof. In Section 7.2, we saw that when S has the unscaled OSNAP distribution

$$\mathbb{E}[\text{tr}((SU)^T(SU) - pm \cdot I_d)^{2q}]^{\frac{1}{2q}} \leq 2 \mathbb{E}[\text{tr}(\Gamma(S_1, S_2))^{2q}]^{\frac{1}{2q}}$$

where S_1 and S_2 are independent random matrices with the unscaled OSNAP distribution and $\Gamma(M_1, M_2) = (M_1 U)^T (M_2 U) + (M_2 U)^T (M_1 U)$. Let $\Gamma(t)$ be as defined in Lemma 7.5. Then, to show the statement of the theorem, it is enough to show that $\mathbb{E}[\text{tr}(\Gamma(1))^{2q}]^{\frac{1}{2q}} \leq pm\varepsilon/2$. Now, by Lemma 6.9, we know that,

$$\mathbb{E}[\text{tr}(\Gamma(0))^{2q}]^{\frac{1}{2q}} = \mathbb{E}[\text{tr}(\Gamma(G_1, G_2))^{2q}]^{\frac{1}{2q}} \leq c_{6.9} p \sqrt{\max\{d, q\}} \sqrt{\max\{m, q\}}$$

We want to find conditions for which $\mathbb{E}[\text{tr}(\Gamma(0))^{2q}]^{\frac{1}{2q}} \leq pm\varepsilon/4$, for which it is enough to ensure $c_{6.9} p \sqrt{\max\{d, q\}} \sqrt{\max\{m, q\}} \leq pm\varepsilon/4$. Clearly, this can only happen when $q \leq m$, and in this case the inequality holds when $m \geq \frac{c(d+q)}{\varepsilon^2}$. Thus, it suffices to show

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

By Lemma 7.5, if q satisfies $q \geq 2$, we have

$$\frac{d}{dt} \mathbb{E}[\text{tr } \Gamma(t)^{2q}] \leq \max_{4 \leq k \leq 2q} (c_5 q)^k ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{qk-2q}{q-1}} (\mathbb{E}[f(S_1(t), S_2(t))])^{1-\frac{k-2}{2q-2}}$$

We use different calculations for two different cases.

In the first case, we consider $pd < q$.

By convexity, the maximum must be attained on one of the two end points. Also, we observe that when $k = 2q$, we have $\frac{k-2}{2q-2} = 1$. Therefore, we have

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[\text{tr } \Gamma(t)^{2q}] &\leq \max_{4 \leq k \leq 2q} (c_5 q)^k ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{qk-2q}{q-1}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{k-2}{2q-2}} \\ &\leq \max\{(c_5 q)^4 ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{q4-2q}{q-1}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{2}{2q-2}}, \\ &\quad (c_5 q)^{2q} ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{q(2q)-2q}{q-1}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{2q-2}{2q-2}}\} \\ &= (c_5 q)^4 ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{2q}{q-1}} \max\{(\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{2}{2q-2}}, \\ &\quad (c_5 q)^{2q-4} ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{2q^2-4q}{q-1}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{2q-2}{2q-2}}\} \\ &= (c_5 q)^4 ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{2q}{q-1}} \max\{(\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{2}{2q-2}}, \\ &\quad (c_5 q)^{2q-4} ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{2q^2-4q}{q-1}}\} \\ &= (c_5 q)^4 ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{2q}{q-1}} \max\{(\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{1}{q-1}}, \\ &\quad (c_5 q)^{2q-4} ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{2q^2-4q}{q-1}}\} \end{aligned}$$

We will use Lemma 6.6 in [13] with $\alpha = \frac{1}{q-1}$. In this case, we have $1 - \alpha = \frac{q-2}{q-1}$, and $\frac{\alpha}{1-\alpha} = \frac{1}{q-2}$. Also, we want $K^{1-\alpha} = (c_5 q)^{2q-4} ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{2q(2q)-8q}{2q-2}}$, so we need $K = ((c_5 q)^{2q-4} ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{2q^2-4q}{q-1}})^{\frac{1}{1-\alpha}}$.

By Lemma 6.6 in [13], we have

$$\begin{aligned} &(\mathbb{E}[\text{tr } \Gamma(S_1, S_2)^{2q}])^\alpha - (\mathbb{E}[\text{tr } \Gamma(G_1, G_2)^{2q}])^\alpha \\ &\leq (c_5 q)^4 ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{2q}{q-1}} \cdot \frac{1}{q-1} + ((c_5 q)^{2q-4} ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{2q^2-4q}{q-1}})^{\frac{\alpha}{1-\alpha}} \\ &\leq (c_5 q)^4 ((pm)^{\frac{1}{q}} \sqrt{q})^{\frac{2q}{q-1}} \cdot \frac{1}{q-1} + ((c_5 q)^{2q-4} ((pm)^{\frac{1}{q}} \sqrt{q})^{\frac{2q^2-4q}{q-1}})^{\frac{1}{q-2}} \\ &= (c_5 q)^4 (pm)^{\frac{2}{q-1}} q(q)^{\frac{1}{q-1}} \cdot \frac{1}{q-1} + (c_5 q)^2 (pm)^{\frac{2}{q-1}} q^{\frac{q}{q-1}} \\ &\leq c_6 (pm)^{\frac{2}{q-1}} q^4 \end{aligned}$$

Therefore, we have

$$(\mathbb{E}[\text{tr } \Gamma(S_1, S_2)^{2q}])^{\frac{1}{q-1}} - (\mathbb{E}[\text{tr } \Gamma(G_1, G_2)^{2q}])^{\frac{1}{q-1}} \leq c_6 (pm)^{\frac{2}{q-1}} q^4$$

By Taylor expansion, we have

$$(a+b)^{\frac{2q}{q-1}} \geq a^{\frac{2q}{q-1}} + b^{\frac{2q}{q-1}}$$

Using change of variable $c = a^{\frac{2q}{q-1}} + b^{\frac{2q}{q-1}}$ and $d = a^{\frac{2q}{q-1}}$, we have

$$c^{\frac{q-1}{2q}} - d^{\frac{q-1}{2q}} \leq (c-d)^{\frac{q-1}{2q}}$$

Therefore, we have

$$(\mathbb{E}[\text{tr } \Gamma(S_1, S_2)^{2q}])^{\frac{1}{2q}} - (\mathbb{E}[\text{tr } \Gamma(G_1, G_2)^{2q}])^{\frac{1}{2q}} \leq (c_6(pm)^{\frac{2}{q-1}}q^4)^{\frac{q-1}{2q}} \leq c_7(pm)^{\frac{1}{q}}q^2$$

By previous analysis, the desired requirement $(\mathbb{E}[\text{tr } \Gamma(S_1, S_2)^{2q}])^{\frac{1}{2q}} - (\mathbb{E}[\text{tr } \Gamma(G_1, G_2)^{2q}])^{\frac{1}{2q}} \leq \frac{1}{4}pm\varepsilon$ will be satisfied when $c_7(pm)^{\frac{1}{q}}q^2 \leq pm\varepsilon$ which means

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

In the second case, we consider $pd > q$. In this case, we have some pd factors in the bound for $\frac{d}{dt} \mathbb{E}[\text{tr } \Gamma(t)^{2q}]$. We will eventually get rid of these pd factors by replacing it by some powers of $\mathbb{E}[\text{tr } \Gamma(t)^{2q}]$ and some extra factors. To illustrate this idea, we start from following lemma to calculate $\mathbb{E}[\text{tr } \Gamma(t)^2]$.

Lemma 7.4 (Second Moments). *Let M and N be $m \times d$ independent random matrices such that the entries of M (and N) are uncorrelated and have variance p . Then,*

- $\mathbb{E}[\text{tr } M^T N M^T N] = p^2 m$.
- $\mathbb{E}[\text{tr } M^T N N^T M] = p^2 m d$

Proof. We have,

$$\mathbb{E}[\text{tr } M^T N M^T N] = \frac{1}{d} \sum_{i=1}^d \sum_{j,k,l} \mathbb{E}[m_{ji} n_{jk} m_{lk} n_{li}]$$

By uncorrelatedness of entries, the expectation is non-zero only when $j = l$ and $k = i$. So,

$$\begin{aligned} \mathbb{E}[\text{tr } M^T N M^T N] &= \frac{1}{d} \sum_{i=1}^d \sum_{j=1}^m \mathbb{E}[m_{ji}^2 n_{ji}^2] \\ &= p^2 m \end{aligned}$$

Similarly,

$$\begin{aligned} \mathbb{E}[\text{tr } M^T N N^T M] &= \frac{1}{d} \sum_{i=1}^d \sum_{j,k,l} \mathbb{E}[m_{ji} n_{jk} n_{lk} m_{li}] \\ &= \frac{1}{d} \sum_{i=1}^d \sum_{j,k} \mathbb{E}[m_{ji}^2 n_{jk}^2] \\ &= p^2 m d \end{aligned}$$

□

By Lemma 7.4, we have $\mathbb{E}[\text{tr } \Gamma(t)^2] = 2p^2 m d + 2p^2 m \geq 2p^2 m d$. By Hölder's inequality,

$$\sqrt{2p^2 m d} \leq \mathbb{E}[\text{tr } \Gamma(t)^2]^{\frac{1}{2}} \leq \mathbb{E}[\text{tr } \Gamma(t)^{2q}]^{\frac{1}{2q}}$$

In this case, we can replace the extra pd factors and obtain

$$\begin{aligned}
\frac{d}{dt} \mathbb{E}[\text{tr } \Gamma(t)^{2q}] &\leq \max_{4 \leq k \leq 2q} (c_5 q)^k ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{2qk-4q}{2q-2}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{k-2}{2q-2}} \\
&\leq \max_{4 \leq k \leq 2q} (c_5 q)^k ((pm)^{\frac{1}{q}} \sqrt{pd})^{\frac{2qk-4q}{2q-2}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{k-2}{2q-2}} \\
&\leq \max_{4 \leq k \leq 2q} (c_5 q)^k ((pm)^{\frac{1}{q}} (2p^2 md)^{1/4} (\frac{1}{2} \frac{d}{m})^{1/4})^{\frac{2qk-4q}{2q-2}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{k-2}{2q-2}} \\
&\leq \max_{4 \leq k \leq 2q} (c_5 q)^k ((pm)^{\frac{1}{q}} (\mathbb{E}[\text{tr } \Gamma(t)^{2q}]^{\frac{1}{2q}})^{1/2} (\frac{1}{2} \frac{d}{m})^{1/4})^{\frac{2qk-4q}{2q-2}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{k-2}{2q-2}} \\
&= \max_{4 \leq k \leq 2q} (c_5 q)^k (pm)^{\frac{k-2}{q-1}} (\mathbb{E}[\text{tr } \Gamma(t)^{2q}])^{\frac{k-2}{4(q-1)}} (\frac{1}{2} \frac{d}{m})^{\frac{qk-2q}{4(q-1)}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{k-2}{2(q-1)}} \\
&= \max_{4 \leq k \leq 2q} (c_5 q)^k (pm)^{\frac{k-2}{q-1}} (\frac{1}{2} \frac{d}{m})^{\frac{qk-2q}{4(q-1)}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{2(k-2)}{4(q-1)} + \frac{k-2}{4(q-1)}} \\
&= \max_{4 \leq k \leq 2q} (c_5 q)^k (pm)^{\frac{k-2}{q-1}} (\frac{1}{2} \frac{d}{m})^{\frac{qk-2q}{4(q-1)}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{(k-2)}{4(q-1)}}
\end{aligned}$$

If we want $1 - \frac{(k-2)}{4(q-1)} = 0$, then we need $(k-2) = 4(q-1)$. Therefore, we can choose $k = 4(q-1) + 2 = 4q - 2 > 2q$ when $q > 1$.

Using the convexity of the function $a^x b^{1-x}$ in x as before, we have

$$\begin{aligned}
\frac{d}{dt} \mathbb{E}[\text{tr } \Gamma(t)^{2q}] &\leq \max_{4 \leq k \leq 2q} (c_5 q)^k (pm)^{\frac{k-2}{q-1}} (\frac{1}{2} \frac{d}{m})^{\frac{qk-2q}{4(q-1)}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{(k-2)}{4(q-1)}} \\
&\leq \max_{4 \leq k \leq (4q-2)} (c_5 q)^k (pm)^{\frac{k-2}{q-1}} (\frac{1}{2} \frac{d}{m})^{\frac{qk-2q}{4(q-1)}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{(k-2)}{4(q-1)}} \\
&\leq \max\{(c_5 q)^4 (pm)^{\frac{4-2}{q-1}} (\frac{1}{2} \frac{d}{m})^{\frac{q(4-2)}{4(q-1)}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{(4-2)}{4(q-1)}}, \\
&\quad (c_5 q)^{4q-2} (pm)^{\frac{4q-2-2}{q-1}} (\frac{1}{2} \frac{d}{m})^{\frac{q(4q-2)-2q}{4(q-1)}}\} \\
&= (c_5 q)^4 (pm)^{\frac{2}{q-1}} (\frac{1}{2} \frac{d}{m})^{\frac{q(4-2)}{4(q-1)}} \max\{(\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{(4-2)}{4(q-1)}}, \\
&\quad (c_5 q)^{4q-2-4} (pm)^{\frac{4q-6}{q-1}} (\frac{1}{2} \frac{d}{m})^{\frac{q(4q-2)-2q}{4(q-1)} - \frac{q(4-2)}{4(q-1)}}\} \\
&= (c_5 q)^4 (pm)^{\frac{2}{q-1}} (\frac{1}{2} \frac{d}{m})^{\frac{q}{2(q-1)}} \max\{(\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{1}{2(q-1)}}, \\
&\quad (c_5 q)^{4q-6} (pm)^{\frac{4q-6}{q-1}} (\frac{1}{2} \frac{d}{m})^{\frac{q(4q-6)}{4(q-1)}}\}
\end{aligned}$$

by convexity.

At this point, we will use Lemma 6.6 in [13] again with $\alpha = \frac{1}{2(q-1)}$. Therefore, we have $1 - \alpha = \frac{2q-3}{2(q-1)}$. Therefore, we have $\frac{\alpha}{1-\alpha} = \frac{1}{2q-3}$. So we need to use Lemma 6.6 in [13] with

$$K = ((c_5 q)^{4q-6} (pm)^{\frac{4q-6}{q-1}} (\frac{1}{2} \frac{d}{m})^{\frac{q(4q-6)}{4(q-1)}})^{\frac{1}{1-\alpha}}$$

in the notation there.

By Lemma 6.6 in [13], we have

$$\begin{aligned}
& (\mathbb{E}[\text{tr } \Gamma(S_1, S_2)^{2q}])^\alpha - (\mathbb{E}[\text{tr } \Gamma(G_1, G_2)^{2q}])^\alpha \\
& \leq (c_5 q)^4 (pm)^{\frac{2}{q-1}} \left(\frac{1}{2} \frac{d}{m}\right)^{\frac{q}{2(q-1)}} \frac{1}{2(q-1)} + ((c_5 q)^{4q-6} (pm)^{\frac{4q-6}{q-1}} \left(\frac{1}{2} \frac{d}{m}\right)^{\frac{q(4q-6)}{4(q-1)}})^{\frac{\alpha}{1-\alpha}} \\
& \leq c_{10} q^3 (pm)^{\frac{2}{q-1}} \left(\frac{d}{m}\right)^{\frac{q}{2(q-1)}} + ((c_5 q)^{4q-6} (pm)^{\frac{4q-6}{q-1}} \left(\frac{1}{2} \frac{d}{m}\right)^{\frac{q(4q-6)}{4(q-1)}})^{\frac{1}{2q-3}} \\
& \leq c_{10} q^3 (pm)^{\frac{2}{q-1}} \left(\frac{d}{m}\right)^{\frac{q}{2(q-1)}} + (c_5 q)^2 (pm)^{\frac{2}{q-1}} \left(\frac{1}{2} \frac{d}{m}\right)^{\frac{2q}{4(q-1)}} \\
& \leq c_{10} q^3 (pm)^{\frac{2}{q-1}} \left(\frac{d}{m}\right)^{\frac{q}{2(q-1)}} + (c_5 q)^2 (pm)^{\frac{2}{q-1}} \left(\frac{1}{2} \frac{d}{m}\right)^{\frac{q}{2(q-1)}} \\
& \leq c_{10} q^3 (pm)^{\frac{2}{q-1}} \left(\frac{d}{m}\right)^{\frac{q}{2(q-1)}} + c_{11} q^2 (pm)^{\frac{2}{q-1}} \left(\frac{d}{m}\right)^{\frac{q}{2(q-1)}} \\
& \leq c_{12} q^3 (pm)^{\frac{2}{q-1}} \left(\frac{d}{m}\right)^{\frac{q}{2(q-1)}}
\end{aligned}$$

In summary, we have

$$\begin{aligned}
& (\mathbb{E}[\text{tr } \Gamma(S_1, S_2)^{2q}])^{\frac{1}{2(q-1)}} - (\mathbb{E}[\text{tr } \Gamma(G_1, G_2)^{2q}])^{\frac{1}{2(q-1)}} \\
& \leq c_{12} q^3 (pm)^{\frac{2}{q-1}} \left(\frac{d}{m}\right)^{\frac{q}{2(q-1)}}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
& (\mathbb{E}[\text{tr } \Gamma(S_1, S_2)^{2q}])^{\frac{1}{2q}} - (\mathbb{E}[\text{tr } \Gamma(G_1, G_2)^{2q}])^{\frac{1}{2q}} \\
& \leq (c_{12} q^3 (pm)^{\frac{2}{q-1}} \left(\frac{d}{m}\right)^{\frac{q}{2(q-1)}})^{\frac{2(q-1)}{(2q)}} \\
& \leq c_{13} q^{\frac{3(q-1)}{q}} (pm)^{\frac{2}{q}} \left(\frac{d}{m}\right)^{\frac{1}{2}} \\
& \leq c_{13} q^3 (pm)^{\frac{2}{q}} \left(\frac{d}{m}\right)^{\frac{1}{2}}
\end{aligned}$$

where we use the fact that

$$(a+b)^{\frac{(2q)}{2(q-1)}} \geq a^{\frac{(2q)}{2(q-1)}} + b^{\frac{(2q)}{2(q-1)}}$$

by Taylor expansion.

We need $(\mathbb{E}[\text{tr } \Gamma(S_1, S_2)^{2q}])^{\frac{1}{2q}} - (\mathbb{E}[\text{tr } \Gamma(G_1, G_2)^{2q}])^{\frac{1}{2q}} \leq pm\varepsilon/4$, so the requirement is

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

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

$$c_{13} q^3 (pm)^{\frac{2}{q}} \left(\frac{d}{m}\right)^{\frac{1}{2}} \leq \frac{pm\sqrt{\frac{c_{14}d}{m}}}{4}$$

which means

$$pm \geq c_{15} (pm)^{\frac{2}{q}} q^3$$

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

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

for some constants $c_{16} > 0$ and $c_{17} > 0$.

