Tensor Random Projection for Low Memory Dimension Reduction

Yiming Sun * Cornell
Yang Guo * UW-Madison
Joel A. Tropp Caltech
Madeleine Udell Cornell

Abstract

Random projections reduce the dimension of a set of vectors while preserving structural information, such as distances between vectors in the set. This paper proposes a novel use of row-product random matrices [18] in random projection, where we call it Tensor Random Projection (TRP). It requires substantially less memory than existing dimension reduction maps. The TRP map is formed as the Khatri-Rao product of several smaller random projections, and is compatible with any base random projection including sparse maps, which enable dimension reduction with very low query cost and no floating point operations. We also develop a reduced variance extension. We provide a theoretical analysis of the bias and variance of the TRP, and a non-asymptotic error analysis for a TRP composed of two smaller maps. Experiments on both synthetic and MNIST data show that our method performs as well as conventional methods with substantially less storage.

1 Introduction

Random projections (RP) are commonly used to reduce the dimension of collections of high dimensional vectors, enabling a broad range of modern applications [23, 7, 2, 5, 9, 10]. In the context of large-scale relational databases, these maps enable applications like information retrieval [17], similarity search [20, 12], and privacy preserving distributed data mining [15]. Consider the problem of detecting plagiarism. We might attempt to solve this problem by comparing the similarity of word-level n-gram profiles for different pairs of documents [3]. To avoid tremendous query cost of this procedure, which scales quadratically with the number of documents, we may instead reduce the dimension of the data vector with a random projection, and cluster the resulting low-dimensional vectors. However, if the dimension of the vectors before reduction (here, the size of the lexicon) is too big, the storage cost of the random map is not negligible. Furthermore, even generating the pseudo-random numbers used to produce the random projection is expensive [16].

To reduce the storage burden, we propose a novel use of the row-product random matrices in random projection, and call it the *Tensor Random Projection* (TRP), formed as the Khatri-Rao product of a list of smaller dimension reduction maps. We show this map is an approximate isometry, with tunable accuracy, and hence can serve as a useful dimension reduction primitive. Furthermore, the storage required to compress d dimension vectors scales as $\sqrt[N]{d}$ where N is the number of smaller maps used to form the TRP. We also develop a reduced variance version of the TRP that allows separate control of the dimension of the range and the quality of the isometry.

Dimension Reduction Map A function f from $\mathbb{R}^d \to \mathbb{R}^k (k \ll d)$ is called a dimension reduction map (DRM) if it approximately preserves the pair-wise distances. More precisely, we call f a *ϵ-Johnson-Lindenstrauss (JL) transform* if for any $\epsilon > 0$ and for any two points $\mathbf{u}, \mathbf{v}$ in a discrete set $\mathcal{X} \subseteq \mathbb{R}^n$, we have

$$(1-\epsilon)\|\mathbf{u}-\mathbf{v}\|^2 \leqslant \|f(\mathbf{u}) - f(\mathbf{v})\|^2 \leqslant (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|^2$$

*Both authors contributed equally.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

The well-known JL Lemma [11] claims for $k = O(\log(|\mathcal{X}|)/\epsilon^2)$, an ϵ-JL transform exists. In fact, the proof shows that a suitable random linear map is an ϵ-JL transform with high probability.

The simplicity of linear maps makes them a favorite choice for dimension reduction. A linear map $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$ for $\mathbf{A} \subseteq \mathbb{R}^{d\times k}$ is a good DRM if has the following properties:

1. *Expected Isometry.* In expectation, the map A is an isometry: $\mathbb{E}\|\mathbf{A}\mathbf{x}\|^2 = \|\mathbf{x}\|^2$.
2. *Vanishing Variance.* $\text{Var}(\|\mathbf{A}\mathbf{x}\|^2)$ decays to zero as k increases. The variance measures the deviations from isometry, and serves as a quality metric for the DRM.
3. *Database-Friendly.* A map is database-friendly if it uses not-too-much storage (and so fits in memory), can be applied to a vector with relatively few queries to vector elements (and so uses few database lookups), and is computationally cheap to construct and apply.

Lemma A.1 in Appendix A shows any linear map that is an expected isometry with vanishing variance is a $\epsilon-$JL transform with high probability, for sufficiently large k.

Sparse random maps for low memory dimension reduction were first proposed by [1], and further work has improved the memory requirements and guarantees of these methods [14, 6]. Most closely related to our work is Rudelson's foundational study [18], which considers how the spectral and geometric properties of the random maps we use in this paper resemble a random map with iid entries, and shows that their largest and smallest singular values are of the same order. These results have been widely used to obtain guarantees for algorithmic privacy, but not for random projection. Battaglino et al. [4] use random projections of Khatri-Rao products to develop a randomized least squares algorithm for tensor factorization; in contrast, our method uses the (full) Khatri-Rao product to enable random projection. Sparse random projections to solve least squares problems were also explored in [21] and [22]. To our knowledge, this paper is the first to consider using the Khatri-Rao product for low memory random projection.

1.1 Notation

We denote *scalar*, *vector*, and *matrix* variables, respectively, by lowercase letters (x), boldface lowercase letters ($\mathbf{x}$), and boldface capital letters ($\mathbf{X}$). Let $[N] = \{1, \ldots, N\}$. For a vector $\mathbf{x}$ of size n, we let $\|\mathbf{x}\|_q = (\sum_{i=1}^n x_i^q)^{1/q}$ be its q norm for $q \geqslant 1$. For a matrix $\mathbf{X}$, we denote its i^{th} row, j^{th} column, and the $(i,j)^{th}$ element as $\mathbf{X}(i,.)$, $\mathbf{X}(.,j)$, and $\mathbf{X}(i,j)$. We let $\mathbf{A} \odot \mathbf{B}$ denote the *Khatri-Rao product*, $\mathbf{A} \in \mathbb{R}^{I\times K}, \mathbf{B} \in \mathbb{R}^{J\times K}$, i.e. the "matching column-wise" Kronecker product. The resulting matrix of size $(IJ) \times K$ is given by:

$$\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} \mathbf{A}(1,1)\mathbf{B}(\cdot,1) & \cdots & \mathbf{A}(1,K)\mathbf{B}(\cdot,K) \\ \vdots & \ddots & \vdots \\ \mathbf{A}(I,1)\mathbf{B}(\cdot,1) & \cdots & \mathbf{A}(I,K)\mathbf{B}(\cdot,K) \end{bmatrix}. \tag{1.1}$$

2 Tensor Random Projection

We seek a random projection map to embed a collection of vectors $\mathcal{X} \subseteq \mathbb{R}^d$ into $\mathbb{R}^k$ with $k \ll d$. Let us take $d = \prod_{n=1}^N d_n$, motivated by the problem of compressing (the vectorization of) an order N tensor with dimensions $d_1, \ldots, d_N$. Conventional random projections use $O(kd)$ random variables. Generating so many random numbers is costly; and storing them can be costly when d is large. Is so much randomness truly necessary for a random projection map?

To reduce randomness and storage requirements, we propose the *tensor random projection* (TRP):

$$f_{\text{TRP}}(\mathbf{x}) := (\mathbf{A}_1 \odot \cdots \odot \mathbf{A}_N)^\top \mathbf{x}, \tag{2.1}$$

where each $\mathbf{A}_i \in \mathbb{R}^{d_i\times k}$, for $i \in [N]$, can be an arbitrary RP map and $\mathbf{A} := (\mathbf{A}_1 \odot \cdots \odot \mathbf{A}_N)^\top$. We call N the *order* of the TRP. We show in this paper that the TRP is an expected isometry, has vanishing variance, and supports database-friendly operations.

The TRP requires only $k\sum_{i=1}^N d_i$ random variables (or $k\sqrt[N]{d}$ by choosing each d_i to be equal), rather than the kd random variables needed by conventional methods. Hence the TRP is database friendly: it significantly reduces storage costs and randomness requirements compared to its constituent DRMs.

In large scale database settings, where computational efficiency is critical and queries of vector elements are costly, practitioners often use sparse RPs. Let δ be the proportion of non-zero elements in the RP map. To achieve a δ-sparse RP, a common construction is the scaled sign random map: each element is distributed as $(-1/\sqrt{\delta}, 0, 1/\sqrt{\delta})$ with probability $(\delta/2, 1-\delta, \delta/2)$. [1] proposed $\delta = 1/3$, while [14] further suggests a sparser scheme with $\delta = 1/\sqrt{d}$ that he calls the *Very Sparse* RP.

To further reduce memory requirements of random projection, we can form a TRP whose constituent submatrices are generated each with sparsity factor δ, which leads to a δ^N-sparse TRP. Under sparse setting, it is a $(1/3)^N$ sparse TRP while under very sparse setting, it is a $1/\sqrt{d}$ sparse TRP. Both TRPs can be applied to a vector using very few queries to vector elements and no multiplications. Below, we show both sparse and very sparse TRP are low-variance approximate isometry empirically.

Variance Reduction One quirk of many DRMs is that the variance of the map is controlled by the range k of the map. However, with the TRP we can reduce the variance without increasing k. We propose the TRP(T), a scaled average of T independent TRPs we call *replicates*, defined as

$$f_{\text{TRP(T)}}(\mathbf{x}) := \frac{1}{\sqrt{T}} \sum_{t=1}^{T} f_{\text{TRP}}^{(t)}(\mathbf{x}). \tag{2.2}$$

(Note that the average of T TRPs is not itself a TRP.) We discuss theoretical properties of this map in the main theory section below.

3 Main Theory

In this section, we will show the TRP and TRP(T) are expected isometries with vanishing variance. We provide a rate for the decrease in variance with k. We also prove a non-asymptotic concentration bound on the quality of the isometry when $N = 2$. Without loss of generality, we state our results only for the TRP(T), since the TRP follows as a special case with $T = 1$. We begin by showing the TRP(T) is an approximate isometry.

Lemma 3.1. *Fix $\mathbf{x} \in \mathbb{R}^{\prod_{n=1}^{N} d_n}$. Form a* TRP*(T) of order N composed of k independent matrices whose columns are independent random vectors of mean zero in isotropic positions, i.e. with identity covariance matrix. Then,*

$$\mathbb{E}\|f_{\text{TRP(T)}}(\mathbf{x})\|^2 = \|\mathbf{x}\|^2.$$

Interestingly, Lemma 3.1 does not require elements of $\mathbf{A}_n$ to be i.i.d.. Now we present an explicit form for the variance of the isometry.

Lemma 3.2. *Fix $\mathbf{x} \in \mathbb{R}^{\prod_{n=1}^{N} d_n}$. Form a* TRP*(T) of order N with range k independent matrices whose entries are i.i.d. with mean zero, variance one, and fourth moment Δ. Then*

$$\mathrm{Var}(\|f_{\text{TRP(T)}}(\mathbf{x})\|^2) = \frac{1}{Tk}(\Delta^N - 3)\|\mathbf{x}\|_4^4 + \frac{2}{k}\|\mathbf{x}\|_2^4.$$

We can see the variance increases with N. In the $N = 1$ Gaussian case, this formula shows a variance of $2/k\|\mathbf{x}\|_2^4$, which agrees with the classic result. Notice the TRP(T) only reduces the first term in the variance bound: as $T \to \infty$, the variance converges to that of a Gaussian random map. Finally we show a non-asymptotic concentration bound for $N = 2$. We leave the parallel result for $N \geqslant 3$ open for future exploration.

Theorem 3.3. *Fix $\mathbf{x} \in \mathbb{R}^{d_1 d_2}$ with sub-Gaussian norm φ_2. Form a* TRP*(T) of order 2 with range k composed of two independent matrices whose entries are drawn i.i.d. from a sub-Gaussian distribution with mean zero and variance one. Then there exists a constant C depending on φ_2 and a universal constant c_1 so that*

$$\mathbb{P}\left(\left|\|f_{\text{TRP}}(\mathbf{x})\|^2 - \|\mathbf{x}\|_2^2\right| \geqslant \epsilon\|x\|^2\right) \leqslant C\exp\left[-c_1\left(\sqrt{k}\epsilon\right)^{1/4}\right],$$

Here φ_2 is the sub-Gaussian norm defined in Definition D.1 in Appendix A. Theorem 3.3 shows that for a TRP to form an ϵ-JL DRM with substantial probability on a dataset with n points, our method requires $k = \mathcal{O}(\epsilon^{-2}\log^8 n)$ while conventional random projections require $k = \mathcal{O}(\epsilon^{-2}\log n)$. Numerical experiments suggest this bound is pessimistic.