This requirement is equivalent to

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

□

7.4. Differential inequality for the derivative of the interpolant. In this section we give the full proof of Lemma 7.5.

Lemma 7.5 (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. Gaussian entries each with variance p , and define the interpolated random matrices,*

$$(7.3) \quad \begin{aligned} S_1(t) &= \sqrt{t}S_1 + \sqrt{1-t}G_1 \\ S_2(t) &= \sqrt{t}S_2 + \sqrt{1-t}G_2 \end{aligned}$$

Let $f(M_1, M_2) = \text{tr}((M_1U)^T(M_2U) + (M_2U)^T(M_1U))^{2q}$. Then, there exists a constant $c_{7.5}$ such that, for any $q \geq 2$, we have

$$\frac{d}{dt} \mathbb{E}[f(S_1(t), S_2(t))] \leq \max_{4 \leq k \leq 2q} (c_{7.5}q)^k ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{qk-2q}{q-1}} (\mathbb{E}[f(S_1(t), S_2(t))])^{1-\frac{k-2}{2q-2}}$$

Remark 7.6. The proof of Lemma 7.5 relies on a technical trace inequality, Lemma 7.8. To illustrate the main idea, we present the proof of Lemma 7.5 using Lemma 7.8 here first, and then prove Lemma 7.8 in the next section.

Proof of Lemma 7.5. For convenience, let $\Gamma(M_1, M_2) = (M_1U)^T(M_2U) + (M_2U)^T(M_1U)$ and $\Gamma(t) = \Gamma(S_1(t), S_2(t))$.

Fix M_2 and view $f_{1,M_2}(M_1) = \text{tr}(\Gamma(M_1, M_2)^{2q})$ as a function of M_1 . We shall first obtain an expression for $\frac{d}{dt_1} \mathbb{E}_{S_1(t_1)}[f(S_1(t_1), M_2)]$. For this purpose, we consider the construction from [13]. Recall that S_1 can be written in the form $\sum_{(l,\gamma) \in \Xi} Z_{(l,\gamma)}$ where $\Xi = [n] \times [pm]$ and $Z_{(l,\gamma)} = \xi_{(l,\gamma)} e_{\mu_{(l,\gamma)}} e_l^\top$ (see definition 3.1). We shall assume that the $Z_{(l,\gamma)}$ are independent, with the additional remark that since all quantities involved are moments of order at most $4q$, the same calculations hold even if the $Z_{(l,\gamma)}$ are $4q$ -wise independent.

Let $k \in \mathbb{N}$ and $\pi \in \mathcal{P}([k])$ be a partition of $[k]$. We construct the following family of random elements

$$\{Z_{(l,\gamma),j|\pi}\}_{(l,\gamma) \in \Xi, j \in [k]}$$

that is independent from the sigma-algebra $\sigma(S_1, S_2, G_1, G_2)$ with the following properties:

1. Each $\{Z_{(l,\gamma),j|\pi}\}_{(l,\gamma) \in \Xi}$ has the same distribution as $\{Z_{(l,\gamma)}\}_{(l,\gamma) \in \Xi}$.
2. $\{Z_{(l,\gamma),j|\pi}\}_{(l,\gamma) \in \Xi} = \{Z_{(l,\gamma),j'|\pi}\}_{(l,\gamma) \in \Xi}$ for indices j, j' that belong to the same element of π .
3. $\{Z_{(l,\gamma),j|\pi}\}_{(l,\gamma) \in \Xi}$ are independent for indices j that belong to distinct elements of π .

More precisely, we construct these random elements in the following way. First, we choose a partition $\pi \in \mathcal{P}([k])$. Without loss of generality, consider the simple case where $\pi = \{A_1, A_2\}$. Then we construct two i.i.d. copies of $\{Z_{(l,\gamma)}\}_{(l,\gamma) \in \Xi}$, and call them $(Z_{(l,\gamma),A_1})_{(l,\gamma) \in \Xi}$ and $(Z_{(l,\gamma),A_2})_{(l,\gamma) \in \Xi}$. For any $j \in A_1$, we just set $\{Z_{(l,\gamma),j|\pi}\}_{(l,\gamma) \in \Xi} = \{Z_{(l,\gamma),A_1}\}_{(l,\gamma) \in \Xi}$. Similarly, for any $j \in A_2$, we just set $\{Z_{(l,\gamma),j|\pi}\}_{(l,\gamma) \in \Xi} = \{Z_{(l,\gamma),A_2}\}_{(l,\gamma) \in \Xi}$.

In addition, we assume that the random elements

$$\{Z_{(l,\gamma),j|\pi}\}_{(l,\gamma) \in \Xi, j \in [k]}$$

for different k and π are mutually independent from each other.

Using these random matrices, we can compute an expression for $\frac{d}{dt_1} \mathbb{E}_{S_1(t_1)}[f(S_1(t_1), M_2)]$ using the following lemma,

Lemma 7.7 (Corollary 6.1, [13]). *For any polynomial $\phi : M_{m \times d}(\mathbb{R}) \rightarrow \mathbb{R}$, we have*

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

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

We remark that Lemma 7.7 is proved in [13] for $S_1(t)$ in the space of $d \times d$ self-adjoint matrices with complex entries, but the result can be seen to hold for arbitrary matrices. Lemma 7.7 relies on the fact that the entries of G_1 have the same mean and covariance as the entries of S_1 for the vanishing of the terms corresponding to $k \leq 2$ in the above expansion. Terms corresponding to $k = 3$ vanish due to our random variables being symmetric, so we start the series with $k = 4$.

Proof of Lemma 7.7. Following the proof of Corollary 6.1 in [13], we have

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[\phi(X(t))] &= \frac{1}{2} \sum_{k=3}^{\infty} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \times \\ & \sum_{(l,\gamma) \in \Xi} \sum_{\substack{(u_j, v_j) \in [m] \times [d] \\ j=1, \dots, k}} \kappa((Z_{(l,\gamma)})_{u_1 v_1}, \dots, (Z_{(l,\gamma)})_{u_k v_k}) \mathbb{E} \left[\frac{\partial^k \phi}{\partial M_{u_1 v_1} \cdots \partial M_{u_k v_k}}(X(t)) \right], \end{aligned}$$

Also, we observe that $\kappa((Z_{(l,\gamma)})_{u_1 v_1}, \dots, (Z_{(l,\gamma)})_{u_k v_k})$ is always 0 when $k = 3$ because the entries of $Z_{(l,\gamma)}$ are symmetric. Therefore, we conclude that

$$\begin{aligned} \frac{d}{dt} \mathbb{E}[\phi(X(t))] &= \frac{1}{2} \sum_{k=4}^{\infty} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \times \\ & \sum_{(l,\gamma) \in \Xi} \sum_{\substack{(u_j, v_j) \in [m] \times [d] \\ j=1, \dots, k}} \kappa((Z_{(l,\gamma)})_{u_1 v_1}, \dots, (Z_{(l,\gamma)})_{u_k v_k}) \mathbb{E} \left[\frac{\partial^k \phi}{\partial M_{u_1 v_1} \cdots \partial M_{u_k v_k}}(X(t)) \right], \end{aligned}$$

Then the result follows by applying Lemma 4.1 in [13]. \square

Using Lemma 7.7 with $\phi = f_{1,M_2}$ and t_1 as the variable name, we have

$$\begin{aligned} & \frac{d}{dt_1} \mathbb{E}[f(S_1(t_1), M_2)] = \frac{d}{dt_1} \mathbb{E}[f_{1,M_2}(S_1(t_1))] \\ &= \frac{1}{2} \sum_{k=4}^{2q} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{\pi \in P([k])} (-1)^{|\pi|-1} (|\pi|-1)! \mathbb{E} \left[\sum_{(l,\gamma) \in [n] \times [pm]} \partial_{Z_{(l,\gamma),1|\pi}} \cdots \partial_{Z_{(l,\gamma),k|\pi}} f_{1,M_2}(S_1(t_1)) \right] \end{aligned}$$

For clarity, we introduce the following notations. Let X, Y be two random elements taking values in measurable spaces \mathcal{S}_1 and \mathcal{S}_2 . Let $\phi : \mathcal{S}_1 \times \mathcal{S}_2 \rightarrow \mathbb{R}$ be a measurable function. Assume that $\phi(X, Y)$ is integrable. Define the function $\psi_2 : \mathcal{S}_2 \rightarrow \mathbb{R}$ such that $\psi_2(y) = \mathbb{E} \phi(X, y)$. Then the notation $\mathbb{E}_X(\phi(X, Y))$ stands for $\psi_2(Y)$. Similarly, we define the function $\psi_1 : \mathcal{S}_1 \rightarrow \mathbb{R}$ such that $\psi_1(x) = \mathbb{E} \phi(x, Y)$ and the notation $\mathbb{E}_Y(\phi(X, Y))$ stands for the random variable $\psi_1(X)$. With

these notation and observing that we can plug the random variable $S_2(t_2)$ into M_2 in the previous identity (since the identity holds for arbitrary M_2), we have

$$\begin{aligned} & \frac{d}{dt_1} \mathbb{E}_{S_1(t_1)} [f(S_1(t_1), S_2(t_2))] \\ &= \frac{1}{2} \sum_{k=4}^{2q} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{\pi \in P([k])} (-1)^{|\pi|-1} (|\pi| - 1)! \\ & \quad \cdot \mathbb{E}_{S_1(t_1)} \left[\sum_{(l, \gamma) \in [n] \times [pm]} \partial_{Z_{(l, \gamma), 1|\pi}} \cdots \partial_{Z_{(l, \gamma), k|\pi}} f_{1, S_2(t_2)}(S_1(t_1)) \right] \end{aligned}$$

By differentiation under integral sign [40, Theorem 2.27] and iterated integration, we have

$$\begin{aligned} & \frac{d}{dt_1} \mathbb{E}_{S_2(t_2)} \mathbb{E}_{S_1(t_1)} [f(S_1(t_1), S_2(t_2))] \\ &= \mathbb{E}_{S_2(t_2)} \frac{d}{dt_1} \mathbb{E}_{S_1(t_1)} [f(S_1(t_1), S_2(t_2))] \\ &= \frac{1}{2} \sum_{k=4}^{2q} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{\pi \in P([k])} (-1)^{|\pi|-1} (|\pi| - 1)! \\ & \quad \mathbb{E}_{S_2(t_2)} \mathbb{E}_{S_1(t_1)} \left[\sum_{(l, \gamma) \in [n] \times [pm]} \partial_{Z_{(l, \gamma), 1|\pi}} \cdots \partial_{Z_{(l, \gamma), k|\pi}} f_{1, S_2(t_2)}(S_1(t_1)) \right], \end{aligned}$$

By the independence between $S_1(t_1)$ and $S_2(t_2)$, we can use iterated expectation and conclude that

$$\mathbb{E}_{S_2(t_2)} \mathbb{E}_{S_1(t_1)} [f(S_1(t_1), S_2(t_2))] = \mathbb{E}[f(S_1(t_1), S_2(t_2))]$$

and

$$\begin{aligned} & \mathbb{E}_{S_2(t_2)} \mathbb{E}_{S_1(t_1)} \left[\sum_{(l, \gamma) \in [n] \times [pm]} \partial_{Z_{(l, \gamma), 1|\pi}} \cdots \partial_{Z_{(l, \gamma), k|\pi}} f_{1, S_2(t_2)}(S_1(t_1)) \right] \\ &= \mathbb{E} \left[\sum_{(l, \gamma) \in [n] \times [pm]} \partial_{Z_{(l, \gamma), 1|\pi}} \cdots \partial_{Z_{(l, \gamma), k|\pi}} f_{1, S_2(t_2)}(S_1(t_1)) \right] \end{aligned}$$

Therefore, we have

$$\begin{aligned} & \frac{d}{dt_1} \mathbb{E}[f(S_1(t_1), S_2(t_2))] \\ &= \frac{1}{2} \sum_{k=4}^{2q} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{\pi \in P([k])} (-1)^{|\pi|-1} (|\pi| - 1)! \\ & \quad \mathbb{E} \left[\sum_{(l, \gamma) \in [n] \times [pm]} \partial_{Z_{(l, \gamma), 1|\pi}} \cdots \partial_{Z_{(l, \gamma), k|\pi}} f_{1, S_2(t_2)}(S_1(t_1)) \right], \end{aligned}$$

Similarly, we can define $f_{2,M_1}(M_2) = \text{tr}(\Gamma(M_1, M_2)^{2q})$ and obtain

$$\begin{aligned} & \frac{d}{dt_2} \mathbb{E}[f(S_1(t_1), S_2(t_2))] \\ &= \frac{1}{2} \sum_{k=4}^{2q} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{\pi \in P([k])} (-1)^{|\pi|-1} (|\pi| - 1)! \\ & \quad \mathbb{E} \left[\sum_{(l, \gamma) \in [n] \times [pm]} \partial_{Z_{(l, \gamma), 1|\pi}} \cdots \partial_{Z_{(l, \gamma), k|\pi}} f_{2, S_1(t_1)}(S_2(t_2)) \right], \end{aligned}$$

We decompose

$$\phi(t) = \frac{d}{dt} \mathbb{E}[f(S_1(t), S_2(t))]$$

as $\phi(t) = \psi_1 \circ \psi_2(t)$ where $\psi_1(t_1, t_2) = \mathbb{E}[f(S_1(t_1), S_2(t_2))]$ and $(t_1, t_2) = \psi_2(t) = (t, t)$.

By chain rule, we have

$$\begin{aligned} & \frac{d}{dt} \mathbb{E}[f(S_1(t), S_2(t))] \\ &= \begin{bmatrix} \frac{d}{dt_1} (\mathbb{E}[f(S_1(t), S_2(t))]) & \frac{d}{dt_2} (\mathbb{E}[f(S_1(t), S_2(t))]) \end{bmatrix} \cdot \begin{bmatrix} \frac{d}{dt_1}(t) \\ \frac{d}{dt_2}(t) \end{bmatrix} \\ &= \frac{1}{2} \sum_{k=4}^{2q} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{\pi \in P([k])} (-1)^{|\pi|-1} (|\pi| - 1)! \\ (7.4) \quad & \mathbb{E} \left[\sum_{(l, \gamma) \in [n] \times [pm]} \partial_{Z_{(l, \gamma), 1|\pi}} \cdots \partial_{Z_{(l, \gamma), k|\pi}} f_{1, S_2(t)}(S_1(t)) \right] \\ &+ \frac{1}{2} \sum_{k=4}^{2q} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{\pi \in P([k])} (-1)^{|\pi|-1} (|\pi| - 1)! \\ & \mathbb{E} \left[\sum_{(l, \gamma) \in [n] \times [pm]} \partial_{Z_{(l, \gamma), 1|\pi}} \cdots \partial_{Z_{(l, \gamma), k|\pi}} f_{2, S_1(t)}(S_2(t)) \right] \\ &=: T_1 + T_2 \end{aligned}$$

where T_1 and T_2 denote the first and second sum respectively.

The next step is to bound $\frac{d}{dt} \mathbb{E}[f(S_1(t), S_2(t))]$.

From Equation (7.4), we notice that the terms inside the expectation in both T_1 and T_2 are directional derivatives of $f_{1, S_2(t_2)}$ and $f_{2, S_1(t_1)}$ along $Z_{(l, \gamma), 1|\pi}, \dots, Z_{(l, \gamma), k|\pi}$. Using a general expression for derivatives of multinomials using product rule, we have, for any deterministic $m \times d$ matrices B_1, \dots, B_k, M_1 and M_2 ,

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

In our case, for each fixed (l, γ) , we have to analyse (in the case of T_1) $\partial_{Z_{(l, \gamma), 1|\pi}} \cdots \partial_{Z_{(l, \gamma), k|\pi}} f_{1, S_2(t_2)}(S_1(t_1))$, which means we have $B_\lambda = Z_{(l, \gamma), \lambda|\pi}$ for $\lambda \in [k]$, and $M_2 = S_2(t)$. So terms of the

form $(B_\lambda U)^T M_2 U$ become,

$$(Z_{(l,\gamma),\lambda|\pi} U)^T S_2(t) U = \xi_{(l,\gamma),\lambda|\pi} \mathbf{u}_l e_{\mu_{(l,\gamma),\lambda|\pi}}^T S_2(t) U$$

where \mathbf{u}_l is the column vector such that \mathbf{u}_l^T is the l th row of U and we use the fact that $Z_{(l,\gamma)} = \xi_{(l,\gamma)} e_{\mu_{(l,\gamma)}}^T$ with $\mu_{(l,\gamma)}$ being a uniformly distributed random variable on $[(m/s)(\gamma-1)+1 : (m/s)\gamma]$ and $\xi_{(l,\gamma)}$ being a random sign.