	Gaussian	Sparse	Very Sparse
RP	0.1198 (0.0147)	0.1198 (0.0150)	0.1189 (0.0108)
TRP	0.1540 (0.0290)	0.1609 (0.0335)	0.1662 (0.0307)
TRP(5)	0.1262 (0.0166)	0.1264 (0.0194)	0.1276 (0.0164)

Table 1: RMSE for the estimate of the pairwise inner product of the MNIST data, where standard error is in the parentheses.

4 Experiment

In this section, we compare the quality of the isometry of conventional RPs, TRP, and TRP(5), for Gaussian, Sparse [1], and Very Sparse random maps [14] on both synthetic data and MNIST data. We also use TRP and TRP(5) to compute pairwise cosine similarity (Table 1 and Appendix B) and to sketch matrices and tensors (Appendix C), although the theory still remains open.

Our first experiment evaluates the quality of the isometry for maps $\mathbb{R}^d \to \mathbb{R}^k$. We generate $n = 10$ independent vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ of sizes $d = 2500, 10000, 40000$ from $\mathcal{N}(\mathbf{0}, \mathbf{I})$. We consider the following three RPs: 1. Gaussian RP; 2. Sparse RP [1]; 3. Very Sparse RP [14]. For each, we compare the performance of RP, TRP, and TRP(5) with order 2 and $d_1 = d_2$. We evaluate the methods by repeatedly generating a RP and computing the reduced vector, and plot the ratio of the pairwise distance $\frac{1}{n(n-1)} \sum_{n \geqslant i \neq j \geqslant 1} \frac{\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\|_2}{\|\mathbf{x}_i - \mathbf{x}_j\|_2}$ and the average standard deviation for different k averaged over 100 replications. In the MNIST example, we choose the first $n = 50$ vectors of size $d = 784$, normalize them, and perform the same experiment. Figure 1 shows results on simulated ($d = 2500$) and MNIST data for the Gaussian and Very Sparse RP. See Section B for additional experiments.

These experiments show that to preserve pairwise distance and cosine similarity, TRP performs nearly as well as RP for all three types of maps. With only five replicates, TRP(5) reduces the variance significantly in real data while not much in the simulation setting. The difference in accuracy between methods diminishes as k increases. When $d = d_1 d_2 = 40000$, the storage for TRP(5) is still $\frac{1}{20}$ of the Gaussian RP. The variance reduction is effective especially in sparse and very sparse setting.

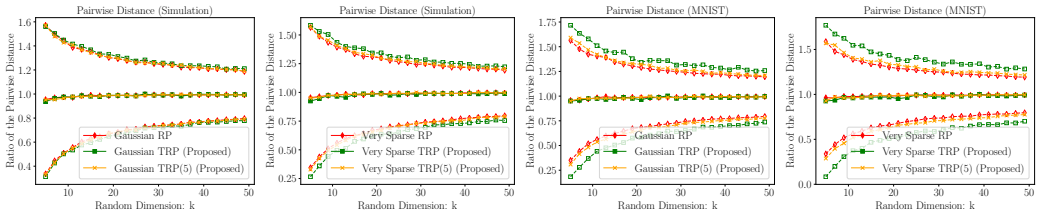

Figure 1: Isometry quality for simulated and MNIST data. The left two plots show results for Gaussian and Very Sparse RP, TRP, TRP(5) respectively applied to $n = 20$ standard normal data vectors in $\mathbb{R}^{2500}$. The right two plots show the same for 50 MNIST image vectors in $\mathbb{R}^{784}$. The dashed line shows the error two standard deviations from the average ratio.

5 Conclusion

The TRP is a novel dimension reduction map composed of smaller DRMs. Compared to its constituent DRMs, it significantly reduces the requirements for randomness and for storage. Numerically, the variance-reduced TRP(5) method with only five replicates achieves accuracy comparable to the conventional RPs for $1/20$ of the original storage. We prove the TRP and TRP(T) are expected isometries with vanishing variance, and provide a non-asymptotic error bound for the order 2 TRP.

For the future work, we will provide a general non-asymptotic bound for the higher order TRP and develop the theory relevant for the application of the TRP in sketching low-rank approximation, given its practical effectiveness (shown in Appendix C).

References

[1] Dimitris Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of computer and System Sciences*, 66(4):671–687, 2003.

[2] Zeyuan Allen-Zhu, Rati Gelashvili, Silvio Micali, and Nir Shavit. Sparse sign-consistent johnson–lindenstrauss matrices: Compression with neuroscience-based constraints. *Proceedings of the National Academy of Sciences*, 111(47):16872–16876, 2014.

[3] Alberto Barrón-Cedeño and Paolo Rosso. On automatic plagiarism detection based on n-grams comparison. In *European Conference on Information Retrieval*, pages 696–700. Springer, 2009.

[4] Casey Battaglino, Grey Ballard, and Tamara G Kolda. A practical randomized cp tensor decomposition. *SIAM Journal on Matrix Analysis and Applications*, 39(2):876–901, 2018.

[5] Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250. ACM, 2001.

[6] Jean Bourgain, Sjoerd Dirksen, and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. *Geometric and Functional Analysis*, 25(4):1009–1088, 2015.

[7] Jeremy Buhler and Martin Tompa. Finding motifs using random projections. *Journal of computational biology*, 9(2):225–242, 2002.

[8] Lieven De Lathauwer, Bart De Moor, and Joos Vandewalle. A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000.

[9] Dmitriy Fradkin and David Madigan. Experiments with random projections for machine learning. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 517–522. ACM, 2003.

[10] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[11] William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

[12] Samuel Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Neural networks proceedings, 1998. ieee world congress on computational intelligence. the 1998 ieee international joint conference on*, volume 1, pages 413–418. IEEE, 1998.

[13] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[14] Ping Li, Trevor J Hastie, and Kenneth W Church. Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 287–296. ACM, 2006.

[15] Kun Liu, Hillol Kargupta, and Jessica Ryan. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE Transactions on knowledge and Data Engineering*, 18(1):92–106, 2006.

[16] Makoto Matsumoto and Takuji Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(1):3–30, 1998.

[17] Christos H Papadimitriou, Prabhakar Raghavan, Hisao Tamaki, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 61(2):217–235, 2000.

[18] Mark Rudelson. Row products of random matrices. *Advances in Mathematics*, 231(6):3199–3231, 2012.

[19] Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[20] Ozgur D Sahin, Aziz Gulbeden, Fatih Emekçi, Divyakant Agrawal, and Amr El Abbadi. Prism: indexing multi-dimensional data in p2p networks using reference vectors. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 946–955. ACM, 2005.

[21] Yining Wang, Hsiao-Yu Tung, Alexander J Smola, and Anima Anandkumar. Fast and guaranteed tensor decomposition via sketching. In *Advances in Neural Information Processing Systems*, pages 991–999, 2015.

[22] David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[23] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2009.

Appendix A Proof for Main Theorems

Before presenting the proof for the main theory, we first define some new notations. Since these notations will only be used in technical proofs, we do not include them in the main body.

Notations for Technical Proofs

For a vector $\mathbf{x}$ with length $\prod_{n=1}^{N} d_n$, for simplicity, we introduce the multi-index for it: let $\mathbf{x}_{r_1,\cdots,r_N}, \forall r_n \in [d_n]$, represents the $(1+\sum_{n=1}^{N}(r_n-1)s_n)^{th}$ element, where $s_n = \prod_{n+1}^{N} d_n$, for $n < N$ and 1 for $n = N$. For vector $\mathbf{r}_1$, $\mathbf{r}_2$, we say $\mathbf{r}_1 = \mathbf{r}_2$ if and only if all their elements are the same.

Also, we let $\mathbf{vec}(\mathbf{A})$ be the vectorization operator for any matrix $\mathbf{A} \in \mathbb{R}^{d\times k}$, which stacks all columns of matrix $\mathbf{A}$ and returns a vector of length kd, $[\mathbf{A}(\cdot,1);\cdots;\mathbf{A}(\cdot,k);]$. Here we use semi-colon to denote the vertical stack of vectors $\mathbf{x}$ and $\mathbf{y}$ as $[\mathbf{x};\mathbf{y}]$. As comparison, we use comma to mean stack row vector horizontally like $[\mathbf{x}^\top,\mathbf{y}^\top]$.

Proof for Lemma 3.1

Proof. We first give a sufficient condition for general random matrix to let (3.1) be held, then we show that Khatri-Rao map with condition in Lemma 3.1 satisfies these two general sufficient conditions.

Consider a general random matrix $\mathbf{A} \in \mathbb{R}^{k\times d}$ and $\mathbf{x} \in \mathbb{R}^d$. we claim if $\mathbb{E}\mathbf{A}^2(r,s) = 1, \forall r,s$ and $\mathbb{E}\mathbf{A}(r,s_1)\mathbf{A}(r,s_2) = 0, \forall r \in [k], s_1 \neq s_2 \in [d]$, then $\mathbb{E}\|\frac{1}{\sqrt{k}}\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2$, when $\mathbf{y} = \mathbf{A}\mathbf{x}$. To see why, it suffices to show that $\mathbb{E}y_r^2 = \|x\|_2^2$.

$$\begin{aligned}\mathbb{E}y_r^2 &= \mathbb{E}\sum_{s_1=1}^{d}\sum_{s_2=1}^{d}\mathbf{A}(r,s_1)\mathbf{A}(r,s_2)x_{s_1}x_{s_2}\\ &= \sum_{s=1}^{d}\mathbf{A}^2(r,s)x_s^2 = \|\mathbf{x}\|_2^2,\end{aligned}$$

where the first equation in the second line comes from the fact that $\mathbb{E}\mathbf{A}(r,s_1)\mathbf{A}(r,s_2) = 0$ for $s_1 \neq s_2$ and the second equation in the second line comes from that $\mathbb{E}\mathbf{A}^2(r,s) = 1$.

Then, we will prove Lemma 3.1 by induction. We first show that for two matrices $\mathbf{B}_1 \in \mathbb{R}^{d_1\times k}, \mathbf{B}_2 \in \mathbb{R}^{d_2\times k}$ whose entries satisfy the two conditions in lemma 3.1: $\mathbb{E}\mathbf{B}_n^2(r,s) = 1$ and $\mathbb{E}[\mathbf{B}_n(r_1,s)\mathbf{B}_n(r_2,s)] = 0$ for $n = 1,2, s \in [d], r, r_1 \neq r_2 \in [d_n]$, we have $\mathbf{A} = (\mathbf{B}_1 \odot \mathbf{B}_2)^\top$ satifies the two sufficient conditions stated previously. It suffices to restrict our focus to the first row of $\mathbf{\Omega}$ and we apply the multi-index to it. For any $1 \leqslant r_1 \leqslant d_1, 1 \leqslant r_2 \leqslant d_2$,

$$\begin{aligned}&\mathbb{E}\mathbf{A}_1^2(k_1,k_2) = \mathbb{E}\mathbf{B}_1^2(k_1,1)\mathbf{B}_2^2(k_2,1)\\ &= \mathbb{E}\mathbf{B}_1^2(k_1,1)\mathbb{E}\mathbf{B}_2^2(k_1,1) = 1. \text{ (independence bewteen } \mathbf{B}_i, i = 1,2)\end{aligned}$$

To avoid confusion in notation, we argue that $\mathbf{A}(1,\cdot)$ is the first row vector of $\mathbf{A}$ of size d_1d_2, and we apply the multi-index to it. Also, for two different elements in the first row of $\mathbf{A}$: $\mathbf{A}_1(k_1,k_2)\mathbf{A}_1(s_1,s_2)$ at least one of $k_1 \neq s_1$, $k_2 \neq s_2$ hold. Without losing generality, assuming $k_1 \neq s_1$,

$$\begin{aligned}&\mathbb{E}\mathbf{A}_1(k_1,k_2)\mathbf{A}_1(s_1,s_2) = \mathbb{E}\mathbf{B}_1(k_1,1)\mathbf{B}_2(k_2,1)\mathbf{B}_1(s_1,1)\mathbf{B}_2(s_2,1)\\ &= \mathbb{E}\mathbb{E}\left[\mathbf{B}_1(k_1,1)\mathbf{B}_1(k_2,1)\mathbf{B}_2(k_2,1)\mathbf{B}_2(s_2,1) \mid \mathbf{B}_2(k_2,1)\mathbf{B}_2(s_2,1)\right]\\ &= \mathbb{E}\mathbf{B}_2(k_2,1)\mathbf{B}_2(s_2,1)\mathbb{E}\left[\mathbf{B}_1(k_1,1)\mathbf{B}_1(s_1,1)\right] = 0,\end{aligned}$$

where we use the fact that entries within/across B_i are independent with each other and have zero expectation.