Let $\Theta_{(l,\gamma),\lambda,1} = \xi_{(l,\gamma)} \mathbf{u}_l^T$ and $\Theta_{(l,\gamma),\lambda,2} = e_{\mu_{(l,\gamma),\lambda|\pi}}(S_2(t) U)$. Then,

$$(Z_{(l,\gamma),\lambda|\pi} U)^T S_2(t) U = \xi_{(l,\gamma)} \mathbf{u}_l e_{\mu_{(l,\gamma),\lambda|\pi}}^T S_2(t) U = \Theta_{(l,\gamma),\lambda,1}^T \Theta_{(l,\gamma),\lambda,2}$$

Thus, terms of the form $(B_\lambda U)^T M_2 U + (M_2 U)^T B_\lambda U$ become

$$\begin{aligned} (Z_{(l,\gamma),\lambda|\pi} U)^T S_2(t) U + (S_2(t) U)^T Z_{(l,\gamma),\lambda|\pi} U &= \Theta_{(l,\gamma),\lambda,1}^T \Theta_{(l,\gamma),\lambda,2} + \Theta_{(l,\gamma),\lambda,2}^T \Theta_{(l,\gamma),\lambda,1} \\ &= \sum_{\tau \in \text{sym}(\{1,2\})} \Theta_{(l,\gamma),\lambda,\tau(1)}^T \Theta_{(l,\gamma),\lambda,\tau(2)} \end{aligned}$$

Doing this for all terms

$$(B_{\sigma(1)} U)^T M_2 U + (M_2 U)^T B_{\sigma(1)} U, \dots, (B_{\sigma(1)} U)^T M_2 U + (M_2 U)^T B_{\sigma(1)} U$$

in the expansion of $\partial_{Z_{(l,\gamma),1|\pi}} \cdots \partial_{Z_{(l,\gamma),k|\pi}} f_{1,S_2(t_2)}(S_1(t_1))$ (for a fixed (l, γ) and $\pi \in P([k])$), we get

$$\begin{aligned} &\partial_{Z_{(l,\gamma),1|\pi}} \cdots \partial_{Z_{(l,\gamma),k|\pi}} f_{1,S_2(t_2)}(S_1(t_1)) \\ &= \sum_{\sigma \in \text{sym}([k])} \sum_{\substack{r_1, \dots, r_{k+1} \geq 0 \\ r_1 + \dots + r_{k+1} = 2q - k}} \sum_{\tau_1, \dots, \tau_k \in \text{sym}(\{1,2\})} \text{tr } \Gamma(t)^{r_1} \Theta_{(l,\gamma),\sigma(1),\tau_1(1)}^T \Theta_{(l,\gamma),\sigma(1),\tau_1(2)} \\ &\quad \cdot \Gamma(t)^{r_2} \Theta_{(l,\gamma),\sigma(2),\tau_2(1)}^T \Theta_{(l,\gamma),\sigma(2),\tau_2(2)} \cdots \Gamma(t)^{r_k} \Theta_{(l,\gamma),\sigma(k),\tau_k(1)}^T \Theta_{(l,\gamma),\sigma(k),\tau_k(2)} \Gamma(t)^{r_{k+1}} \end{aligned}$$

Fixing k and π , summing over $(l, \gamma) \in [n] \times [pm]$, and taking expectations, we have

$$\begin{aligned} &\mathbb{E} \left[\sum_{(l,\gamma) \in [n] \times [pm]} \partial_{Z_{(l,\gamma),1|\pi}} \cdots \partial_{Z_{(l,\gamma),k|\pi}} f_{1,S_2(t_2)}(S_1(t_1)) \right] \\ &= \sum_{\sigma \in \text{sym}([k])} \sum_{\substack{r_1, \dots, r_{k+1} \geq 0 \\ r_1 + \dots + r_{k+1} = 2q - k}} \sum_{\tau_1, \dots, \tau_k \in \text{sym}(\{1,2\})} \sum_{(l,\gamma) \in [n] \times [pm]} \\ &\quad \mathbb{E} [\text{tr } \Gamma(t)^{r_1} \Theta_{(l,\gamma),\sigma(1),\tau_1(1)}^T \Theta_{(l,\gamma),\sigma(1),\tau_1(2)} \cdot \Gamma(t)^{r_2} \\ &\quad \cdots \Gamma(t)^{r_k} \Theta_{(l,\gamma),\sigma(k),\tau_k(1)}^T \Theta_{(l,\gamma),\sigma(k),\tau_k(2)} \Gamma(t)^{r_{k+1}}] \end{aligned}$$

By Lemma 7.8 (when S_1 and S_2 have the unscaled OSNAP distribution) and by Lemma 9.5 (when S_1 and S_2 have the unscaled OSE-IE distribution) with $\Upsilon_1 = \Gamma(t)^{r_2}, \dots, \Upsilon_k = \Gamma(t)^{r_k + r_1}$, we

have

$$\begin{aligned}
& \sum_{(l,\gamma) \in [n] \times [pm]} \mathbb{E}[\text{tr } \Gamma(t)^{r_1} \Theta_{(l,\gamma),\sigma(1),\tau_1(1)}^T \Theta_{(l,\gamma),\sigma(1),\tau_1(2)} \cdot \Gamma(t)^{r_2} \\
& \quad \cdots \Gamma(t)^{r_k} \Theta_{(l,\gamma),\sigma(k),\tau_k(1)}^T \Theta_{(l,\gamma),\sigma(k),\tau_k(2)} \Gamma(t)^{r_{k+1}}] \\
(7.5) \quad &= \sum_{(l,\gamma) \in [n] \times [pm]} \mathbb{E}[\text{tr } \Theta_{(l,\gamma),\sigma(1),\tau_1(1)}^T \Theta_{(l,\gamma),\sigma(1),\tau_1(2)} \cdot \Gamma(t)^{r_2} \\
& \quad \cdots \Gamma(t)^{r_k} \Theta_{(l,\gamma),\sigma(k),\tau_k(1)}^T \Theta_{(l,\gamma),\sigma(k),\tau_k(2)} \Gamma(t)^{r_{k+1}} \Gamma(t)^{r_1}] \\
&= \sum_{(l,\gamma) \in [n] \times [pm]} \mathbb{E}[\text{tr } \Theta_{(l,\gamma),1,\tau_1(1)}^T \Theta_{(l,\gamma),1,\tau_1(2)} \Upsilon_1 \Theta_{(l,\gamma),2,\tau_2(1)}^T \Theta_{(l,\gamma),2,\tau_2(2)} \\
& \quad \cdots \Upsilon_2 \Theta_{(l,\gamma),k,\tau_k(1)}^T \Theta_{(l,\gamma),k,\tau_k(2)} \Upsilon_k] \\
&\leq (c_{7.8}(pm))^{\frac{1}{q}} \sqrt{\max\{pd, q\}}^{\frac{qk-2q}{q-1}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{\frac{1}{q} \cdot \frac{2q-k}{2q-2}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{k}{2q}}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
& \mathbb{E} \left[\sum_{(l,\gamma) \in [n] \times [pm]} \partial_{Z_{(l,\gamma),1|\pi}} \cdots \partial_{Z_{(l,\gamma),k|\pi}} f_{1,S_2(t_2)}(S_1(t_1)) \right] \\
&\leq k! \binom{2q}{k} 2^k (c_{7.8}(pm))^{\frac{1}{q}} \sqrt{\max\{pd, q\}}^{\frac{qk-2q}{q-1}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{\frac{1}{q} \cdot \frac{2q-k}{2q-2}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{k}{2q}} \\
&\leq (4q)^k (c_{7.8}(pm))^{\frac{1}{q}} \sqrt{\max\{pd, q\}}^{\frac{qk-2q}{q-1}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{1}{2q}(k-2/r)}
\end{aligned}$$

where $r = \frac{2q-2}{2q-k}$.

Going back to the expression for T_1 and using [13, Lemma 6.4], we get,

$$\begin{aligned}
T_1 &= \frac{1}{2} \sum_{k=4}^{2q} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{\pi \in P([k])} (-1)^{|\pi|-1} (|\pi|-1)! \\
&\quad \cdot \mathbb{E} \left[\sum_{(l,\gamma) \in [n] \times [pm]} \partial_{Z_{(l,\gamma),1|\pi}} \cdots \partial_{Z_{(l,\gamma),k|\pi}} f_{1,S_2(t_2)}(S_1(t_1)) \right] \\
&\leq \frac{1}{2} \sum_{k=4}^{2q} \frac{1}{(k-1)!} 2^k (k-1)! \\
&\quad \cdot (4q)^k (c_{7.8}(pm))^{\frac{1}{q}} \sqrt{\max\{pd, q\}}^{\frac{qk-2q}{q-1}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{1}{2q}(k-2/r)} \\
&\leq \frac{1}{2} \sum_{k=4}^{2q} (8q)^k c_1^k ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{qk-2q}{q-1}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{1}{2q}(k-2/r)}
\end{aligned}$$

We similarly get $T_2 \leq \frac{1}{2} \sum_{k=4}^{2q} (8q)^k c_1^k ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{qk-2q}{q-1}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{1}{2q}(k-2/r)}$.

Going back, we have,

$$\begin{aligned}
\frac{d}{dt} \mathbb{E}[\text{tr} \Gamma(t)^{2q}] &\leq \sum_{k=4}^{2q} (8c_1 q)^k ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{qk-2q}{q-1}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{1}{2q}(k-2/r)} \\
&= \sum_{k=4}^{2q} 2^{-k} (16c_1 q)^k ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{qk-2q}{q-1}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{1}{2q}(k-2/r)} \\
&\leq \max_{4 \leq k \leq 2q} (16c_1 q)^k ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{qk-2q}{q-1}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{1-\frac{1}{2q}(k-2/r)} \\
&\leq \max_{4 \leq k \leq 2q} (c_5 q)^k ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{qk-2q}{q-1}} (\mathbb{E}[f(S_1(t), S_2(t))])^{1-\frac{k-2}{2q-2}}
\end{aligned}$$

□

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 (7.5) in the proof of Lemma 7.5.

Lemma 7.8 (Trace Inequalities for OSNAP). *Let $S_1(t)$ and $S_2(t)$ be as in Lemma 7.5 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 $q \geq 2$ 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

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

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 $\{\Upsilon_1, \dots, \Upsilon_k\}$ 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 $\{\Upsilon_1, \dots, \Upsilon_k\}$. (In other words, $\{\Upsilon_1, \dots, \Upsilon_k\}$ 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

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

where $e_{\mu_{(l,\gamma),\lambda}}$ represents the $\mu_{(l,\gamma),\lambda}$ th coordinate vector. Then, given $0 \leq \beta_1, \dots, \beta_k \leq +\infty$ such that

$$\sum_{\lambda=1}^k \frac{1}{\beta_\lambda} = 1 - \frac{k}{2q}, \tau_1, \dots, \tau_k \in \text{sym}(\{1, 2\}), \text{ there exists } c_{7.8} > 0 \text{ such that}$$

$$\begin{aligned}
(7.6) \quad &\sum_{(l,\gamma) \in [n] \times [pm]} \mathbb{E}[\text{tr} \Theta_{(l,\gamma),1,\tau_1(1)}^T \Theta_{(l,\gamma),1,\tau_1(2)} \Upsilon_1 \Theta_{(l,\gamma),2,\tau_2(1)}^T \Theta_{(l,\gamma),2,\tau_2(2)} \\
&\cdots \Upsilon_2 \Theta_{(l,\gamma),k,\tau_k(1)}^T \Theta_{(l,\gamma),k,\tau_k(2)} \Upsilon_k] \\
&\leq (c_{7.8} (pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{2qk-4q}{2q-2}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{\frac{1}{q} \cdot \frac{2q-k}{2q-2}} \prod_{\lambda=1}^k \|\Upsilon_\lambda\|_{\beta_\lambda}
\end{aligned}$$

Remark 7.9. 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

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

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 $(\mathbb{E}[\text{tr } \Gamma(1)^{2q}])^{\frac{1}{2q}} - (\mathbb{E}[\text{tr } \Gamma(0)^{2q}])^{\frac{1}{2q}}$ by just a constant times some power of $\log(d)$.

To bound $(\mathbb{E}[\text{tr } \Gamma(1)^{2q}])^{\frac{1}{2q}} - (\mathbb{E}[\text{tr } \Gamma(0)^{2q}])^{\frac{1}{2q}}$, 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

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

if we have a differential inequality of the form

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

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 (7.6) 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 7.8, 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 $(\mathbb{E}[\text{tr } \Gamma(t)^{2q}])^{\frac{1}{2q}}$, 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 (7.6) 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 $\sqrt{\max\{pd, q\}} \cdot \sqrt{\max\{pm, q\}}$ 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 $\sqrt{\max\{pd, q\}} \cdot \sqrt{\max\{pm, q\}}$, we develop a completely different method. We combine several factors at early stage and then use a key observation, Lemma 7.11, to replace this combined factor by $(\mathbb{E}[\text{tr } \Gamma(t)^{2q}])^{\frac{1}{2q}}$ directly. Eventually this method allowed us to get a more precise upper estimation, namely the right hand side of (7.6). 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 $\sqrt{pdp m}$ (see the proof of Proposition 7.3 for details).

Proof of Lemma 7.8. Let

$$\begin{aligned} & F(\Upsilon_1, \dots, \Upsilon_k) \\ &= \sum_{(l, \gamma) \in [n] \times [pm]} \mathbb{E}[\text{tr } \Theta_{(l, \gamma), 1, \tau_1(1)}^T \Theta_{(l, \gamma), 1, \tau_1(2)} \Upsilon_1 \Theta_{(l, \gamma), 2, \tau_2(1)}^T \Theta_{(l, \gamma), 2, \tau_2(2)} \\ & \quad \cdots \Upsilon_2 \Theta_{(l, \gamma), k, \tau_k(1)}^T \Theta_{(l, \gamma), k, \tau_k(2)} \Upsilon_k] \end{aligned}$$

Then, by Holder and the definition of $\Theta_{(l,\gamma),\lambda,1}, \Theta_{(l,\gamma),\lambda,2} \in \mathbb{R}^d$, we can show that $F(\Upsilon_1, \dots, \Upsilon_k)$ defines a multilinear functional on $L_\infty(S_\infty^d)$.

By Corollary 6.6, it suffices to prove the claim for $(\beta_1, \dots, \beta_k)$ such that $(\frac{1}{\beta_1}, \dots, \frac{1}{\beta_k})$ are extreme points of the set $\{(x_1, \dots, x_k) \in [0, 1]^k : \sum_{\lambda=1}^k x_\lambda = 1 - \frac{k}{2q}\}$, because all the other $(\frac{1}{\beta_1}, \dots, \frac{1}{\beta_k})$ are convex combinations of the extreme points and the result follows from interpolation. Therefore, we only need to prove the claim for the extreme case when one of the β_λ 's is $\frac{q}{q-k}$ and all the others are ∞ . By symmetry, we only need to consider the case $\beta_1 = \dots = \beta_{k-1} = \infty$ and $\beta_k = \frac{q}{q-k}$.

Let $\eta = (\eta(1), \eta(2))$ be a uniformly distributed random vector in $[n] \times [pm]$ such that η is independent with $\{\mathcal{Z}_\lambda\}_{\lambda \in [k]}$, S_1, S_2, G_1, G_2 , and $\{\Upsilon_1, \dots, \Upsilon_k\}$.

For all $\lambda \in [k]$, define random vectors $\Theta_{1,\lambda} = \Theta_{\eta,\lambda,1}$ and $\Theta_{2,\lambda} = \Theta_{\eta,\lambda,2}$. Then,

$$\begin{aligned} & \sum_{(l,\gamma) \in [n] \times [pm]} \mathbb{E}[\text{tr} \Theta_{(l,\gamma),1,\tau_1(1)}^T \Theta_{(l,\gamma),1,\tau_1(2)} \Upsilon_1 \Theta_{(l,\gamma),2,\tau_2(1)}^T \Theta_{(l,\gamma),2,\tau_2(2)} \\ & \quad \cdots \Upsilon_2 \Theta_{(l,\gamma),k,\tau_k(1)}^T \Theta_{(l,\gamma),k,\tau_k(2)} \Upsilon_k] \\ &= pmn \cdot \mathbb{E}[\text{tr} \Theta_{\tau_1(1),1}^T \Theta_{\tau_1(2),1} \Upsilon_1 \Theta_{\tau_2(1),2}^T \Theta_{\tau_2(2),2} \cdots \Upsilon_2 \Theta_{\tau_k(1),k}^T \Theta_{\tau_k(2),k} \Upsilon_k] \\ &= \mathcal{S} \cdot \mathbb{E}[\text{tr} \Theta_{\tau_1(1),1}^T \Theta_{\tau_1(2),1} \Upsilon_1 \Theta_{\tau_2(1),2}^T \Theta_{\tau_2(2),2} \cdots \Upsilon_2 \Theta_{\tau_k(1),k}^T \Theta_{\tau_k(2),k} \Upsilon_k] \end{aligned}$$

where $\mathcal{S} := pmn$ is the number of non-zero entries in S . Essentially, we have written the sum over indices (l, γ) as an expectation over uniformly chosen (l, γ) times the number of tuples (l, γ) .

Let $\Upsilon_k = V_k |\Upsilon_k|$ be the polar decomposition. By matrix Holder inequality (Lemma 5.3. in [13]), we have

$$\begin{aligned} & |\mathbb{E}[\text{tr} \Theta_{\tau_1(1),1}^T \Theta_{\tau_1(2),1} \Upsilon_1 \Theta_{\tau_2(1),2}^T \Theta_{\tau_2(2),2} \cdots \Upsilon_{k-1} \Theta_{\tau_k(1),k}^T \Theta_{\tau_k(2),k} \Upsilon_k]| \\ &= |\mathbb{E}[\text{tr} |\Upsilon_k|^{1/2} \Theta_{\tau_1(1),1}^T \Theta_{\tau_1(2),1} \Upsilon_1 \Theta_{\tau_2(1),2}^T \Theta_{\tau_2(2),2} \cdots \Upsilon_{k-1} \Theta_{\tau_k(1),k}^T \Theta_{\tau_k(2),k} V_k |\Upsilon_k|^{1/2}]| \\ &\leq |\mathbb{E}[\text{tr} \Theta_{\tau_1(1),1}^T \Theta_{\tau_1(2),1} \Upsilon_1 \cdots \Theta_{\tau_{k/2}(1),k/2}^T \Theta_{\tau_{k/2}(2),k/2} \Upsilon_{k/2} \\ & \quad \cdot \Upsilon_{k/2}^T \Theta_{\tau_{k/2}(2),k/2} \Theta_{\tau_{k/2}(1),k/2} \cdots \Upsilon_1^T \Theta_{\tau_1(2),1} \Theta_{\tau_1(1),1} |\Upsilon_k|]|^{1/2} \\ & \quad \cdot |\mathbb{E}[\text{tr} \Theta_{\tau_k(2),k}^T \Theta_{\tau_k(1),k} \Upsilon_{k-1} \cdots \Upsilon_{k/2+1}^T \Theta_{\tau_{k/2+1}(2),k/2+1} \Theta_{\tau_{k/2+1}(1),k/2+1} \\ & \quad \cdot \Theta_{\tau_{k/2+1}(1),k/2+1}^T \Theta_{\tau_{k/2+1}(2),k/2+1} \Upsilon_{k/2+1} \cdots \Upsilon_{k-1} \Theta_{\tau_k(1),k}^T \Theta_{\tau_k(2),k} V_k |\Upsilon_k| V_k^T]|^{1/2} \\ &= (\text{factor 1}) \cdot (\text{factor 2}) \end{aligned}$$

where, assuming k is even, in the notation of Proposition 5.1 in [13], we set

$$Y_1 = |\Upsilon_k|^{1/2} \Theta_{\tau_1(1),1}^T \Theta_{\tau_1(2),1} \Upsilon_1 \cdots \Theta_{\tau_{k/2}(1),k/2}^T \Theta_{\tau_{k/2}(2),k/2} \Upsilon_{k/2}$$

and

$$Y_2 = \Theta_{\tau_{k/2+1}(1),k/2+1}^T \Theta_{\tau_{k/2+1}(2),k/2+1} \Upsilon_{k/2+1} \cdots \Upsilon_{k-1} \Theta_{\tau_k(1),k}^T \Theta_{\tau_k(2),k} V_k |\Upsilon_k|^{1/2}$$

with the exponents being $p_1 = p_2 = 2$. When k is odd, the argument can be modified as suggested in Step 4 of the proof of Proposition 5.1 in [13].

We focus on the factor

$$\begin{aligned} (\text{factor 1}) &= |\mathbb{E}[\text{tr} \Theta_{\tau_1(1),1}^T \Theta_{\tau_1(2),1} \Upsilon_1 \cdots \Theta_{\tau_{k/2}(1),k/2}^T \Theta_{\tau_{k/2}(2),k/2} \Upsilon_{k/2} \\ & \quad \cdot \Upsilon_{k/2}^T \Theta_{\tau_{k/2}(2),k/2} \Theta_{\tau_{k/2}(1),k/2} \cdots \Upsilon_1^T \Theta_{\tau_1(2),1} \Theta_{\tau_1(1),1} |\Upsilon_k|]|^{1/2} \end{aligned}$$

and the analysis of the other factor will be similar.

Let

$$\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1} = \Lambda_1 |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|$$

be the polar decomposition.

Therefore, we have, for some $r \geq 1$ to be fixed later,

$$\begin{aligned}
& (\text{factor } 1)^2 \\
& = |\mathbb{E}[\text{tr} \Theta_{\tau_1(1),1}^T \Theta_{\tau_1(2),1} \Upsilon_1 \cdots \Theta_{\tau_{k/2}(1),k/2}^T \Theta_{\tau_{k/2}(2),k/2} \Upsilon_{k/2} \\
& \quad \cdot \Upsilon_{k/2}^T \Theta_{\tau_{k/2}(2),k/2}^T \Theta_{\tau_{k/2}(1),k/2} \cdots \Upsilon_1^T \Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1} |\Upsilon_k|]| \\
& = |\mathbb{E}[\text{tr} |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1-1/r} \Lambda_1^T \Upsilon_1 \cdots \Theta_{\tau_{k/2}(1),k/2}^T \Theta_{\tau_{k/2}(2),k/2} \Upsilon_{k/2} \\
& \quad \cdot \Upsilon_{k/2}^T \Theta_{\tau_{k/2}(2),k/2}^T \Theta_{\tau_{k/2}(1),k/2} \cdots \Upsilon_1^T \Lambda_1 |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1-1/r} \\
& \quad \cdot |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r} |\Upsilon_k| |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r}]| \\
& \leq \|\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}\|_{2q}^{2-2/r} \prod_{\lambda=2}^{k/2} \|\Theta_{\tau_\lambda(1),\lambda}^T \Theta_{\tau_\lambda(2),\lambda}\|_{2q}^2 \prod_{\lambda=1}^{k/2} \|\Upsilon_\lambda\|_\infty^2 \\
& \quad \cdot \| |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r} |\Upsilon_k| |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r} \|_r
\end{aligned} \tag{7.7}$$

where the last assertion follows from using the Hölder inequality with the exponents,

- $2q$ for $\Theta_{\tau_j(1),j}^T \Theta_{\tau_j(2),j}$ and $\Theta_{\tau_j(2),j}^T \Theta_{\tau_j(1),j}$ for $j = 2, \dots, k/2$. Note that there are $2(\frac{k}{2} - 1) = k - 2$ many such terms.
- ∞ for Υ_j or Υ_j^T for $j = 1, \dots, k/2$, Λ_1 and Λ_1^T .
- $\frac{2qr}{r-1}$ for each term $|\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1-1/r}$.
- r for the term $|\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r} |\Upsilon_k| |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r}$.

Then, the condition on exponents for Hölder inequality determines the value of r ,

$$\frac{k-2}{2q} + \frac{2(r-1)}{2rq} + \frac{1}{r} = 1$$

giving, $r = \frac{2q-2}{2q-k}$.

Starting from here, we bound all the factors that appear in the final line of (7.7). First, the factor $\|\Theta_{\tau_\lambda(1),\lambda}^T \Theta_{\tau_\lambda(2),\lambda}\|_{2q} = \|\Theta_{1,\lambda}^T \Theta_{2,\lambda}\|_{2q}$ can be bounded directly by using the following lemma.

Lemma 7.10 (Product of Random Rows). *Let $\Theta_{(l,\gamma),\lambda,1}, \Theta_{(l,\gamma),\lambda,2} \in \mathbb{R}^d$ be as in Lemma 7.8. Let $\eta = (\eta(1), \eta(2))$ be a uniformly distributed random vector in $[n] \times [pm]$ such that η is independent with $\{\mathcal{Z}_\lambda\}_{\lambda \in [k]}$, S_1, S_2, G_1, G_2 . For all $\lambda \in [k]$, define random vectors $\Theta_{1,\lambda} = \Theta_{\eta,\lambda,1}$ and $\Theta_{2,\lambda} = \Theta_{\eta,\lambda,2}$. Then there exists a constant $c_{7.10}$ such that, for any integer $q \geq 2$, we have*

$$\|\Theta_{1,\lambda}^T \Theta_{2,\lambda}\|_q \leq \frac{c_{7.10}(pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}}}{\mathcal{S}^{\frac{1}{q}}}$$

where $\mathcal{S} = pmn$.

Proof. Note that for each fixed $1 \leq \lambda \leq k/2$,

$$\begin{aligned} \|\Theta_{1,\lambda}^T \Theta_{2,\lambda}\|_q &= \|\xi_{\eta,\lambda} \mathbf{u}_{\eta(1)} e_{\mu_{\eta,\lambda}}^T S_2(t) U\|_q \\ &\leq \left(\mathbb{E}[\text{tr}|\xi_{\eta,\lambda} \mathbf{u}_{\eta(1)} e_{\mu_{\eta,\lambda}}^T S_2(t) U|^q] \right)^{\frac{1}{q}} \\ &= \left(\mathbb{E}[\text{tr}|\mathbf{u}_{\eta(1)} e_{\mu_{\eta,\lambda}}^T S_2(t) U|^q] \right)^{\frac{1}{q}} \\ &= \left(\mathbb{E} \left[\frac{1}{d} \|\mathbf{u}_{\eta(1)}\|^q \|e_{\mu_{\eta,\lambda}}^T S_2(t) U\|^q \right] \right)^{\frac{1}{q}} \end{aligned}$$

Conditioning on η , we take expectation of the second factor over $\mu_{\eta,\lambda}$ and S_2 ,

$$\|\Theta_{1,\lambda}^T \Theta_{2,\lambda}\|_q \leq \left(\mathbb{E}_\eta \left[\frac{1}{d} \|\mathbf{u}_{\eta(1)}\|^q \mathbb{E}_{\mu_{\eta,\lambda}, S_2(t)} [\|e_{\mu_{\eta,\lambda}}^T S_2(t) U\|^q] \right] \right)^{\frac{1}{q}}$$

Note that, conditioned on η , $\mu_{\eta,\lambda}$ is uniformly distributed over a subset of $[m]$. Therefore, by Lemma 6.2, we have

$$\begin{aligned} \|\Theta_{1,\lambda}^T \Theta_{2,\lambda}\|_q &\leq \left(\mathbb{E}_\eta \left[\frac{1}{d} \|\mathbf{u}_{\eta(1)}\|^q \mathbb{E}_{\mu_{\eta,\lambda}, S_2(t)} [\|e_{\mu_{\eta,\lambda}}^T S_2(t) U\|^q] \right] \right)^{\frac{1}{q}} \\ &\leq c_{6.2} \sqrt{\max\{pd, q\}} \left(\mathbb{E}_\eta \left[\frac{1}{d} \|\mathbf{u}_{\eta(1)}\|^q \right] \right)^{\frac{1}{q}} \\ &= c_{6.2} \sqrt{\max\{pd, q\}} \left(\frac{1}{\mathcal{S}d} \sum_{(l,\gamma) \in [n] \times [pm]} \|u_l\|^q \right)^{\frac{1}{q}} \\ &= c_{6.2} \sqrt{\max\{pd, q\}} \left(\frac{pm}{\mathcal{S}d} \sum_{l=1}^n \|u_l\|^q \right)^{\frac{1}{q}} \\ &\leq c_{6.2} \sqrt{\max\{pd, q\}} \left(\frac{pm}{\mathcal{S}d} \sum_{l=1}^n \|u_l\|^2 \right)^{\frac{1}{q}} \\ &\leq \frac{c_{7.10} (pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}}}{\mathcal{S}^{\frac{1}{q}}} \end{aligned}$$

where in the last line we use the fact that $\sum_{l=1}^n \|u_l\|^2 = d$.

□

Next, we have the following bound for the factor $\| |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r} |\Upsilon_k| |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r} \|_r$.

$$\begin{aligned} &\| |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r} |\Upsilon_k| |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r} \|_r^r \\ &= \mathbb{E} \text{tr} ((|\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r} |\Upsilon_k| |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r})^r) \\ &\leq \mathbb{E} \text{tr} (|\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}| |\Upsilon_k|^r |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|) \\ &= \mathbb{E} \text{tr} (|\Upsilon_k|^r |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}| |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|) \\ &= \mathbb{E} \text{tr} (|\Upsilon_k|^r (\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1})^T \Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}) \end{aligned}$$

where the inequality is due to the Lieb Thirring Inequality ([13, Lemma 5.4]).

Here, we have two possible cases. The first case (CASE I) is $\tau_1(1) = 1$. In this case, we have

$$\begin{aligned} & (\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1})^T (\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}) \\ &= \Theta_{1,1}^T \Theta_{2,1} (\Theta_{2,1}^T \Theta_{1,1}) \\ &= (Z_{\eta,1} U)^T (S_2(t) U) (S_2(t) U)^T Z_{\eta,1} U \end{aligned}$$

Therefore, we have

$$\begin{aligned} & \mathbb{E} \operatorname{tr}(|\Upsilon_k|^r (\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1})^T \Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}) \\ &= \mathbb{E} \operatorname{tr}(|\Upsilon_k|^r (Z_{\eta,1} U)^T (S_2(t) U) (S_2(t) U)^T Z_{\eta,1} U) \\ &= \mathbb{E} \operatorname{tr}(|\Upsilon_k|^r \mathbb{E}_\eta((Z_{\eta,1} U)^T (S_2(t) U) (S_2(t) U)^T Z_{\eta,1} U)) \\ &= \mathbb{E} \operatorname{tr}(|\Upsilon_k|^r \frac{1}{S} (\sum_{(l,\gamma)} (Z_{(l,\gamma),1} U)^T (S_2(t) U) (S_2(t) U)^T Z_{(l,\gamma),1} U)) \\ &= \frac{1}{S} \mathbb{E} \operatorname{tr}(|\Upsilon_k|^r \mathbb{E}_Z (\sum_{(l,\gamma)} (Z_{(l,\gamma),1} U)^T (S_2(t) U) (S_2(t) U)^T Z_{(l,\gamma),1} U)) \\ &\leq \frac{1}{S} \| |\Upsilon_k|^r \|_{\frac{2q}{2q-2}} \cdot \|\mathbb{E}_Z (\sum_{(l,\gamma)} (Z_{(l,\gamma),1} U)^T (S_2(t) U) (S_2(t) U)^T Z_{(l,\gamma),1} U)\|_q \end{aligned}$$

Plugging this estimate and the previous estimate on $\|\Theta_{\tau_j(1),j}^T \Theta_{\tau_j(2),j}\|_q$ into inequality 7.7, we have

$$\begin{aligned} & (\text{factor } 1)^2 \\ & \leq \|\Theta_{2,1}^T \Theta_{1,1}\|_{2q}^{2-2/r} \prod_{\lambda=2}^{k/2} \|\Theta_{1,\lambda}^T \Theta_{2,\lambda}\|_{2q}^2 \prod_{\lambda=1}^{k/2} \|\Upsilon_\lambda\|_\infty^2 \|\Theta_{2,1}^T \Theta_{1,1}\|^{1/r} |\Upsilon_k| \|\Theta_{2,1}^T \Theta_{1,1}\|^{1/r} \|r \\ & \leq \left(\frac{c_{7.10}(pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}}}{S^{\frac{1}{q}}} \right)^{2-2/r} \left(\frac{c_{7.10}(pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}}}{S^{\frac{1}{q}}} \right)^{k-2} \prod_{\lambda=1}^{k/2} \|\Upsilon_\lambda\|_\infty^2 \\ (7.8) \quad & \cdot \left(\frac{1}{S} \| |\Upsilon_k|^r \|_{\frac{2q}{2q-2}} \cdot \left\| \mathbb{E}_Z \left(\sum_{(l,\gamma)} (Z_{(l,\gamma),1} U)^T (S_2(t) U) (S_2(t) U)^T Z_{(l,\gamma),1} U \right) \right\|_q \right)^{1/r} \\ & = \left(\frac{c_{7.10}(pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}}}{S^{\frac{1}{q}}} \right)^{\frac{2qk-4q}{2q-2}} \prod_{\lambda=1}^{k/2} \|\Upsilon_\lambda\|_\infty^2 \\ & \cdot \left(\frac{1}{S} \| |\Upsilon_k|^r \|_{\frac{2q}{2q-2}} \cdot \left\| \mathbb{E}_Z \left(\sum_{(l,\gamma)} (Z_{(l,\gamma),1} U)^T (S_2(t) U) (S_2(t) U)^T Z_{(l,\gamma),1} U \right) \right\|_q \right)^{1/r} \end{aligned}$$

To bound the term $\|\mathbb{E}_Z (\sum_{(l,\gamma)} (Z_{(l,\gamma),1} U)^T (S_2(t) U) (S_2(t) U)^T Z_{(l,\gamma),1} U)\|_q$, we can use the following lemma to relate it to $\Gamma(t)$.

Lemma 7.11 (Replacement by the Original Matrix). *Let $S_1(t)$, $S_2(t)$, $\Gamma(t)$, and $Z_{(l,\gamma),\lambda}$ be as in Lemma 7.8.*

We have

$$\begin{aligned} & \|\mathbb{E}_{Z_{(l,\gamma),1}} (\sum_{(l,\gamma)} (Z_{(l,\gamma),1} U)^T (S_2(t) U) (S_2(t) U)^T Z_{(l,\gamma),1} U)\|_q^q \\ & \leq \mathbb{E} \operatorname{tr}((\Gamma(t))^{2q}) \end{aligned}$$

and

$$\begin{aligned} & \|\mathbb{E}_{Z_{(l,\gamma),1}} \left(\sum_{(l,\gamma)} (S_2(t)U)^T (Z_{(l,\gamma),1}U) (Z_{(l,\gamma),1}U)^T (S_2(t)U) \right)\|_q^q \\ & \leq \mathbb{E} \operatorname{tr}((\Gamma(t))^{2q}) \end{aligned}$$

Proof. Let $W_{(l,\gamma),1}$ be the gaussian model for $Z_{(l,\gamma),1}$, i.e., the entries of $W_{(l,\gamma),1}$ are gaussian and the covariance structure between entries of $W_{(l,\gamma),1}$ is the same as $Z_{(l,\gamma),1}$, such that the family $\{W_{(l,\gamma),1} : (l,\gamma) \in \Xi\} \cup \{Z_{(l,\gamma),1} : (l,\gamma) \in \Xi\}$ are mutually independent. Let $Z_{(l,\gamma),1}(t) = \sqrt{t}Z_{(l,\gamma),1} + \sqrt{1-t}W_{(l,\gamma),1}$. With this notation, we know that $\sum_{(l,\gamma) \in \Xi} Z_{(l,\gamma),1}(t)$ has the same distribution with

$S_1(t)$.

We observe that

$$\begin{aligned} & \mathbb{E}_{Z_{(l,\gamma),1}} \left(\sum_{(l,\gamma)} (Z_{(l,\gamma),1}U)^T (S_2(t)U) (S_2(t)U)^T Z_{(l,\gamma),1}U \right) \\ & = \mathbb{E}_{Z_{(l,\gamma),1}(t)} \left(\sum_{(l,\gamma)} (Z_{(l,\gamma),1}(t)U)^T (S_2(t)U) (S_2(t)U)^T Z_{(l,\gamma),1}(t)U \right) \end{aligned}$$

because for fixed $(S_2(t)U)(S_2(t)U)^T$, the value

$$\mathbb{E}_{Z_{(l,\gamma),1}} \left(\sum_{(l,\gamma)} (Z_{(l,\gamma),1}U)^T (S_2(t)U) (S_2(t)U)^T Z_{(l,\gamma),1}U \right)$$

only depends on the covariance structure of $Z_{(l,\gamma),1}$.

Also, we have

$$\begin{aligned} & \mathbb{E}_{Z_{(l,\gamma),1}(t)} \left(\sum_{(l,\gamma)} (Z_{(l,\gamma),1}(t)U)^T (S_2(t)U) (S_2(t)U)^T Z_{(l,\gamma),1}(t)U \right) \\ & = \mathbb{E}_{Z_{(l,\gamma),1}(t)} \left(\left(\sum_{(l,\gamma)} Z_{(l,\gamma),1}(t)U \right)^T (S_2(t)U) (S_2(t)U)^T \left(\sum_{(l,\gamma)} Z_{(l,\gamma),1}(t)U \right) \right) \\ & = \mathbb{E}_{S_1(t)} \left((S_1(t)U)^T (S_2(t)U) (S_2(t)U)^T S_1(t)U \right) \end{aligned}$$

because the cross terms have zero expectation by independence.

Therefore, we have

$$\begin{aligned} & \mathbb{E}_{Z_{(l,\gamma),1}} \left(\sum_{(l,\gamma)} (Z_{(l,\gamma),1}U)^T (S_2(t)U) (S_2(t)U)^T Z_{(l,\gamma),1}U \right) \\ & = \mathbb{E}_{S_1(t)} \left((S_1(t)U)^T (S_2(t)U) (S_2(t)U)^T S_1(t)U \right) \end{aligned}$$

Similarly, we also have

$$\begin{aligned} & \mathbb{E}_{Z_{(l,\gamma),1}} \left(\sum_{(l,\gamma)} (S_2(t)U)^T (Z_{(l,\gamma),1}U) (Z_{(l,\gamma),1}U)^T (S_2(t)U) \right) \\ & = \mathbb{E}_{S_1(t)} \left((S_2(t)U)^T (S_1(t)U) (S_1(t)U)^T S_2(t)U \right) \end{aligned}$$

Now we look at $\mathbb{E}_{S_1(t)}(\Gamma(t)^2)$. We have

$$\begin{aligned} & \mathbb{E}_{S_1(t)}(\Gamma(t)^2) \\ &= \mathbb{E}_{S_1(t)}((S_1(t)U)^T(S_2(t)U)(S_2(t)U)^T S_1(t)U) \\ & \quad + \mathbb{E}_{S_1(t)}((S_2(t)U)^T(S_1(t)U)(S_1(t)U)^T S_2(t)U) \\ & \quad + \mathbb{E}_{S_1(t)}((S_1(t)U)^T(S_2(t)U)(S_1(t)U)^T S_2(t)U) \\ & \quad + \mathbb{E}_{S_1(t)}((S_2(t)U)^T(S_1(t)U)(S_2(t)U)^T S_1(t)U) \end{aligned}$$

To simplify the last two terms, we need the following observation. Consider a random matrix X whose entries have variance 1 and zero covariances. Let Y be a deterministic matrix. Then direct calculation can show that $\mathbb{E}_X XYX = Y^T$.

In fact, we have

$$(XY)_{i,\beta} = \sum_{\alpha} X_{i,\alpha} Y_{\alpha,\beta}$$

and therefore

$$\mathbb{E}_X(XYX)_{i,j} = \mathbb{E}_X\left(\sum_{\beta} \sum_{\alpha} X_{i,\alpha} Y_{\alpha,\beta} X_{\beta,j}\right) = \sum_{\beta} \sum_{\alpha} \mathbb{E}_X(X_{i,\alpha} Y_{\alpha,\beta} X_{\beta,j})$$

We observe that $\mathbb{E}_X(X_{i,\alpha} Y_{\alpha,\beta} X_{\beta,j}) = Y_{j,i}$ if $\alpha = j$ and $\beta = i$. When $(\alpha, \beta) \neq (j, i)$, we always have $\mathbb{E}_X(X_{i,\alpha} Y_{\alpha,\beta} X_{\beta,j}) = 0$. Therefore, we have $\mathbb{E}_X(XYX)_{i,j} = Y_{j,i}$, and then we conclude $\mathbb{E}_X(XYX) = Y^T$.