Notice that two conditions for $\mathbf{A} = (\mathbf{B}_1 \odot \mathbf{B}_2)^\top$ directly show that $\mathbf{B}_1 \odot \mathbf{B}_2$ satisfies two conditions in Lemma 3.1, we could use a standard mathematical induction argument to finish the proof for f_{TRP}.

For TRP(T),

$$\mathbb{E}\|f_{\text{TRP(T)}}(\mathbf{x})\|_2^2 = \frac{1}{T}\mathbb{E}\|\sum_{t=1}^{T} f_{\text{TRP}}^{(t)}(\mathbf{x})\|_2^2$$

$$= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|f_{\text{TRP}}^{(t)}(\mathbf{x})\|_2^2 = \|x\|_2^2,$$

where in the second line we use the fact that each $f_{\text{TRP}}^{(t)}$ is independent with each other. □

Next we introduce a lemma which shows that by bounding the deviation for the norm square of each vector, we could also bound the deviation for inner product. Although it is commonly known in any random projection literature, for completeness, we still list the lemma with proof here.

Lemma A.1. *For a linear mapping from $\mathbb{R}^d \to \mathbb{R}^k$: $f(\mathbf{x}) = \frac{1}{\sqrt{k}}\mathbf{\Omega}\mathbf{x}$,*

$$\mathbb{P}(|\langle f(\mathbf{x}), f(\mathbf{y})\rangle - \langle \mathbf{x}, \mathbf{y}\rangle| \geqslant \epsilon|\langle \mathbf{x}, \mathbf{y}\rangle|) \leqslant 2\sup_{\mathbf{x}\in\mathbb{R}^d}\mathbb{P}(|\|f(\mathbf{x})\|^2 - \|\mathbf{x}\|^2| \geqslant \epsilon\|\mathbf{x}\|_2^2).$$

Proof. Since f is a linear mapping, we have

$$4f(\mathbf{x})f(\mathbf{y}) = \|f(\mathbf{x}+\mathbf{y})\|_2^2 - \|f(\mathbf{x}-\mathbf{y})\|_2^2.$$

Consider the event

$$\mathcal{A}_1 = \left\{|\|f(\mathbf{x}+\mathbf{y})\|_2^2 - \|\mathbf{x}+\mathbf{y}\|_2^2| \geqslant \epsilon\|\mathbf{x}+\mathbf{y}\|_2^2\right\}$$
$$\mathcal{A}_2 = \left\{|\|f(\mathbf{x}-\mathbf{y})\|_2^2 - \|\mathbf{x}-\mathbf{y}\|_2^2| \geqslant \epsilon\|\mathbf{x}-\mathbf{y}\|_2^2\right\}$$

On the event $\mathcal{A}_1^{\complement} \cap \mathcal{A}_2^{\complement}$,

$$4f(\mathbf{x})f(\mathbf{y}) \geqslant (1-\epsilon)(\mathbf{x}+\mathbf{y})^2 - (1+\epsilon)(\mathbf{x}-\mathbf{y})^2 = 4\langle \mathbf{x}, \mathbf{y}\rangle - 2\epsilon(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2),$$

noticing $\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 \geqslant 2\langle \mathbf{x}, \mathbf{y}\rangle$, and by similar argument on the other side of the inequality, we could claim that

$$\{|\langle f(\mathbf{x}), f(\mathbf{y})\rangle - \langle \mathbf{x}, \mathbf{y}\rangle| \geqslant \epsilon|\langle \mathbf{x}, \mathbf{y}\rangle|\} \subseteq \mathcal{A}_1 \cup \mathcal{A}_2.$$

Then we finish the proof by simply applying an union bound of two events. □

Remark. *The key element of classic random projections is the dimension-free bound. Similarly, according to Prop. 3.3, our TRP has a norm preservation bound independent of the particular vector $\mathbf{x}$ and dimension d and thus a dimension-free inner product preservation bound according to Lemma A.1.*

Proof for Lemma 3.2

Proof. Let $\mathbf{y} = \mathbf{A}\mathbf{x}$. We know from Lemma 3.1 that $\mathbb{E}\|f_{\text{TRP}}(\mathbf{x})\|_2^2 = \frac{1}{k}\mathbb{E}\|\mathbf{A}\mathbf{x}\|^2 = \|\mathbf{x}\|_2^2$. Notice

$$\mathbb{E}(\|f_{\text{TRP(T)}}(\mathbf{x})\|_2^2) = \|\mathbf{x}\|_2^2,$$

and $\mathbb{E}y_1^2 = \|x\|_2^2$ as shown in the poof of Lemma 3.1. It is easy to see that

$$\mathbb{E}\|\mathbf{y}\|_2^4 = \sum_{i=1}^{k}\mathbb{E}y_i^4 + \sum_{i\neq j}\mathbb{E}y_i^2 y_j^2.$$

Again, as shown in Lemma 3.1, $\mathbb{E}y_i^2y_j^2 = \mathbb{E}y_i^2\mathbb{E}y_j^2 = \|\mathbf{x}\|^4$. To find $\mathbb{E}\|\mathbf{y}\|_2^4$, it suffices to find $\mathbb{E}y_1^4$ by noticing that y_i are i.i.d. random variables. Let Ω be the set containing all corresponding multi-index vector for $\{1, \cdots, \prod_{n=1}^{N} d_n\}$.