Observing that the matrix $(S_1(t)U)^T$ has entries with variance 1 and zero covariances, the above analysis gives that

$$\begin{aligned} & \mathbb{E}_{S_1(t)}((S_1(t)U)^T(S_2(t)U)(S_1(t)U)^T S_2(t)U) \\ &= (\mathbb{E}_{S_1(t)}((S_1(t)U)^T(S_2(t)U)(S_1(t)U)^T))(S_2(t)U) \\ &= (S_2(t)U)^T(S_2(t)U) \end{aligned}$$

and

$$\begin{aligned} & \mathbb{E}_{S_1(t)}((S_2(t)U)^T(S_1(t)U)(S_2(t)U)^T S_1(t)U) \\ &= (S_2(t)U)^T(\mathbb{E}_{S_1(t)}((S_1(t)U)(S_2(t)U)^T S_1(t)U)) \\ &= (S_2(t)U)^T(S_2(t)U) \end{aligned}$$

Therefore, we have

$$\begin{aligned} & \mathbb{E}_{S_1(t)}(\Gamma(t)^2) \\ &= \mathbb{E}_{S_1(t)}((S_1(t)U)^T(S_2(t)U)(S_2(t)U)^T S_1(t)U) \\ & \quad + \mathbb{E}_{S_1(t)}((S_2(t)U)^T(S_1(t)U)(S_1(t)U)^T S_2(t)U) \\ & \quad + \mathbb{E}_{S_1(t)}((S_1(t)U)^T(S_2(t)U)(S_1(t)U)^T S_2(t)U) \\ & \quad + \mathbb{E}_{S_1(t)}((S_2(t)U)^T(S_1(t)U)(S_2(t)U)^T S_1(t)U) \\ &= \mathbb{E}_{S_1(t)}((S_1(t)U)^T(S_2(t)U)(S_2(t)U)^T S_1(t)U) \\ & \quad + \mathbb{E}_{S_1(t)}((S_2(t)U)^T(S_1(t)U)(S_1(t)U)^T S_2(t)U) \\ & \quad + 2(S_2(t)U)^T(S_2(t)U) \end{aligned}$$

Since all the matrices in the sum

$$\begin{aligned} & \mathbb{E}_{S_1(t)}((S_1(t)U)^T(S_2(t)U)(S_2(t)U)^T S_1(t)U) \\ & + \mathbb{E}_{S_1(t)}((S_2(t)U)^T(S_1(t)U)(S_1(t)U)^T S_2(t)U) \\ & + 2(S_2(t)U)^T(S_2(t)U) \end{aligned}$$

are positively semidefinite, we can conclude that

$$\mathbb{E}_{S_1(t)}((S_1(t)U)^T(S_2(t)U)(S_2(t)U)^T S_1(t)U) \leq \mathbb{E}_{S_1(t)}(\Gamma(t)^2)$$

and

$$\mathbb{E}_{S_1(t)}((S_2(t)U)^T(S_1(t)U)(S_1(t)U)^T S_2(t)U) \leq \mathbb{E}_{S_1(t)}(\Gamma(t)^2)$$

Therefore, by [41, Theorem 2.10], we have

$$\begin{aligned} & \|\mathbb{E}_{Z_{(l,\gamma),1}}(\sum_{(l,\gamma)} (Z_{(l,\gamma),1}U)^T(S_2(t)U)(S_2(t)U)^T Z_{(l,\gamma),1}U)\|_q^q \\ & = \|\mathbb{E}_{S_1(t)}((S_1(t)U)^T(S_2(t)U)(S_2(t)U)^T S_1(t)U)\|_q^q \\ & = \mathbb{E} \operatorname{tr}((\mathbb{E}_{S_1(t)}((S_1(t)U)^T(S_2(t)U)(S_2(t)U)^T S_1(t)U))^q) \\ & \leq \mathbb{E} \operatorname{tr}((\mathbb{E}_{S_1(t)}(\Gamma(t)^2))^q) \\ & \leq \mathbb{E} \operatorname{tr}((\Gamma(t)^2)^q) \\ & \leq \mathbb{E} \operatorname{tr}((\Gamma(t))^{2q}) \end{aligned}$$

Similarly, we also have

$$\begin{aligned} & \|\mathbb{E}_{Z_{(l,\gamma),1}}(\sum_{(l,\gamma)} (S_2(t)U)^T(Z_{(l,\gamma),1}U)Z_{(l,\gamma),1}U)^T(S_2(t)U)\|_q^q \\ & \leq \mathbb{E} \operatorname{tr}((\Gamma(t))^{2q}) \end{aligned}$$

□

Therefore, we have bounded all the factors in the last line of (7.7), and the calculations give

$$\begin{aligned} & (\text{factor } 1)^2 \\ & \leq \left(\frac{c_{7.10}(pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}}}{\mathcal{S}^{\frac{1}{q}}} \right)^{\frac{2qk-4q}{2q-2}} \prod_{\lambda=1}^{k/2} \|\Upsilon_\lambda\|_\infty^2 \\ & \quad \cdot \left(\frac{1}{\mathcal{S}} \|\Upsilon_k\|^r \right)^{\frac{2q}{2q-2}} \cdot \|\mathbb{E}_{Z_{(l,\gamma),1}(t)}(\sum_{(l,\gamma)} (Z_{(l,\gamma),1}(t)U)^T(S_2(t)U)(S_2(t)U)^T Z_{(l,\gamma),1}(t)U)\|_q^{1/r} \\ & = \mathcal{S}^{-1} (c_{7.10}(pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{2qk-4q}{2q-2}} \prod_{\lambda=1}^{k/2} \|\Upsilon_\lambda\|_\infty^2 \cdot \|\Upsilon_k\|_{\frac{2q}{2q-k}} (\mathbb{E} \operatorname{tr}((\Gamma(t))^{2q}))^{(1/q)(1/r)} \end{aligned}$$

The second case (CASE II) is $\tau_1(1) = 2$, or $\tau_1(2) = 1$. In this case, we have

$$\begin{aligned} & (\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1})^T (\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}) \\ & = (\Theta_{1,1}^T \Theta_{2,1})^T (\Theta_{1,1}^T \Theta_{2,1}) \\ & = (S_2(t)U)^T Z_{\eta,1} U (Z_{\eta,1} U)^T S_2(t)U \end{aligned}$$

Therefore, we have

$$\begin{aligned} & \| |\Theta_{1,1}^T \Theta_{2,1}|^{1/r} |\Upsilon_k| |\Theta_{1,1}^T \Theta_{2,1}|^{1/r} \|_{r'}^{r'} \\ & \leq \mathbb{E} \operatorname{tr}(|\Upsilon_k|^r (S_2(t)U)^T Z_{\eta,1} U (Z_{\eta,1} U)^T S_2(t)U) \end{aligned}$$

Following similar calculations as in CASE (I), we obtain

$$\begin{aligned} & \mathbb{E} \operatorname{tr}(|\Upsilon_k|^r (S_2(t)U)^T Z_{\eta,1} U (Z_{\eta,1} U)^T S_2(t)U) \\ & = \mathbb{E} \operatorname{tr}(|\Upsilon_k|^r \mathbb{E}_\eta (S_2(t)U)^T Z_{\eta,1} U (Z_{\eta,1} U)^T S_2(t)U) \\ & = \mathbb{E} \operatorname{tr}(|\Upsilon_k|^r \frac{1}{S} (\sum_{(l,\gamma)} (S_2(t)U)^T Z_{(l,\gamma),1} U (Z_{(l,\gamma),1} U)^T S_2(t)U)) \\ & = \frac{1}{S} \mathbb{E} \operatorname{tr}(|\Upsilon_k|^r \mathbb{E}_{Z_{(l,\gamma),1}} (\sum_{(l,\gamma)} (S_2(t)U)^T Z_{(l,\gamma),1} U (Z_{(l,\gamma),1} U)^T S_2(t)U)) \\ & \leq \frac{1}{S} \| |\Upsilon_k|^r \|_{\frac{2q}{2q-2}} \cdot \| \mathbb{E}_{Z_{(l,\gamma),1}} (\sum_{(l,\gamma)} (S_2(t)U)^T Z_{(l,\gamma),1} U (Z_{(l,\gamma),1} U)^T S_2(t)U) \|_q \end{aligned}$$

because η and $Z_{(l,\gamma),1}$ are independent with Υ_k and $S_2(t)$.

By Lemma 7.11, we still get

$$\begin{aligned} & (\text{factor } 1)^2 \\ & \leq S^{-1} (c_{7.10}(pm))^{\frac{1}{q}} \sqrt{\max\{pd, q\}}^{\frac{2qk-4q}{2q-2}} \prod_{\lambda=1}^{k/2} \|\Upsilon_\lambda\|_\infty^2 \cdot \|\Upsilon_k\|_{\frac{2q}{2q-k}} (\mathbb{E} \operatorname{tr}((\Gamma(t))^{2q}))^{(1/q)(1/r)} \end{aligned}$$

We can repeat the same argument for (factor 2) and show that

$$\begin{aligned} & (\text{factor } 2)^2 \\ & \leq S^{-1} (c_{7.10}(pm))^{\frac{1}{q}} \sqrt{\max\{pd, q\}}^{\frac{2qk-4q}{2q-2}} \prod_{\lambda=k/2+1}^{k-1} \|\Upsilon_\lambda\|_\infty^2 \cdot \|\Upsilon_k\|_{\frac{2q}{2q-k}} (\mathbb{E} \operatorname{tr}((\Gamma(t))^{2q}))^{(1/q)(1/r)} \end{aligned}$$

Combining the estimates for (factor 1) and (factor 2) together, we have

$$\begin{aligned} & \sum_{(l,\gamma) \in [n] \times [pm]} \mathbb{E} [\operatorname{tr} \Theta_{(l,\gamma),1,\tau_1(1)}^T \Theta_{(l,\gamma),1,\tau_1(2)} \Upsilon_1 \Theta_{(l,\gamma),2,\tau_2(1)}^T \Theta_{(l,\gamma),2,\tau_2(2)} \\ & \quad \cdots \Upsilon_2 \Theta_{(l,\gamma),k,\tau_k(1)}^T \Theta_{(l,\gamma),k,\tau_k(2)} \Upsilon_k] \\ & \leq (c_{7.10}(pm))^{\frac{1}{q}} \sqrt{\max\{pd, q\}}^{\frac{2qk-4q}{2q-2}} (\mathbb{E} \operatorname{tr}((\Gamma(t))^{2q}))^{\frac{1}{q} \cdot \frac{2q-k}{2q-2}} \prod_{\lambda=1}^{k-1} \|\Upsilon_\lambda\|_\infty \cdot \|\Upsilon_k\|_{\frac{2q}{2q-k}} \end{aligned}$$

Now, we have shown that

$$\begin{aligned} & \sup_{(\Upsilon_1, \dots, \Upsilon_k) \in L_\infty(S_\infty^d)^k} \frac{F(\Upsilon_1, \dots, \Upsilon_k)}{(\prod_{\lambda=1}^{k-1} \|\Upsilon_\lambda\|_\infty) \cdot \|\Upsilon_k\|_{\frac{2q}{2q-k}}} \\ & \leq (c_{7.10}(pm))^{\frac{1}{q}} \sqrt{\max\{pd, q\}}^{\frac{2qk-4q}{2q-2}} (\mathbb{E} \operatorname{tr}((\Gamma(t))^{2q}))^{\frac{1}{q} \cdot \frac{2q-k}{2q-2}} \end{aligned}$$

The final result follows from Corollary 6.6. \square

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 $(S_1 U)^T S_2 U + (S_2 U)^T S_1 U$ 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 7.5 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 7.8 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 8.9, 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, \dots, 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, \dots, 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 $[b_l(\gamma_l - 1) + 1 : \min\{b_l \gamma_l, m\}]$;
- $\alpha_{(l, \gamma_l)} := \sqrt{p(\min\{b_l \gamma_l, m\} - b_l(\gamma_l - 1))}$;
- $e_{\mu_{(l, \gamma_l)}}$ 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 $\{\xi_{(l, \gamma_l)}\}_{l \in [n], \gamma_l \in [s_l]}$ and $\{\mu_{(l, \gamma_l)}\}_{l \in [n], \gamma_l \in [s_l]}$ 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

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

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

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

The matrix Π has $O(n + \beta pmd)$ many non-zero entries and can be applied to an $n \times d$ matrix A in $O(\text{nnz}(A) + \beta pmd^2)$ 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 $pm \geq \frac{c \log(\frac{d}{\varepsilon\delta})}{\varepsilon^2}$.

Proof. The proof of the main statement using Theorem 8.5 is identical to the proof of Theorem 3.2. As in the proof of Theorem 3.2, we shall assume that the collection of random variables $\{\xi_{l,\gamma_l}\}_{l \in [n], \gamma_l \in [s_l]}$ and $\{\mu_{l,\gamma_l}\}_{l \in [n], \gamma_l \in [s_l]}$ (See Definition 8.1) are fully independent in our calculations, and since the proof proceeds via looking at moments of order $O(\log(\frac{d}{\varepsilon\delta}))$, the same calculations still hold when we have log-wise independence.

In this case, we require q to satisfy

$$pm \geq \left(\max \left\{ c_{8.5.2} \sqrt{e}(q)^{5/2}/\varepsilon, c_{8.5.3}(q)^3 \right\} \right)^{1 + \frac{1}{q-2}}$$

. Just as in the case of OSNAP, it is enough to ensure that $c_{8.5.1} \frac{d + \log(d/\varepsilon\delta)}{(\varepsilon/\sqrt{e})^2} \leq m$, and

$$pm \geq C_1 \max \left\{ \frac{C_2 (\log(\frac{d}{\varepsilon\delta}))^{2.5}}{\varepsilon}, C_3 (\log(\frac{d}{\varepsilon\delta}))^3 \right\}$$

where the form of the constants C_1, C_2 and C_3 are analogous to those in (7.1), and set $q = \lceil 2 \log(\frac{d}{\varepsilon\delta}) \rceil + 2$.

However, the value of p satisfying $pm \geq C_1 \max \left\{ \frac{C_2 (\log(\frac{d}{\varepsilon\delta}))^{2.5}}{\varepsilon^{1+4/d}}, C_3 (\log(\frac{d}{\varepsilon\delta}))^3 \right\}$ has to be smaller than 1 for our construction to be defined. So we must also have

$$m \geq C_1 \max \left\{ \frac{C_2 (\log(\frac{d}{\varepsilon\delta}))^{2.5}}{\varepsilon}, C_3 (\log(\frac{d}{\varepsilon\delta}))^3 \right\}$$

It is enough to ensure,

$$\begin{aligned} m &\geq \frac{\max\{C_1 C_2, C_1 C_3\} (\log(d/\delta\varepsilon))^3}{\varepsilon} \\ &= \max\{C_1 C_2, C_1 C_3\} \left(\frac{\log(1/\varepsilon)^3}{\varepsilon} + \frac{3 \log(1/\varepsilon)^2 \log(d/\delta)}{\varepsilon} + \frac{3 \log(1/\varepsilon) \log(d/\delta)^2}{\varepsilon} + \frac{\log(d/\delta)^3}{\varepsilon} \right) \end{aligned}$$

Now, we observe that

$$\begin{aligned} \frac{\log(1/\varepsilon)^2}{\varepsilon} + \frac{3 \log(1/\varepsilon)^2 \log(d/\delta)}{\varepsilon} &\leq \frac{1}{\varepsilon^2} + \frac{3 \log(d/\delta)}{\varepsilon^2} \\ \text{and, } \frac{3 \log(1/\varepsilon) \log(d/\delta)^2}{\varepsilon} &\leq \frac{3 \log(d/\delta)^2}{\varepsilon^2} \end{aligned}$$

So it is enough to have, for some $C_4 > 0$,

$$m \geq C_4 \frac{d + \log(d/\delta)^2 + \log(1/\varepsilon)}{\varepsilon^2} + \frac{\log(d/\delta)^3}{\varepsilon}$$

We proceed to estimate the number of non-zero entries. With reference to Definition 8.1, we have,

$$b_j = \left\lfloor \frac{1}{\beta_1 p z_j} \right\rfloor \geq \frac{1}{2\beta_1 p z_j}$$

when $\frac{1}{\beta_1 p z_j} \geq 1$. Thus,

$$s_j = \left\lceil \frac{m}{b_j} \right\rceil \leq \lceil 2\beta_1 p m z_j \rceil \leq \max\{1, 4\beta_1 p m z_j\}$$

When $\frac{1}{\beta_1 p z_j} \leq 1$,

$$s_j = \left\lceil \frac{m}{b_j} \right\rceil = m \leq m\beta_1 p z_j$$

The total number of non-zero entries in S is,

$$\sum_{j=1}^n s_j \leq n + \sum_{j=1}^n 4\beta_1 p m z_j \leq n + 4\beta_1 \beta_2 p m d$$

□

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

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

Consequently,

$$\mathbb{E}[\text{tr}(U^T S^T S U - pm \cdot I_d)^{2q}] \leq \mathbb{E}_{S, S'} \left[\text{tr} \left(2((SU)^T S' U + (SU)^T S' U) \right)^{2q} \right]$$

where S' is an independent copy of S .

Proof. Letting s_{ij} denote the entries of S ,

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

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

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

By construction, $\sum_{i=1}^m s_{ij}^2 = \sum_{\gamma_j=1}^{s_j} \alpha_{(j, \gamma_j)}^2 = pm$, so,

$$U^T S^T S U - pm \cdot I_d = \sum_{i=1}^m \sum_{\substack{j, j'=1 \\ j \neq j'}}^n s_{ij} s_{ij'} u_j u_{j'}^T$$

From this point on, the proof is exactly the same as Lemma 7.2. \square

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 $\text{Var}(s_{ij}) = p$ for all $i \in [m], j \in [n]$, and $\text{Cov}(s_{i_1 j_1}, s_{i_2 j_2}) = 0$ 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. We have

$$\begin{aligned} \mathbb{E}(s_{ij}) &= \alpha_{(j, \gamma_j)}^2 \mathbb{E}(\xi_{j, \gamma_j}) \mathbb{P}(\mu_{(j, \gamma)} = i) \\ &= \alpha_{(j, \gamma_j)}^2 \cdot 0 \cdot \mathbb{P}(\mu_{(j, \gamma)} = i) \\ &= 0 \end{aligned}$$

because $\mathbb{E}(\xi_{j, \gamma_j}) = 0$.

$$\begin{aligned} \mathbb{E}(s_{ij}^2) &= \alpha_{(j, \gamma_j)}^2 \mathbb{E}(\xi_{j, \gamma_j}^2) \mathbb{P}(\mu_{(j, \gamma)} = i) \\ &= (\sqrt{p(\min\{b_l \gamma_l, m\} - b_l(\gamma_l - 1))})^2 \cdot 1 \cdot \frac{1}{(\min\{b_l \gamma_l, m\} - b_l(\gamma_l - 1))} \\ &= p \end{aligned}$$

For the covariances, we first observe that $\text{Cov}(s_{i_1 j_1}, s_{i_2 j_2}) = 0$ if (i_1, j_1) and (i_2, j_2) belong to two different subcolumns by independence. If (i_1, j_1) and (i_2, j_2) belong to the same subcolumn, we have

$$\text{Cov}(s_{i_1 j_1}, s_{i_2 j_2}) = \mathbb{E}(s_{i_1 j_2} s_{i_2 j_1}) = \mathbb{E}(0) = 0$$

because each subcolumn has one hot distribution which means at most one of $s_{i_1 j_1}$ and $s_{i_2 j_2}$ can be nonzero. \square

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 $U^T U = I$ with given (β_1, β_2) -approximate leverage scores. Define $X = \frac{1}{\sqrt{pm}} S U$.*

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 $pm \geq \left(\max \left\{ \frac{c_{8.5.2} q^{5/2}}{\varepsilon}, c_{8.5.3} q^3 \right\} \right)^{1 + \frac{2}{q-2}}$,

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

Proof. The structure of the proof is the same as the proof of Theorem 7.3 and we only highlight changes in the specific values. By Lemma 8.6, we have,

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

As in the proof of Theorem 7.3, we divide the analysis into two cases - (i) $pd < q^2$ and (ii) $pd \geq q^2$.

In the first case, following the steps in the proof of Theorem 7.3, we get,

$$\begin{aligned} & (\mathbb{E}[\text{tr} \Gamma(S_1, S_2)^{2q}]^\alpha - (\mathbb{E}[\text{tr} \Gamma(G_1, G_2)^{2q}]^\alpha) \\ & \leq (c_5 q)^4 ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q^2\}})^{\frac{2q}{q-1}} \cdot \frac{1}{q-1} + ((c_5 q)^{2q-4} ((pm)^{\frac{1}{q}} \sqrt{\max\{pd, q^2\}})^{\frac{2q^2-4q}{q-1}})^{\frac{\alpha}{1-\alpha}} \\ & \leq (c_5 q)^4 (q)^{\frac{2q}{q-1}} (pm)^{\frac{2}{q-1}} \cdot \frac{1}{q-1} + ((c_5 q)^{2q-4} ((pm)^{\frac{1}{q}} q)^{\frac{2q^2-4q}{q-1}})^{\frac{1}{q-2}} \\ & = (c_5 q)^4 q^2 (q)^{\frac{2}{q-1}} (pm)^{\frac{2}{q-1}} \cdot \frac{1}{q-1} + ((c_5 q)^2 q^{\frac{2q}{q-1}} (pm)^{\frac{2}{q-1}}) \\ & \leq c_6 q^5 (pm)^{\frac{2}{q-1}} \end{aligned}$$

Therefore, we have

$$(\mathbb{E}[\text{tr} \Gamma(S_1, S_2)^{2q}]^{\frac{1}{q-1}} - (\mathbb{E}[\text{tr} \Gamma(G_1, G_2)^{2q}]^{\frac{1}{q-1}}) \leq c_6 q^5 (pm)^{\frac{2}{q-1}}$$

which gives,

$$(\mathbb{E}[\text{tr} \Gamma(S_1, S_2)^{2q}]^{\frac{1}{2q}} - (\mathbb{E}[\text{tr} \Gamma(G_1, G_2)^{2q}]^{\frac{1}{2q}}) \leq (c_6 q^5)^{\frac{q-1}{2q}} \leq c_7 q^{5/2} (pm)^{\frac{1}{q}}$$

so in this case, we need $pm \geq c_8(q)^{5/2}/\varepsilon$.

The case when $pd > q^2$ gives the same lower bound on pm as the corresponding case in Theorem 7.3. □

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 $U^T U = I$ with given (β_1, β_2) -approximate leverage scores. Let G_1 and G_2 be independent random matrices with i.i.d. Gaussian entries each with variance p , and define the interpolated random matrices,*

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

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

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

Proof. Following the steps in the proof of Lemma 7.5, we get,

$$\begin{aligned}
& \frac{d}{dt} \mathbb{E}[f(S_1(t), S_2(t))] \\
&= \frac{1}{2} \sum_{k=4}^{2q} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{\pi \in P([k])} (-1)^{|\pi|-1} (|\pi| - 1)! \\
& \quad \mathbb{E} \left[\sum_{l \in [n], \gamma_l \in [s_l]} \partial_{Z_{(l, \gamma_l), 1|\pi}} \cdots \partial_{Z_{(l, \gamma_l), k|\pi}} f_{1, S_2(t)}(S_1(t)) \right] \\
& \quad + \frac{1}{2} \sum_{k=4}^{2q} \frac{t^{\frac{k}{2}-1}}{(k-1)!} \sum_{\pi \in P([k])} (-1)^{|\pi|-1} (|\pi| - 1)! \\
& \quad \mathbb{E} \left[\sum_{(l, \gamma_l) \in [n] \times [pm]} \partial_{Z_{(l, \gamma_l), 1|\pi}} \cdots \partial_{Z_{(l, \gamma_l), k|\pi}} f_{2, S_1(t)}(S_2(t)) \right] \\
&=: T_1 + T_2
\end{aligned} \tag{8.2}$$

where $f_{1, S_2(t)}(S_1(t))$ and $f_{2, S_1(t)}(S_2(t))$ are as in the proof of Lemma 7.5 and $Z_{(l, \gamma_l), j|\pi}$ are constructed from $Z_{(l, \gamma_l)}$ appearing in S written as the sum $\sum_{l=1}^n \sum_{\gamma_l=1}^{s_l} Z_{l, \gamma_l}$ when S has the unscaled LESS-IC distribution. Once again we remind the reader about our assumption that the $Z_{(l, \gamma_l)}$ are independent, with the additional remark that since all quantities involved are moments of order at most $4q$, the same calculations hold even if the $Z_{(l, \gamma)}$ are $4q$ -wise independent.

In this case, we use Lemma 8.7 to estimate (8.2) as in the proof of Lemma 7.5 to get,

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

□

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 $Z_{(l, \gamma)} = \xi_{(l, \gamma)} e_{\mu_{(l, \gamma)}}^T$ and $Z_{(l, \gamma), \lambda} = \xi_{(l, \gamma), \lambda} e_{\mu_{(l, \gamma), \lambda}}^T$. Let $\{\Upsilon_1, \dots, \Upsilon_k\}$ 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 $\{\Upsilon_1, \dots, \Upsilon_k\}$. (In other words, $\{\Upsilon_1, \dots, \Upsilon_k\}$ 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

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

where $e_{\mu(l,\gamma),\lambda}$ represents the $\mu(l,\gamma),\lambda$ th coordinate vector. Then, given $0 \leq \rho_1, \dots, \rho_k \leq +\infty$ such that

$$\sum_{\lambda=1}^k \frac{1}{\rho_\lambda} = 1 - \frac{k}{2q}, \quad \tau_1, \dots, \tau_k \in \text{sym}(\{1, 2\}), \quad \text{there exists } c_{8.7} > 0 \text{ such that}$$

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

Proof. The structure of the proof is exactly the same as the proof of Lemma 7.8, and only the specific expressions differ. We define the functional $F(\Upsilon_1, \dots, \Upsilon_k)$ exactly as in the proof of Lemma 7.8, and proceed to prove the claim for when $\rho_1 = \dots = \rho_{k-1} = \infty$ and $\rho_k = \frac{q}{q-k}$.

Let $\eta = (\eta(1), \eta(2))$ be a uniformly distributed random variable in $\{(l, \gamma_l) | l \in [n], \gamma_l \in [s_l]\}$ and for all $\lambda \in [k]$, define random variables $\Theta_{1,\lambda} = \Theta_{\eta,\lambda,1}$ and $\Theta_{2,\lambda} = \Theta_{\eta,\lambda,2}$. Then, for $\mathcal{S} = \sum_{l=1}^n s_l$,

$$\begin{aligned} & \sum_{l \in [n], \gamma_l \in [s_l]} \mathbb{E}[\text{tr} \Theta_{(l,\gamma_l),1,\tau_1(1)}^T \Theta_{(l,\gamma_l),1,\tau_1(2)} \Upsilon_1 \Theta_{(l,\gamma_l),2,\tau_2(1)}^T \Theta_{(l,\gamma_l),2,\tau_2(2)} \\ & \quad \cdots \Upsilon_2 \Theta_{(l,\gamma_l),k,\tau_k(1)}^T \Theta_{(l,\gamma_l),k,\tau_k(2)} \Upsilon_k] \\ & = \mathcal{S} \cdot \mathbb{E}[\text{tr} \Theta_{\tau_1(1),1}^T \Theta_{\tau_1(2),1} \Upsilon_1 \Theta_{\tau_2(1),2}^T \Theta_{\tau_2(2),2} \cdots \Upsilon_2 \Theta_{\tau_k(1),k}^T \Theta_{\tau_k(2),k} \Upsilon_k] \end{aligned}$$

Defining factor 1 exactly as in Lemma 7.8, we have,

$$\begin{aligned} (\text{factor } 1)^2 & \leq \|\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}\|_{2q}^{2-2/r} \prod_{\lambda=2}^{k/2} \|\Theta_{\tau_\lambda(1),\lambda}^T \Theta_{\tau_\lambda(2),\lambda}\|_{2q}^2 \prod_{\lambda=1}^{k/2} \|\Upsilon_\lambda\|_\infty^2 \\ & \quad \cdot \|\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}\|^{1/r} |\Upsilon_k| \|\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}\|^{1/r} \|r \end{aligned}$$

To proceed, we prove an analogue of Lemma 7.10 to bound $\|\Theta_{\tau_\lambda(1),\lambda}^T \Theta_{\tau_\lambda(2),\lambda}\|_q = \|\Theta_{1,\lambda}^T \Theta_{2,\lambda}\|_q$.

Lemma 8.8 (Product of Random Rows for LESS). *Let $\Theta_{(l,\gamma),\lambda,1}, \Theta_{(l,\gamma),\lambda,2} \in \mathbb{R}^d$ be as in Lemma 8.7. Let $\eta = (\eta(1), \eta(2))$ be a uniformly distributed random variable in $\{(l, \gamma_l) | l \in [n], \gamma_l \in [s_l]\}$ such that η is independent with $\{\mathcal{Z}_\lambda\}_{\lambda \in [k]}$, S_1, S_2, G_1, G_2 . For all $\lambda \in [k]$, define random vectors $\Theta_{1,\lambda} = \Theta_{\eta,\lambda,1}$ and $\Theta_{2,\lambda} = \Theta_{\eta,\lambda,2}$. Let $q \geq 2$. Then there exists a constant $c_{8.8} > 0$ such that,*

$$\|\Theta_{1,\lambda}^T \Theta_{2,\lambda}\|_q \leq \frac{c_{8.8}(pm)^{\frac{1}{q}} \sqrt{\max\{pd, q^2\}}}{\mathcal{S}^{\frac{1}{q}}}$$

Proof. Note that for each fixed $1 \leq \lambda \leq k/2$,

$$\begin{aligned} \|\Theta_{1,\lambda}^T \Theta_{2,\lambda}\|_q & = \|\xi_{\eta,\lambda} \alpha_\eta u_{\eta(1)} e_{\mu_{\eta,\lambda}}^T S_2(t) U\|_q \\ & \leq \left(\mathbb{E}[\text{tr} |\xi_{\eta,\lambda} \alpha_\eta u_{\eta(1)} e_{\mu_{\eta,\lambda}}^T S_2(t) U|^q] \right)^{\frac{1}{q}} \\ & = \left(\mathbb{E}[\text{tr} |\alpha_\eta u_{\eta(1)} e_{\mu_{\eta,\lambda}}^T S_2(t) U|^q] \right)^{\frac{1}{q}} \\ & \leq \left(\mathbb{E} \left[\frac{1}{d} \|\alpha_\eta u_{\eta(1)}\|^q \|e_{\mu_{\eta,\lambda}}^T S_2(t) U\|^q \right] \right)^{\frac{1}{q}} \end{aligned}$$

Conditioning on η , we take expectation of the second factor over $\mu_{\eta,\lambda}$ and S_2 ,

$$\|\Theta_{1,\lambda}^T \Theta_{2,\lambda}\|_q \leq \left(\mathbb{E}_\eta \left[\frac{1}{d} \|\alpha_\eta u_{\eta(1)}\|^q \mathbb{E}_{\mu_{\eta,\lambda}, S_2(t)} [\|e_{\mu_{\eta,\lambda}}^T S_2(t) U\|^q] \right] \right)^{\frac{1}{q}}$$

Note that, conditioned on η , $\mu_{\eta,\lambda}$ is uniformly distributed over a subset of $[m]$. Therefore, by Lemma 8.9, we have

$$\begin{aligned} \|\Theta_{1,\lambda}^T \Theta_{2,\lambda}\|_q &\leq \left(\mathbb{E}_\eta \left[\frac{1}{d} \|\alpha_\eta u_{\eta(1)}\|^q \mathbb{E}_{\mu_{\eta,\lambda}, S_2(t)} [\|e_{\mu_{\eta,\lambda}}^T S_2(t) U\|^q] \right] \right)^{\frac{1}{q}} \\ &\leq c_{8.9} \sqrt{\max\{pd, q^2\}} \left(\mathbb{E}_\eta \left[\frac{1}{d} \|\alpha_\eta u_{\eta(1)}\|^q \right] \right)^{\frac{1}{q}} \\ &\leq c_{8.9} \sqrt{\max\{pd, q^2\}} \left(\frac{1}{\mathcal{S}d} \sum_{l=1}^n \sum_{\gamma_l=1}^{s_l} |\alpha_{(l,\gamma_l)}|^q \|u_l\|^q \right)^{\frac{1}{q}} \end{aligned}$$

Note that when $b_l \geq m$, we have $s_l = 1$ and $|\alpha_{(l,\gamma_l)}| \leq \sqrt{pm}$. Moreover, $b_l \geq m$ only when $1/\beta_1 p z_l \geq m$, which means we have $pm \|u_l\|^2 \leq \|u_l\|^2 / \beta_1 z_l \leq 1$. This means, $|\alpha_{(l,\gamma_l)}|^q \|u_l\|^q \leq (pm \|u_l\|^2)^{q/2} \leq pm \|u_l\|^2$.

When $b_l < m$, we have $|\alpha_{(l,\gamma_l)}| \|u_l\| \leq \sqrt{pb_j} \|u_l\| \leq \|u_l\| / \sqrt{\beta_1 z_j} \leq 1$ when $\lfloor 1/\beta_1 p z_j \rfloor \geq 1$ and $b_j = \lfloor 1/\beta_1 p z_j \rfloor$ because $pb_j \leq 1/\beta_1 z_j \leq 1/\|u_j\|^2$. When $b_j = 1$, we still have $|\alpha_{(l,\gamma_l)}| \|u_l\| \leq \sqrt{pb_j} \|u_l\| \leq \sqrt{p} \|u_l\| \leq 1$ since $p \leq 1$.

So, we have $\sum_{\gamma_l=1}^{s_l} |\alpha_{(l,\gamma_l)}|^q \|u_l\|^q \leq \sum_{\gamma_l=1}^{s_l} |\alpha_{(l,\gamma_l)}|^2 \|u_l\|^2 \leq pb_l s_l \|u_l\|^2$ (Since $\alpha_{(l,\gamma_l)} \leq \sqrt{pb_l}$ by definition). But $s_l = \lceil m/b_l \rceil \leq 2m/b_l$. So, $pb_l s_l \|u_l\|^2 \leq 2pm \|u_l\|^2$.

We can now continue to bound,

$$\begin{aligned} \|\Theta_{\tau_j(1),j}^T \Theta_{\tau_j(2),j}\|_q &\leq c_{8.9} \sqrt{\max\{pd, q^2\}} \left(\frac{1}{\mathcal{S}d} \sum_{l=1}^n \sum_{\gamma_l=1}^{s_l} |\alpha_{(l,\gamma_l)}|^q \|u_l\|^q \right)^{\frac{1}{q}} \\ &\leq c_{8.9} \sqrt{\max\{pd, q^2\}} \left(\frac{1}{\mathcal{S}d} \sum_{l=1}^n 2pm \|u_l\|^2 \right)^{\frac{1}{q}} \\ &\leq c_{8.9} \sqrt{\max\{pd, q^2\}} \left(\frac{2pmd}{\mathcal{S}d} \right)^{\frac{1}{q}} \\ &\leq c_{8.9} \sqrt{\max\{pd, q^2\}} \left(\frac{1}{\mathcal{S}} \right)^{\frac{1}{q}} (2pm)^{\frac{1}{q}} \end{aligned}$$

□

Using the same calculations as Lemma 7.8 for the term

$$\| |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r} |\Upsilon_k| |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r} \|_r$$

we get

$$\begin{aligned}
(\text{factor } 1)^2 &\leq \left(\frac{c_{8.8}(pm)^{\frac{1}{q}} \sqrt{\max\{pd, q^2\}}}{\mathcal{S}^{\frac{1}{2q}}} \right)^{k-2/r} \prod_{\lambda=1}^{k/2} \|\Upsilon_\lambda\|_\infty^2 \cdot \left(\frac{1}{\mathcal{S}} \|\Upsilon_k\|^r_{\frac{2q}{2q-2}} \cdot (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{\frac{1}{q}} \right)^{\frac{1}{r}} \\
&\leq \frac{1}{\mathcal{S}} \left(c_{8.8}(pm)^{\frac{1}{q}} \sqrt{\max\{pd, q^2\}} \right)^{k-2/r} \prod_{\lambda=1}^{k/2} \|\Upsilon_\lambda\|_\infty^2 \cdot \|\Upsilon_k\|_{\frac{2q}{2q-k}}^{\frac{2q}{2q-k}} \cdot (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{\frac{1}{qr}}
\end{aligned}$$

By repeating the same argument for factor 2, we get,

$$\begin{aligned}
&\sum_{l \in [n], \gamma_l \in [s_l]} \mathbb{E}[\text{tr} \Theta_{(l, \gamma_l), 1, \tau_1(1)}^T \Theta_{(l, \gamma_l), 1, \tau_1(2)} \Upsilon_1 \Theta_{(l, \gamma_l), 2, \tau_2(1)}^T \Theta_{(l, \gamma_l), 2, \tau_2(2)} \\
&\quad \cdots \Upsilon_2 \Theta_{(l, \gamma_l), k, \tau_k(1)}^T \Theta_{(l, \gamma_l), k, \tau_k(2)} \Upsilon_k] \\
&\leq (c_{8.8}(pm)^{\frac{1}{q}} \sqrt{\max\{pd, q^2\}})^{\frac{2qk-4q}{2q-2}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{\frac{1}{q} \cdot \frac{2q-k}{2q-2}} \prod_{\lambda=1}^{k-1} \|\Upsilon_\lambda\|_\infty \cdot \|\Upsilon_k\|_{\frac{2q}{2q-k}}^{\frac{2q}{2q-k}}
\end{aligned}$$

□

Lemma 8.9 (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. Gaussian entries with variance p . Let μ be a random variable uniformly distributed in $\phi \neq I \subset [m]$ and independent of S and G . Then, there exists $c_{8.9} > 0$, such that*

$$\mathbb{E}_{\mu, S(t)}[\|e_\mu^T S(t)U\|^q]^{\frac{1}{q}} \leq c_{8.9} \sqrt{\max\{pd, q^2\}}$$

Proof. By Hölder's inequality, it suffices to prove these bounds for moments of the order of the smallest even integer bigger than q , so without loss of generality, we may assume that q is an even integer. Moreover,

$$\mathbb{E}_{\mu, S(t)}[\|e_\mu^T S(t)U\|^q] = \frac{1}{|I|} \sum_{i \in I} \mathbb{E}_{S(t)}[\|e_i^T S(t)U\|^q]$$

In what follows, we fix $i \in I$, and obtain a bound for this fixed i . We observe that,

$$\begin{aligned}
\mathbb{E}_{S(t)}[\|e_i^T S(t)U\|^q]^{\frac{1}{q}} &= \mathbb{E}_{S(t)}[\|e_i^T (\sqrt{t}S + \sqrt{1-t}G)U\|^q]^{\frac{1}{q}} \\
&\leq \mathbb{E}_{S(t)}[(\|e_i^T SU\| + \|e_i^T GU\|)^q]^{\frac{1}{q}} \\
&\leq \mathbb{E}_S[\|e_i^T SU\|^q]^{\frac{1}{q}} + \mathbb{E}_G[\|e_i^T GU\|^q]^{\frac{1}{q}}
\end{aligned}$$

Since $\mathbb{E}_G[\|e_i^T GU\|^q]^{\frac{1}{q}}$ is simply the q^{th} moment of a d -dimensional Gaussian random vector whose independent components have variance p , $\mathbb{E}_G[\|e_i^T GU\|^q]^{\frac{1}{q}} \leq c_1 \sqrt{\max\{pd, pq\}}$, and it is enough to prove the bound in the statement of the lemma for $\mathbb{E}_S[\|e_i^T SU\|^q]^{\frac{1}{q}}$.