$$y_1^4 = \left[\sum_{\mathbf{r}\in\Omega}\mathbf{A}(1,\mathbf{r})x_{\mathbf{r}}\right]^4 = \sum_{\mathbf{r}\in\Omega}\mathbf{A}^4(1,\mathbf{r})x_{\mathbf{r}}^4 + 3\sum_{\mathbf{r}_1\neq\mathbf{r}_2\in\Omega}\mathbf{A}^2(1,\mathbf{r}_1)x_{\mathbf{r}_1}^2\mathbf{A}^2(1,\mathbf{r}_2)x_{\mathbf{r}_2}^2$$
$$+ 6\sum_{\mathbf{r}_1\neq\mathbf{r}_2\neq\mathbf{r}_3\in\Omega}\mathbf{A}^2(1,\mathbf{r}_1)x_{\mathbf{r}_1}\mathbf{A}(2,\mathbf{r}_2)x_{\mathbf{r}_2}\mathbf{A}(3,\mathbf{r}_3)x_{\mathbf{r}_3} + 4\sum_{\mathbf{r}_1\neq\mathbf{r}_2\in\Omega}\mathbf{A}^3(1,\mathbf{r}_1)x_{\mathbf{r}_1}^3\mathbf{A}(1,\mathbf{r}_2)x_{\mathbf{r}_2}$$
$$+ \sum_{\mathbf{r}_1\neq\mathbf{r}_2\neq\mathbf{r}_3\neq\mathbf{r}_4\in\Omega}\mathbf{A}(1,\mathbf{r}_1)x_{\mathbf{r}_1}\mathbf{A}(1,\mathbf{r}_2)x_{\mathbf{r}_2}\mathbf{A}(1,\mathbf{r}_3)x_{\mathbf{r}_3}\mathbf{A}(1,\mathbf{r}_4)x_{\mathbf{r}_4}.$$

It is not hard to see that except for the first line, the expectation of second and third line is zero.

$$\mathbb{E}\mathbf{A}^4(1,\mathbf{r}) = \mathbb{E}\mathbf{A}_1^4(1,r_1)\cdots\mathbf{A}_N^4(1,r_N) = \Delta^N.$$

Also with proof in Lemma 3.1,

$$\mathbb{E}\mathbf{A}^2(1,\mathbf{r}_1)\mathbf{A}^2(1,\mathbf{r}_2) = \mathbb{E}\mathbf{A}^2(1,\mathbf{r}_1)\mathbb{E}\mathbf{A}^2(1,\mathbf{r}_2) = 1.$$

Combining these two together, we have

$$\begin{aligned}\mathbb{E}\|f_{\text{TRP}}(\mathbf{x})\|^4 &= \frac{1}{k^2}\left[k(\Delta^N-3)\|\mathbf{x}\|_4^4 + 3k\|\mathbf{x}\|_2^4 + (k-1)k\|\mathbf{x}\|_2^4\right]\\ &= \frac{1}{k}\left[(\Delta^N-3)\|\mathbf{x}\|_4^4 + 2\|\mathbf{x}\|_2^4\right] + \|\mathbf{x}\|_2^4.\end{aligned}$$

Therefore,

$$\text{Var}(\|f_{\text{TRP}}(\mathbf{x})\|_2^2) = \mathbb{E}\|f_{\text{TRP}}(\mathbf{x})\|_2^4 - (\mathbb{E}\|f_{\text{TRP}}(\mathbf{x})\|_2^2)^2 = \frac{1}{k}\left[(\Delta^N-3)\|\mathbf{x}\|_4^4 + 2\|\mathbf{x}\|_2^4\right].$$

Now we switch to see how much variance could be reduced by the variance reduction method. With Lemma 3.1, we already know that $\mathbb{E}\|f_{\text{TRP(T)}}(\mathbf{x})\|_2^2 = \|\mathbf{x}\|_2^2$. The rest is to calculate $\mathbb{E}\|f_{\text{TRP(T)}}(\mathbf{x})\|_2^4$ out.

$$\begin{aligned}\|f_{\text{TRP(T)}}(\mathbf{x})\|_2^4 &= \frac{1}{T^2}\left[\sum_{t=1}^T \|f_{\text{TRP}}^{(t)}(\mathbf{x})\|_2^2 + \sum_{t_1\neq t_2}\langle f_{\text{TRP}}^{(t_1)}(\mathbf{x}), f_{\text{TRP}}^{(t_2)}(\mathbf{x})\rangle\right]^2\\ &= \frac{1}{T^2}\left[\sum_{t=1}^T \|f_{\text{TRP}}^{(t)}(\mathbf{x})\|_2^4 + \sum_{t_1\neq t_2}\|f_{\text{TRP}}^{(t_1)}(\mathbf{x})\|_2^2\|f_{\text{TRP}}^{(t_2)}(\mathbf{x})\|_2^2 + 2\sum_{t_1\neq t_2}\langle f_{\text{TRP}}^{(t_1)}(\mathbf{x}), f_{\text{TRP}}^{(t_2)}(\mathbf{x})\rangle^2 + \text{rest}\right].\end{aligned}$$

It is not hard to show that $\mathbb{E}(\text{rest}) = 0$. Following the definition of $\mathbf{y}$,

$$\mathbb{E}\|f_{\text{TRP}}^{(t_1)}(\mathbf{x})\|_2^2\|f_{\text{TRP}}^{(t_2)}(\mathbf{x})\|_2^2 = \|\mathbf{x}\|_2^4,$$

and

$$\begin{aligned}&\mathbb{E}\langle f_{\text{TRP}}^{(t_1)}(\mathbf{x}), f_{\text{TRP}}^{(t_2)}(\mathbf{x})\rangle^2\\ &= \frac{1}{k^2}\mathbb{E}\left[\sum_{i=1}^k y_i^{(t_1)}y_i^{(t_2)}\right]^2 = \frac{1}{k}\mathbb{E}[y_1^{(t_1)}y_1^{(t_2)}]^2 = \frac{1}{k}\|x\|_2^4.\end{aligned}$$

Combining all these together, we could show that

$$\begin{aligned}\text{Var}(\|f_{\text{TRP(T)}}(\mathbf{x})\|_2^2) &= \mathbb{E}\|f_{\text{TRP(T)}}(\mathbf{x})\|_2^4 - (\mathbb{E}\|f_{\text{TRP(T)}}(\mathbf{x})\|_2^2)^2\\ &= \frac{1}{T^2}\left[\frac{T}{k}\left[(\Delta^N-3)\|\mathbf{x}\|_4^4 + 2\|\mathbf{x}\|_2^4\right]\right.\\ &\quad \left. + T(T-1)\|\mathbf{x}\|_2^4 + T\|\mathbf{x}\|_2^4 + \frac{2T(T-1)}{k}\|\mathbf{x}\|_2^4\right] - \|\mathbf{x}\|_2^4\\ &= \frac{1}{Tk}(\Delta^N-3)\|\mathbf{x}\|_4^4 + \frac{2}{k}\|\mathbf{x}\|_2^4.\end{aligned}$$

□

Definition A.1. *A random variable x is said to satisfy the generalized-sub-exponential moment condition with constant α, if for general positive integer k, there exists a general constant C(not depending on k), s.t.*

$$\mathbb{E}|x|^k \leqslant (Ck)^{k\alpha} \tag{A.1}$$

Proof for Theorem 3.3

Proof. From now on, with losing generality, we will assume $\|x\| = 1$. Let

$$\mathbf{y} = \frac{1}{\sqrt{k}}(\mathbf{A}_1 \odot \mathbf{A}_2)^\top \mathbf{x},$$

Lemma 3.1 asserts that $\mathbb{E}\|\mathbf{y}\|_2^2 = \|\mathbf{x}\|_2^2$ (conditions in lemma 3.1 naturally hold for i.i.d. random variables in our setting). The key observation is that $y_i, i \in [k]$ is quadratic form of elements of $\mathbf{A}_i, i = 1, 2$. Then as quadratic form of sub-Gaussian variables, y_i are identically independently distributed generalized sub-exponential random variable. Then we could use Hanson-Wright inequality to determine the constants in moments condition A.1 which shall present tighter bound compared to directly citing results of linear combination of sub-exponential random variable defined in (A.1)

We aim to write y_i as a quadratic form of $\mathbf{z}_i := [\mathbf{vec}(\mathbf{A}_1(\cdot, i)); \mathbf{vec}(\mathbf{A}_2(\cdot, i))]$. Also, for convenience, we partition $\mathbf{x}$ into d_1 sub-vectors with equal length d_2 i.e., $\mathbf{x} = [\mathbf{x}_1; \cdots; \mathbf{x}_{d_1}]$. To make it clear, we consider writing y_1 as quadratic form of $\mathbf{z}_1$ first.

$$y_1 = \langle [\mathbf{A}_1(1,1)\mathbf{A}_2(\cdot,1); \cdots; \mathbf{A}_1(d_1,1)\mathbf{A}_2(\cdot,1)], [\mathbf{x}_1; \cdots; \mathbf{x}_{d_1}] \rangle$$

which indicates that we could write

$$y_1 = \mathbf{z}_1^\top \mathbf{M} \mathbf{z}_1,$$

where

$$\mathbf{M} = \begin{bmatrix} \mathbf{0} & \mathbf{D} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \quad \mathbf{D} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_{d_1}^\top. \end{bmatrix}$$

It is easy to see that $\|\mathbf{M}\| \leqslant \|\mathbf{D}\| \leqslant \|\mathbf{D}\|_F = \|\mathbf{M}\|_F = 1$ by assuming $\|\mathbf{x}\| = 1$. Then applying the Hanson Wright inequality in Lemma D.1, we could have for any positive number η, there exists a general constant c_1 s.t.

$$\begin{aligned} \mathbb{P}(|y_i| \geqslant \eta) &\leqslant 2\exp\left[-c_1 \min\left\{-\frac{\eta}{\varphi_2^2\|M\|}, \frac{\eta^2}{\varphi_2^4\|M\|_F^2}\right\}\right] \\ &\leqslant 2\exp\left[-c_1 \min\left\{-\frac{\eta}{\varphi_2^2}, \frac{\eta^2}{\varphi_2^4}\right\}\right]. \end{aligned}$$

Then by Lemma D.2, we could find a constant C depending on sub-Gaussian norm and general constant c_1 s.t.

$$\mathbb{E}|y_i|^k \leqslant (Ck)^k,$$

where in fact we could give the explicit form of C as

$$C = 1 + \frac{c_1}{\min\{\varphi_2^2, \varphi_2^4\}}. \tag{A.2}$$

Notice y_i has mean zero and variance 1 (assuming $\|x\| = 1$), then apply Lemma D.3, we could assert that there exists a general constant c_2

$$\mathbb{P}\left(\left|\frac{1}{k}\mathbf{y}^\top \mathbf{I}_{k,k}\mathbf{y} - 1\right| \geqslant \epsilon\right) \leqslant C\exp\left(-c_2\left[\sqrt{k}\epsilon\right]^{1/4}\right),$$

where C is defined in (A.2) and we use the fact $\alpha = 1$ in our case which is defined in moments condition.

□

Appendix B More Simulation Results

Pairwise Distance Estimation In Figure 2, 3, 4, we compare the performance of Gaussian, Sparse, Very Sparse random maps on the pairwise distance estimation problem with $d = 2500, 10000, 40000$, $N = 2$. Additionally, we compare their performance for $d = 125000$, $N = 3$ in Figure 5.

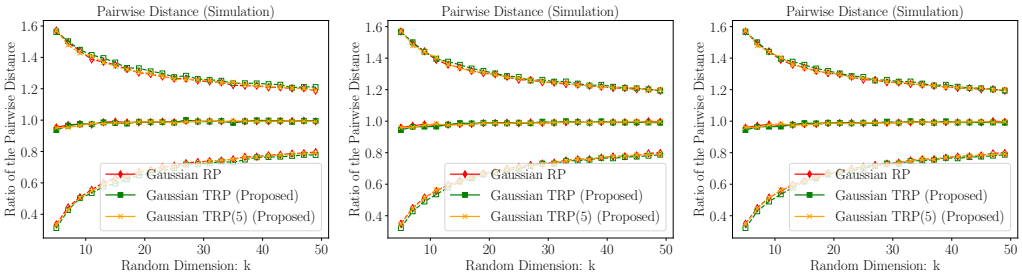

Figure 2: Average ratio of the pairwise distance for simulation data using Gaussian RP: *The plots correspond to the simulation for Gaussian RP, TRP, TRP(5) respectively with* $n = 20, d = 2500, 10000, 40000$ *and each data vector comes from* $N(\mathbf{0}, \mathbf{I})$. *The dashed line represents the error bar 2 standard deviation away from the average ratio.*

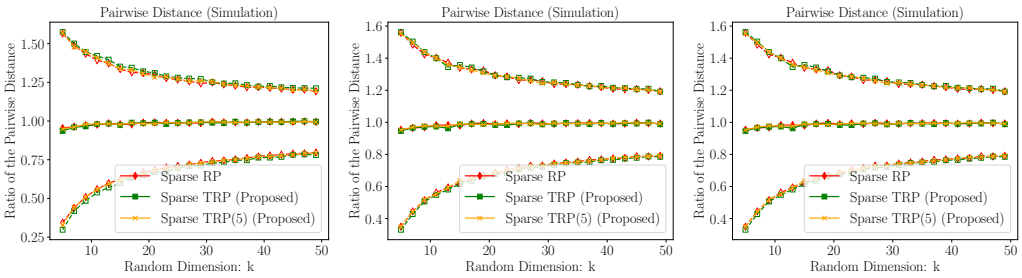

Figure 3: Average ratio of the pairwise distance for simulation data using Sparse RP: *The plots correspond to the simulation for Sparse RP, TRP, TRP(5) respectively with* $n = 20, d = 2500, 10000, 40000$ *and each data vector comes from* $N(\mathbf{0}, \mathbf{I})$. *The dashed line represents the error bar 2 standard deviation away from the average ratio.*