Note that in the case when S has the unscaled LESS-IC distribution, $S_{1,1}$ is non-zero when the one hot distribution on the column submatrix $S_{[1:\min\{b_1, m\}] \times 1}$ has its non zero entry on the first row. The probability that this happens is $1/\min\{b_1, m\} = p/\alpha_{1,1}^2$. To generalise this to the entry $S_{i,j}$ for a given (i, j) , observe that for each (i, j) , we can associate a unique tuple (j, r) with $r = r(i, j)$ such that the random variable $\mu_{j,r}$ determines whether S_{ij} is non zero. $\mu_{j,r}$ is uniformly distributed in an interval of size $\alpha_{j,r}^2/p$, so the probability that S_{ij} is non-zero is $p/\alpha_{j,r}^2$. Moreover, we assume the columns of S are independent and remark that the calculations hold even for log-wise independence

since $\|e_i^T S(t)U\|^q$ is a polynomial of order $q = O(\log(d/\varepsilon\delta))$ in the entries of the first row of S only sees products of $O(\log(d/\varepsilon\delta))$ many random variables.

From the above discussion, we can write,

$$\begin{aligned} e_i^T S U &= \sum_{j=1}^n \delta_{ij} \alpha_{j,r(i,j)} \xi_{i,j} u_j^T \\ &=: \sum_{j=1}^n Z_{ij} \end{aligned}$$

where,

- $\{\delta_{ij}\}_{j \in [n]}$ are independent Bernoulli random variables which are non zero with probability $p/\alpha_{j,r(i,j)}^2$ respectively.
- $\{\xi_{i,j}\}_{j \in [n]}$ is a collection of independent Rademacher random variables.
- $\{u_j^T\}_{j \in [n]}$ are rows of the matrix U .

We shall use a recursive relation to estimate $\mathbb{E} \left[\left\| \sum_{j=1}^n Z_{ij} \right\|^{2q} \right]^{\frac{1}{q}}$, and proceed to obtain this recursive inequality,

$$\begin{aligned} &\mathbb{E} \left[\left\| \sum_{j=1}^n Z_{ij} \right\|^{2q} \right]^{\frac{1}{q}} \\ &= \mathbb{E} \left[\left(\left\| \sum_{j=1}^n Z_{ij} \right\|^2 \right)^q \right]^{\frac{1}{q}} \\ (8.3) \quad &= \mathbb{E} \left[\left(\sum_{j=1}^n \delta_{ij}^2 \alpha_{j,r(i,j)}^2 \|u_j\|^2 + \sum_{\substack{j_1, j_2=1 \\ j_1 \neq j_2}}^n \alpha_{j_1,r(i,j_1)} \alpha_{j_2,r(i,j_2)} \delta_{ij_1} \delta_{ij_2} \xi_{i,j_1} \xi_{i,j_2} u_{j_1}^\top u_{j_2} \right)^q \right]^{\frac{1}{q}} \\ &\leq \mathbb{E} \left[\left(\sum_{j=1}^n \delta_{ij}^2 \alpha_{j,r(i,j)}^2 \|u_j\|^2 \right)^q \right]^{\frac{1}{q}} + \mathbb{E} \left[\left(\sum_{\substack{j_1, j_2=1 \\ j_1 \neq j_2}}^n \alpha_{j_1,r(i,j_1)} \alpha_{j_2,r(i,j_2)} \delta_{ij_1} \delta_{ij_2} \xi_{i,j_1} \xi_{i,j_2} u_{j_1}^\top u_{j_2} \right)^q \right]^{\frac{1}{q}} \end{aligned}$$

Using decoupling ([38, Theorem 6.1.1]),

$$\begin{aligned} &\mathbb{E} \left[\left(\sum_{\substack{j_1, j_2=1 \\ j_1 \neq j_2}}^n \alpha_{j_1,r(i,j_1)} \alpha_{j_2,r(i,j_2)} \delta_{ij_1} \delta_{ij_2} \xi_{i,j_1} \xi_{i,j_2} u_{j_1}^\top u_{j_2} \right)^q \right] \\ &\leq \mathbb{E} \left[\left(4 \sum_{j_1, j_2=1}^n \alpha_{j_1,r(i,j_1)} \alpha_{j_2,r(i,j_2)} \delta_{ij_1} \tilde{\delta}_{ij_2} \xi_{i,j_1} \tilde{\xi}_{i,j_2} u_{j_1}^\top u_{j_2} \right)^q \right] \\ &= \mathbb{E} \left[\left(4 \left\langle \alpha_{j,r(i,j)} \sum_{j=1}^n \delta_{ij} \xi_{ij} u_j, \sum_{j=1}^n \alpha_{j,r(i,j)} \tilde{\delta}_{ij} \tilde{\xi}_{ij} u_j \right\rangle \right)^q \right] \end{aligned}$$

where $\{\tilde{\delta}_{ij}, \tilde{\xi}_{ij}\}_{i \in [m], j \in [n]}$ are independent copies of $\{\delta_{ij}, \xi_{ij}\}_{i \in [m], j \in [n]}$.

Conditioning on $\{\tilde{\delta}_{ij}, \tilde{\xi}_{ij}\}_{i \in [m], j \in [n]}$, we have

$$\begin{aligned} & \mathbb{E} \left[\left(4 \left\langle \sum_{j=1}^n \alpha_{j,r(i,j)} \delta_{ij} \xi_{ij} u_j, \sum_{j=1}^n \alpha_{j,r(i,j)} \tilde{\delta}_{ij} \tilde{\xi}_{ij} u_j \right\rangle \right)^q \right] \\ &= \mathbb{E} \left[\mathbb{E} \left[\left(4 \left\langle \sum_{j=1}^n \alpha_{j,r(i,j)} \delta_{ij} \xi_{ij} u_j, \sum_{j=1}^n \alpha_{j,r(i,j)} \tilde{\delta}_{ij} \tilde{\xi}_{ij} u_j \right\rangle \right)^q \middle| \{\tilde{\delta}_{ij}, \tilde{\xi}_{ij}\}_{i \in [m], j \in [n]} \right] \right] \\ &= \mathbb{E} \left[\left\| \sum_{j=1}^n \alpha_{j,r(i,j)} \tilde{\delta}_{ij} \tilde{\xi}_{ij} u_j \right\|^q \mathbb{E} \left[\left(4 \left\langle \sum_{j=1}^n \alpha_{j,r(i,j)} \delta_{ij} \xi_{ij} u_j, v \right\rangle \right)^q \right] \right] \end{aligned}$$

for some fixed unit vector $v \in \mathbb{R}^d$. Plugging this estimate back into (8.3),

$$\begin{aligned} \mathbb{E} \left[\left\| \sum_{j=1}^n Z_{ij} \right\|^{2q} \right]^{\frac{1}{q}} &\leq \mathbb{E} \left[\left(\sum_{j=1}^n \delta_{ij}^2 \alpha_{j,r(i,j)}^2 \|u_j\|^2 \right)^q \right]^{\frac{1}{q}} \\ &\quad + \mathbb{E} \left[\left\| \sum_{j=1}^n Z_{ij} \right\|^q \right]^{\frac{1}{q}} \mathbb{E} \left[\left(4 \left\langle \sum_{j=1}^n \alpha_{j,r(i,j)} \delta_{ij} \xi_{ij} u_j, v \right\rangle \right)^q \right]^{\frac{1}{q}} \end{aligned}$$

Both $\mathbb{E} \left[\left(\sum_{j=1}^n \delta_{ij}^2 \alpha_{j,r(i,j)}^2 \|u_j\|^2 \right)^q \right]^{\frac{1}{q}}$ and $\mathbb{E} \left[\left(4 \left\langle \sum_{j=1}^n \alpha_{j,r(i,j)} \delta_{ij} \xi_{ij} u_j, v \right\rangle \right)^q \right]^{\frac{1}{q}}$ shall be computed using tail probabilities obtained via Chernoff bounds. To this end, we first look at $\mathbb{E}[\exp(4\lambda \langle \alpha_{j,r(i,j)} \delta_{ij} \xi_{ij} u_j, v \rangle)]$ for a fixed j ,

$$\begin{aligned} \mathbb{E} [\exp(4\lambda \langle \alpha_{j,r(i,j)} \delta_{ij} \xi_{ij} u_j, v \rangle)] &= 1 + \frac{p}{\alpha_{j,r(i,j)}^2} (\cosh(4\lambda \alpha_{j,r(i,j)} \langle u_j, v \rangle) - 1) \\ &= 1 + \frac{\cosh(4\lambda \alpha_{j,r(i,j)} \langle u_j, v \rangle) - 1}{16\lambda^2 \alpha_{j,r(i,j)}^2 \langle u_j, v \rangle^2} \cdot 16\lambda^2 \langle u_j, v \rangle^2 p \end{aligned}$$

Since \cosh is an even function, we may assume that $\langle u_j, v \rangle \geq 0$. Then, $4\lambda \alpha_{j,r(i,j)} \langle u_j, v \rangle \leq 4\lambda \alpha_{j,r(i,j)} \|u_j\| \leq 4\lambda \sqrt{pb_j} \|u_j\| \leq 4\lambda$ when $\lfloor 1/\beta_1 p z_j \rfloor \geq 1$ and $b_j = \lfloor 1/\beta_1 p z_j \rfloor$ because $pb_j \leq 1/\beta_1 z_j \leq 1/\|u_j\|^2$. When $b_j = 1$, we still have $4\lambda \alpha_{j,r(i,j)} \langle u_j, v \rangle \leq 4\lambda \sqrt{p} \cdot 1 \leq 4\lambda$.

Using the fact that $\frac{\cosh(x)-1}{x^2}$ is increasing in x for $x \geq 0$, we get,

$$\begin{aligned} \mathbb{E} [\exp(4\lambda \langle \alpha_{j,r(i,j)} \delta_{ij} \xi_{ij} u_j, v \rangle)] &\leq 1 + \frac{\cosh(4\lambda) - 1}{16\lambda^2} \cdot 16\lambda^2 \langle u_j, v \rangle^2 p \\ &\leq 1 + \langle u_j, v \rangle^2 p e^{4\lambda} \\ &\leq \exp(\langle u_j, v \rangle^2 p e^{4\lambda}) \end{aligned}$$

So,

$$\begin{aligned} \mathbb{E} \left[\exp \left(4\lambda \left\langle \sum_{j=1}^n \alpha_{j,r(i,j)} \delta_{ij} \xi_{ij} u_j, v \right\rangle \right) \right] &\leq \exp \left(\sum_{j=1}^n \langle u_j, v \rangle^2 p e^{4\lambda} \right) \\ &\leq \exp(p e^{4\lambda}) \end{aligned}$$

This gives a Chernoff tail bound,

$$\mathbb{P} \left[\left| \left\langle \sum_{j=1}^n \alpha_{j,r(i,j)} \delta_{ij} \xi_{ij} u_j, v \right\rangle \right| \geq t \right] \leq 2 \exp \left(\frac{t}{4} \left(1 - \log \left(\frac{t}{4p} \right) \right) \right)$$

and a standard moment computation gives, $\mathbb{E} \left[\left(4 \left\langle \sum_{j=1}^n \alpha_{j,r(i,j)} \delta_{ij} \xi_{ij} u_j, v \right\rangle \right)^q \right]^{\frac{1}{q}} \leq C_1 q$.

To deal with $\mathbb{E} \left[\left(\sum_{j=1}^n \delta_{ij}^2 \alpha_{j,r(i,j)}^2 \|u_j\|^2 \right)^q \right]^{\frac{1}{q}}$, we use the following Rosenthal's inequality from [32, p.442].

Lemma 8.10 (Theorem 14.10 from [32]). *Let $Z = \sum_{j=1}^n X_j$ where X_1, \dots, X_n are independent and nonnegative random variables. Then there exists a constant 8.10 such that, for all integers $q \geq 1$, we have*

$$(\mathbb{E}(Z^q))^{1/q} \leq 2 \mathbb{E}(Z) + c_{8.10} q (\mathbb{E}(\max_{j=1, \dots, n} X_j^q))^{1/q}$$

We use this lemma with $X_j = \delta_{ij}^2 \alpha_{j,r(i,j)}^2 \|u_j\|^2$. We first calculate that

$$\mathbb{E} \left(\sum_{j=1}^n \delta_{ij}^2 \alpha_{j,r(i,j)}^2 \|u_j\|^2 \right) = \sum_{j=1}^n (p / \alpha_{j,r(i,j)}^2) \alpha_{j,r(i,j)}^2 \|u_j\|^2 = pd$$

Next, we calculate $(\mathbb{E}(\max_{j=1, \dots, n} X_j^q))^{1/q}$. Since $\alpha_{j,r(i,j)} \leq \sqrt{pb_j}$ and $\delta_{ij} \leq 1$, we have $X_j = \delta_{ij}^2 \alpha_{j,r(i,j)}^2 \|u_j\|^2 \leq pb_j \|u_j\|^2$. Now we consider two cases depending on whether $b_j = \left\lfloor \frac{1}{\beta_1 p z_j} \right\rfloor$ or $b_j = 1$. First, if $b_j = \left\lfloor \frac{1}{\beta_1 p z_j} \right\rfloor$, then we have $X_j \leq pb_j \|u_j\|^2 \leq p \frac{1}{\beta_1 p z_j} \beta_1 z_j \leq 1$. Second, if $b_j = 1$, then we have $X_j \leq pb_j \|u_j\|^2 \leq p \leq 1$. In conclusion, we have $\max_{j=1, \dots, n} \delta_{ij}^2 \alpha_{j,r(i,j)}^2 \|u_j\|^2 \leq 1$ almost surely, and therefore we have $(\mathbb{E}(\max_{j=1, \dots, n} \delta_{ij}^2 \alpha_{j,r(i,j)}^2 \|u_j\|^2)^q)^{1/q} \leq 1$.

Therefore, we have

$$\mathbb{E} \left[\left(\sum_{j=1}^n \delta_{ij}^2 \alpha_{j,r(i,j)}^2 \|u_j\|^2 \right)^q \right]^{\frac{1}{q}} \leq 2pd + c_2 q$$

for some constant c_2 .

We thus have,

$$\mathbb{E} \left[\left\| \sum_{j=1}^n Z_{ij} \right\|^{2q} \right]^{\frac{1}{2q}} \leq \sqrt{2pd} + \sqrt{c_2 q} + \mathbb{E} \left[\left\| \sum_{j=1}^n Z_{ij} \right\|^q \right]^{\frac{1}{2q}} \sqrt{C_1 q}$$

Let $C_2 = \max\{2, c_2, (3C_1)^2\}$. Observe that $\mathbb{E} \left[\left\| \sum_{j=1}^n Z_{ij} \right\|^2 \right]^{\frac{1}{2}} = \sqrt{pd}$. Assuming that

$$\mathbb{E} \left[\left\| \sum_{j=1}^n Z_{ij} \right\|^q \right]^{\frac{1}{q}} \leq 3\sqrt{C_2} \sqrt{\max\{pd, q^2\}}$$

for some q , the above relation gives,

$$\begin{aligned} \mathbb{E} \left[\left\| \sum_{j=1}^n Z_{ij} \right\|^{2q} \right]^{\frac{1}{2q}} &\leq \sqrt{C_2 p d} + \sqrt{C_2 q} + \left(3\sqrt{C_2} \sqrt{\max\{pd, q^2\}} \right)^{\frac{1}{2}} \sqrt{C_1 q} \\ &\leq \left(2\sqrt{C_2} + \sqrt{3C_1 C_2^{1/4}} \right) \sqrt{\max\{pd, q^2\}} \\ &\leq 3\sqrt{C_2} \sqrt{\max\{pd, q^2\}} \end{aligned}$$

□

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 called 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 $U^T U = I$. 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*

$$(9.1) \quad \mathbb{P}(1 - \varepsilon \leq s_{\min}((1/\sqrt{pm})SU) \leq s_{\max}((1/\sqrt{pm})SU) \leq 1 + \varepsilon) \geq 1 - \delta$$

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

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

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 $U^T S^T S U - pm \cdot I_d$ as

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

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,

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

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 $q \geq \log(d)$, there exist constants $c_{9.3}$ such that for $m \geq d \geq 20$ and $pm \geq c_{9.3} \frac{q}{\varepsilon^2}$,*

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

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

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

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. First, note that,

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

and we shall use Matrix Bernstein's inequality to estimate this moment. We shall only use a specific case of this theorem that holds more generally.

Lemma 9.4 (Matrix Bernstein, Theorem 6.1.1 [12]). *For a finite sequence of $d \times d$ independent mean zero symmetric bounded random matrices $\{S_k\}$ with $\|S_k\| \leq R$, let,*

$$\sigma = \left\| \sum_k \mathbb{E}[S_k^2] \right\|$$

Then,

$$\mathbb{P} \left(\left\| \sum_k S_k \right\| \geq t \right) \leq 2d \exp \left(\frac{-t^2/2}{\sigma + \frac{Rt}{3}} \right)$$

Recall that we are assuming that the entries of S are fully independent, but since we are only dealing with moments of order up to $4q$, the same bounds will hold even if the entries of S are $4q$ -wise independent.

In our case, for the sum $\sum_{i=1}^m \sum_{j=1}^n (s_{ij}^2 - p)u_j u_j^T$, we have $R = 1$ and,

$$\begin{aligned}\sigma &= \left\| \sum_{i=1}^m \sum_{j=1}^n \mathbb{E}[(s_{ij}^2 - p)^2] \|u_j\|^2 u_j u_j^T \right\| \\ &= \left\| \sum_{j=1}^n pm(1-p) \|u_j\|^2 u_j u_j^T \right\| \\ &\leq pm\end{aligned}$$

where we use the fact that $\sum_{j=1}^n (1-p) \|u_j\|^2 u_j u_j^T \prec \sum_{j=1}^n u_j u_j^T = I_d$. Then,

$$\begin{aligned}\mathbb{P}\left(\left\| \sum_{i=1}^m \sum_{j=1}^n (s_{ij}^2 - p)u_j u_j^T \right\| \geq t'\right) &\leq 2d \exp\left(\frac{-(t')^2/2}{pm + \frac{t}{3}}\right) \\ \mathbb{P}\left(\left\| \left(\frac{1}{pm} \sum_{i=1}^m \sum_{j=1}^n s_{ij}^2 u_j u_j^T\right) - I_d \right\| \geq \frac{t'}{pm}\right) &\leq 2d \exp\left(\frac{-(t')^2/2}{pm + \frac{t}{3}}\right) \\ \mathbb{P}\left(\left\| \left(\frac{1}{pm} \sum_{i=1}^m \sum_{j=1}^n s_{ij}^2 u_j u_j^T\right) - I_d \right\| \geq t\right) &\leq 2d \exp\left(\frac{-(pmt)^2/2}{pm + \frac{pmt}{3}}\right) \\ &= 2d \exp\left(\frac{-pmt^2/2}{1 + \frac{t}{3}}\right)\end{aligned}$$

Let X denote $\left\| \left(\frac{1}{pm} \sum_{i=1}^m \sum_{j=1}^n s_{ij}^2 u_j u_j^T\right) - I_d \right\|$ for convenience. Then we have the mixed tail,

$$\mathbb{P}(X \geq t) \leq \begin{cases} 2d \exp(-pmt^2/4) & \text{for } 0 \leq t \leq 3 \\ 2d \exp(-3pmt/4) & \text{for } 3 < t \end{cases}$$

We are interested in $\mathbb{E}[X^{2q}]^{\frac{1}{2q}}$, and, for $\varepsilon < 1$,

$$\begin{aligned}\mathbb{E}[X^{2q}] &= 2q \int_0^\infty t^{2q-1} \mathbb{P}(X \geq t) dt \\ &= 2q \int_0^\varepsilon t^{2q-1} \mathbb{P}(X \geq t) dt + 2q \int_\varepsilon^3 t^{2q-1} \mathbb{P}(X \geq t) dt + 2q \int_3^\infty t^{2q-1} \mathbb{P}(X \geq t) dt \\ &\leq 2q \cdot \varepsilon^{2q} + 2q \int_\varepsilon^3 t^{2q-1} \mathbb{P}(X \geq t) dt + 2q \int_3^\infty t^{2q-1} \mathbb{P}(X \geq t) dt\end{aligned}$$

When $pm \geq 4q/\varepsilon^2$,

$$\begin{aligned}2q \int_\varepsilon^3 t^{2q-1} \mathbb{P}(X \geq t) dt &\leq 2q \int_\varepsilon^3 t^{2q-1} 2d \exp(-pmt^2/4) dt \\ &\leq 2q \int_\varepsilon^3 t^{2q-1} 2d \exp(-2qt^2/2\varepsilon^2) dt\end{aligned}$$

Using change of variables $\hat{t} = \sqrt{2q}t/\varepsilon$,

$$\begin{aligned}
2q \int_{\varepsilon}^3 t^{2q-1} \mathbb{P}(X \geq t) dt &\leq 2q \cdot 2d \int_0^{\infty} \left(\frac{\varepsilon}{\sqrt{2q}} \right)^{2q} \hat{t}^{2q-1} \exp(-\hat{t}^2/2) d\hat{t} \\
&\leq 4qd \left(\frac{\varepsilon}{\sqrt{2q}} \right)^{2q} \int_0^{\infty} \hat{t}^{2q-1} \exp(-\hat{t}^2/2) d\hat{t} \\
&\leq \sqrt{2\pi} 4qd \left(\frac{\varepsilon \sqrt{2q-1}}{\sqrt{2q}} \right)^{2q} \\
&\leq \sqrt{2\pi} 4qd (\varepsilon)^{2q}
\end{aligned}$$

Continuing,

$$\begin{aligned}
2q \int_3^{\infty} t^{2q-1} \mathbb{P}(X \geq t) dt &\leq 4qd \int_3^{\infty} t^{2q-1} \exp(-3pmt/4) dt \\
&\leq 4qd \int_{\varepsilon}^3 t^{2q-1} \exp(-3qt/\varepsilon^2) dt
\end{aligned}$$

with the change of variables $\tilde{t} = 3qt/\varepsilon^2$,

$$\begin{aligned}
2q \int_3^{\infty} t^{2q-1} \mathbb{P}(X \geq t) dt &\leq 4qd \left(\frac{\varepsilon^2}{3q} \right)^{2q} \int_0^{\infty} (\tilde{t})^{2q-1} \exp(-\tilde{t}) d\tilde{t} \\
&\leq 4qd \left(\frac{\varepsilon^2(2q-1)}{3q} \right)^{2q} \\
&\leq 4qd (\varepsilon^2)^{2q}
\end{aligned}$$

So,

$$\mathbb{E}[X^{2q}]^{\frac{1}{2q}} \leq (18qd)^{\frac{1}{2q}} \varepsilon \leq \exp\left(\frac{2\log d + \log 18}{2q}\right) \varepsilon \leq e^{1.5} \varepsilon$$

By starting with $\varepsilon/e^{1.5}$ instead, we can get the claim made in the statement. \square

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 $U^T U = I$. 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*

$$(9.1) \quad \mathbb{P}(1 - \varepsilon \leq s_{\min}((1/\sqrt{pm})SU) \leq s_{\max}((1/\sqrt{pm})SU) \leq 1 + \varepsilon) \geq 1 - \delta$$

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

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

Proof. By the proof of Theorem [3.2](#), we see that it is enough to show that $\mathbb{E}[\text{tr}(X^T X - I_d)^{2q}]^{\frac{1}{2q}} \leq \varepsilon$, or, equivalently, $\mathbb{E}[\text{tr}(U^T S^T S U - pm \cdot I_d)^{2q}]^{\frac{1}{2q}} \leq pm\varepsilon$ whenever p and q satisfy $C\left(\frac{q^2}{\varepsilon} + q^3\right) \leq pm$

and $m \geq C_1 \frac{d+q}{\varepsilon^2}$. At the beginning of Section 9, we saw that when S has the OSE-IE distribution,

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

By Proposition 9.3, the former term is bounded by $pm\varepsilon/2$ when $pm \geq 4c_{9.3} \frac{q}{\varepsilon^2}$ and $m \geq 20$. For the latter term, we see that the proof of Lemma 7.2 still applies and we have,

$$\mathbb{E} \left[\text{tr} \left(\sum_{i=1}^m \sum_{j,j'=1, j \neq j'}^n s_{ij} s_{ij'} u_j u_{j'}^T \right)^{2q} \right]^{\frac{1}{2q}} \leq 2 \mathbb{E}[\text{tr}(\Gamma(S_1, S_2))^{2q}]^{\frac{1}{2q}}$$

Following the proof of Theorem 7.3 and using Lemma 9.5, we see that $\mathbb{E}[\text{tr}(\Gamma(S_1, S_2))^{2q}]^{\frac{1}{2q}} \leq \varepsilon/4$ when $C \left(\frac{q^2}{\varepsilon} + q^3 \right)^{1+\frac{2}{q-2}} \leq pm$ and $m \geq C_1 \frac{d+q}{\varepsilon^2}$. The claim follows as in the proof of Theorem 3.2. \square

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 $(S_1 U)^T S_2 U$ and $(G_1 U)^T G_2 U$ has a slightly different proof in the OSE-IE case than in the OSNAP case, even though both bounds are the same.

Lemma 9.5. *Let $S_1(t)$ and $S_2(t)$ be as in Lemma 7.5 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 $q \geq 2$ 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*

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

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 $\{\Upsilon_1, \dots, \Upsilon_k\}$ 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 $\{\Upsilon_1, \dots, \Upsilon_k\}$. (In other words, $\{\Upsilon_1, \dots, \Upsilon_k\}$ 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 $0 \leq \beta_1, \dots, \beta_k \leq +\infty$ such that $\sum_{\lambda=1}^k \frac{1}{\beta_\lambda} = 1 - \frac{k}{2q}$, $\tau_1, \dots, \tau_k \in \text{sym}(\{1, 2\})$, there exists $c_{9.5} > 0$ such that

$$\begin{aligned} & \sum_{(l, \gamma) \in [n] \times [m]} \mathbb{E}[\text{tr} \Theta_{(l, \gamma), 1, \tau_1(1)}^T \Theta_{(l, \gamma), 1, \tau_1(2)} \Upsilon_1 \Theta_{(l, \gamma), 2, \tau_2(1)}^T \Theta_{(l, \gamma), 2, \tau_2(2)} \\ & \quad \dots \Upsilon_2 \Theta_{(l, \gamma), k, \tau_k(1)}^T \Theta_{(l, \gamma), k, \tau_k(2)} \Upsilon_k] \\ & \leq (c_{9.5}(pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}})^{\frac{2qk-4q}{2q-2}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{\frac{1}{q} \cdot \frac{2q-k}{2q-2}} \prod_{\lambda=1}^k \|\Upsilon_\lambda\|_{\beta_\lambda} \end{aligned}$$

Proof. The structure of the proof is exactly the same as the proof of Lemma 7.8, and only the specific expressions differ. We define the functional $F(\Upsilon_1, \dots, \Upsilon_k)$ exactly as in the proof of Lemma 7.8, and proceed to prove the claim for when $\beta_1 = \dots = \beta_{k-1} = \infty$ and $\beta_k = \frac{q}{q-k}$.

Let $\eta = (\eta(1), \eta(2))$ be a uniformly distributed random variable in $[n] \times [m]$ and for all $\lambda \in [k]$, define random variables $\Theta_{1, \lambda} = \Theta_{\eta, \lambda, 1}$ and $\Theta_{2, \lambda} = \Theta_{\eta, \lambda, 2}$. Then, for $\mathcal{S} = mn$,

$$\begin{aligned} & \sum_{l \in [n], \gamma \in [m]} \mathbb{E}[\text{tr} \Theta_{(l, \gamma), 1, \tau_1(1)}^T \Theta_{(l, \gamma), 1, \tau_1(2)} \Upsilon_1 \Theta_{(l, \gamma), 2, \tau_2(1)}^T \Theta_{(l, \gamma), 2, \tau_2(2)} \\ & \quad \dots \Upsilon_2 \Theta_{(l, \gamma), k, \tau_k(1)}^T \Theta_{(l, \gamma), k, \tau_k(2)} \Upsilon_k] \\ & = \mathcal{S} \cdot \mathbb{E}[\text{tr} \Theta_{\tau_1(1), 1}^T \Theta_{\tau_1(2), 1} \Upsilon_1 \Theta_{\tau_2(1), 2}^T \Theta_{\tau_2(2), 2} \dots \Upsilon_2 \Theta_{\tau_k(1), k}^T \Theta_{\tau_k(2), k} \Upsilon_k] \end{aligned}$$

Defining factor 1 exactly as in Lemma 7.8, we have,

$$\begin{aligned} (\text{factor } 1)^2 & \leq \|\Theta_{\tau_1(2), 1}^T \Theta_{\tau_1(1), 1}\|_{2q}^{2-2/r} \prod_{\lambda=2}^{k/2} \|\Theta_{\tau_\lambda(1), \lambda}^T \Theta_{\tau_\lambda(2), \lambda}\|_{2q}^2 \prod_{\lambda=1}^{k/2} \|\Upsilon_\lambda\|_\infty^2 \\ & \quad \cdot \|\Theta_{\tau_1(2), 1}^T \Theta_{\tau_1(1), 1}\|^{1/r} \|\Upsilon_k\| \|\Theta_{\tau_1(2), 1}^T \Theta_{\tau_1(1), 1}\|^{1/r} \|\Upsilon_k\| \end{aligned}$$

To proceed, we prove an analogue of Lemma 7.10 to bound $\|\Theta_{\tau_j(1), j}^T \Theta_{\tau_j(2), j}\|_q$,

Lemma 9.6. *Let $\Theta_{\tau_j(1), j}, \Theta_{\tau_j(2), j} \in \mathbb{R}^d$ be as in Lemma 9.5. Let $\eta = (\eta(1), \eta(2))$ be a uniformly distributed random variable in $[n] \times [m]$ such that η is independent with $\{\mathcal{Z}_\lambda\}_{\lambda \in [k]}$, S_1, S_2, G_1, G_2 . For all $\lambda \in [k]$, define random vectors $\Theta_{1, \lambda} = \Theta_{\eta, \lambda, 1}$ and $\Theta_{2, \lambda} = \Theta_{\eta, \lambda, 2}$. Let $q \geq \log(pm)$. Then there exists a constant $c_{9.6} > 0$ such that,*

$$\|\Theta_{\tau_j(1), j}^T \Theta_{\tau_j(2), j}\|_q \leq \frac{c_{9.6} \sqrt{\max\{pd, q\}}}{\mathcal{S}^{\frac{1}{q}}}$$

Proof. Note that for each fixed $1 \leq j \leq k/2$,

$$\begin{aligned} \|\Theta_{\tau_j(1), j}^T \Theta_{\tau_j(2), j}\|_q & = \|\xi_\eta \delta_\eta \mathbf{u}_{\eta(1)} e_{\eta(2)}^T S_2(t) U\|_q \\ & \leq \left(\mathbb{E}[\text{tr} |\xi_\eta \delta_\eta \mathbf{u}_{\eta(1)} e_{\eta(2)}^T S_2(t) U|^q] \right)^{\frac{1}{q}} \\ & = \left(\mathbb{E}[\text{tr} |\delta_\eta \mathbf{u}_{\eta(1)} e_{\eta(2)}^T S_2(t) U|^q] \right)^{\frac{1}{q}} \\ & \leq \left(\mathbb{E} \left[\frac{1}{d} \|\delta_\eta \mathbf{u}_{\eta(1)}\|^q \|e_{\eta(2)}^T S_2(t) U\|^q \right] \right)^{\frac{1}{q}} \end{aligned}$$

Conditioning on η , we take expectation of the first factor over δ_η and the second factor over S_2 ,

$$\|\Theta_{\tau_j(1),j}^T \Theta_{\tau_j(2),j}\|_q \leq \left(\mathbb{E}_\eta \left[\frac{1}{d} \mathbb{E}_{\delta_\eta} [\|\delta_\eta \mathbf{u}_{\eta(1)}\|^q] \mathbb{E}_{S_2(t)} [\|e_{\eta(2)}^T S_2(t) U\|^q] \right] \right)^{\frac{1}{q}}$$

Note that, after conditioning on η , $\eta(2)$ is fixed. In other words, $\eta(2)$ is uniformly distributed over a one element subset of $[m]$. Therefore, by Lemma 6.2, we have

$$\begin{aligned} \|\Theta_{\tau_j(1),j}^T \Theta_{\tau_j(2),j}\|_q &\leq \left(\mathbb{E}_\eta \left[\frac{1}{d} \mathbb{E}_{\delta_\eta} [\|\delta_\eta \mathbf{u}_{\eta(1)}\|^q] \mathbb{E}_{S_2(t)} [\|e_{\eta(2)}^T S_2(t) U\|^q] \right] \right)^{\frac{1}{q}} \\ &\leq c_{6.2} \sqrt{\max\{pd, q\}} \left(\mathbb{E}_\eta \left[\frac{p}{d} \|\mathbf{u}_{\eta(1)}\|^q \right] \right)^{\frac{1}{q}} \\ &\leq c_{6.2} \sqrt{\max\{pd, q\}} \left(\frac{p}{\mathcal{S}d} \sum_{l=1}^n \sum_{\gamma=1}^m \|\mathbf{u}_l\|^q \right)^{\frac{1}{q}} \\ &= c_{6.2} \sqrt{\max\{pd, q\}} \left(\frac{pm}{\mathcal{S}d} \sum_{l=1}^n \|u_l\|^q \right)^{\frac{1}{q}} \\ &\leq c_{6.2} \sqrt{\max\{pd, q\}} \left(\frac{pm}{\mathcal{S}d} \sum_{l=1}^n \|u_l\|^2 \right)^{\frac{1}{q}} \\ &\leq \frac{c_{9.6} (pm)^{\frac{1}{q}} \sqrt{\max\{pd, q\}}}{\mathcal{S}^{\frac{1}{q}}} \end{aligned}$$

where in the last line we use the fact that $\sum_{l=1}^n \|u_l\|^2 = d$.

□

Using the same calculations as Lemma 7.8 for the term

$$\| |\Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r} \Upsilon_k | \Theta_{\tau_1(2),1}^T \Theta_{\tau_1(1),1}|^{1/r} \|_r$$

we get

$$\begin{aligned} (\text{factor 1})^2 &\leq \left(\frac{c_{9.6} \sqrt{\max\{pd, q\}}}{\mathcal{S}^{\frac{1}{2q}}} \right)^{k-2/r} \prod_{\lambda=1}^{k/2} \|\Upsilon_\lambda\|_\infty^2 \cdot \left(\frac{1}{\mathcal{S}} \|\Upsilon_k\|^r_{\frac{2q}{2q-2}} \cdot (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{\frac{1}{q}} \right)^{\frac{1}{r}} \\ &\leq \frac{1}{\mathcal{S}} \left(c_{9.6} \sqrt{\max\{pd, q\}} \right)^{k-2/r} \prod_{\lambda=1}^{k/2} \|\Upsilon_\lambda\|_\infty^2 \cdot \|\Upsilon_k\|_{\frac{2q}{2q-k}}^{\frac{2q}{2q-k}} \cdot (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{\frac{1}{qr}} \end{aligned}$$

By repeating the same argument for factor 2, we get,

$$\begin{aligned} &\sum_{l \in [n], \gamma_l \in [s_l]} \mathbb{E}[\text{tr} \Theta_{(l, \gamma_l), 1, \tau_1(1)}^T \Theta_{(l, \gamma_l), 1, \tau_1(2)} \Upsilon_1 \Theta_{(l, \gamma_l), 2, \tau_2(1)}^T \Theta_{(l, \gamma_l), 2, \tau_2(2)} \\ &\quad \cdots \Upsilon_2 \Theta_{(l, \gamma_l), k, \tau_k(1)}^T \Theta_{(l, \gamma_l), k, \tau_k(2)} \Upsilon_k] \\ &\leq (c_{9.6} \sqrt{\max\{pd, q\}})^{\frac{2qk-4q}{2q-2}} (\mathbb{E} \text{tr}((\Gamma(t))^{2q}))^{\frac{1}{q} \cdot \frac{2q-k}{2q-2}} \left(\prod_{\lambda=1}^{k-1} \|\Upsilon_\lambda\|_\infty \right) \cdot \|\Upsilon_k\|_{\frac{2q}{2q-k}} \end{aligned}$$

□

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. & Nguyễn, 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. & Nguyễn, 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).