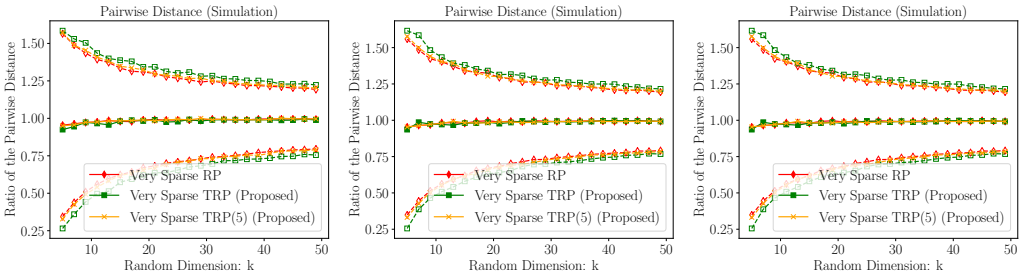

Figure 4: Average ratio of the pairwise distance for simulation data using Very Sparse RP: *The plots correspond to the simulation for Very Sparse RP, TRP, TRP(5) respectively with* $n = 20, d = 2500, 10000, 40000$ *and each data vector comes from* $N(\mathbf{0}, \mathbf{I})$. *The dashed line represents the error bar 2 standard deviation away from the average ratio.*

Pairwise Cosine Similarity Estimation The second experiment is to estimate the pairwise cosine similarity, i.e. $\frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2}$ for $\mathbf{x}_i, \mathbf{x}_j$. We use both the simulation data ($d = 10000$) and the MNIST data ($d = 784, n = 60000$). We experiment with Gaussian, Sparse, Very Sparse RP, TRP, and TRP(5)

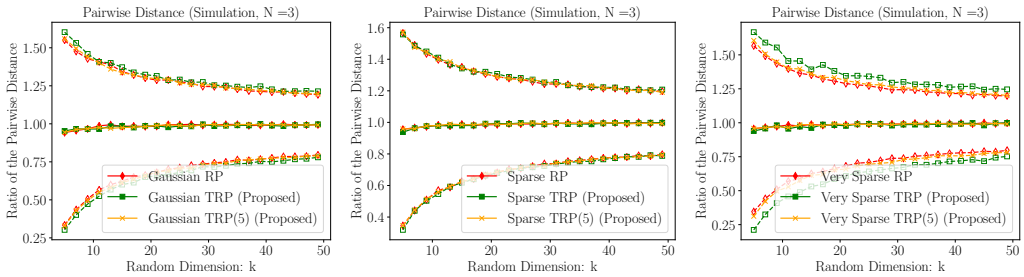

Figure 5: Average ratio of the pairwise distance for simulation data using: *The plots correspond to the simulation for Gaussian, Sparase, Very Sparse RP, TRP, TRP(5) respectively with* $n = 20, d = d_1 d_2 d_3 = 50 \times 50 \times 50 = 125000$ *and each data vector comes from* $N(\mathbf{0}, \mathbf{I})$. *The dashed line represents the error bar 2 standard deviation away from the average ratio.*

with the same setting as above ($k = 50$). We evaluate the performance by the average root mean square error (RMSE). The results is given in Table 1, 2.

	Gaussian	Sparse	Very Sparse
RP	0.1409 (0.0015)	0.1407 (0.0013)	0.1412 (0.0014)
TRP	0.1431 (0.0016)	0.1431 (0.0015)	0.1520 (0.0033)
TRP(5)	0.1412 (0.0012)	0.1411 (0.0015)	0.1427 (0.0014)

Table 2: RMSE for the estimate of the pairwise inner product of the simulation data ($d = 10000, k = 50, n = 100$), where standard error is in the parentheses.

Appendix C Application: Sketching

Beyond random projection, our novel TRP also has an important application in sketching. Sketching is an important technique to accelerate expensive computations with widespread applications, such as regression, low-rank approximation, and graph sparsification, etc. [10, 22] The core idea behind sketching is to compress a large dataset, typically a matrix or tensor, into a smaller one by multiplying a random matrix. In this section, we will mainly focus on the low-rank matrix approximation problem. Consider a matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ with rank r, we want to find the best rank-r approximation with the minimal amount of time. The most common method is the randomized singular value decomposition (SVD), whose underlying idea is sketching.

First, we compute the linear sketch $\mathbf{Z} \in \mathbb{R}^{m \times k}$ by $\mathbf{Z} = \mathbf{X}\mathbf{\Omega}$, where $\mathbf{\Omega} \in \mathbb{R}^{d \times r}$ is the random map. Then we compute the QR decomposition of $\mathbf{X}\mathbf{\Omega}$ by $\mathbf{Q}\mathbf{R} = \mathbf{Z}$, where $\mathbf{Q} \in \mathbb{R}^{m \times k}, \mathbf{R} \in \mathbb{R}^{r \times r}$. At the end, we project $\mathbf{X}$ onto the column space of $\mathbf{Q}$, and obtain the approximation $\hat{\mathbf{X}} = \mathbf{Q}\mathbf{Q}^\top \mathbf{X}$.

With our TRP, we can significantly reduce the storage of the random map, while achieving similar rate of convergence as demonstrated in Figure 6. With further variance reduction by taking the geometric-median over multiple runs, our TRP with variance reduction can achieve even better performance. The detailed implementation is given in Algorithm 1. And we will delay the theoretical analysis of this method for future works.

Algorithm 1 Tensor Sketching with Variance Reduction

Input: $\mathbf{X} \in \mathbb{R}^{m\times d}$, where $d = \prod_{i=1}^{N} d_n$ and RMAP is a user-specified function that generates a random dimension reduction map. T is the number of runs for variance reduction averaging.

function $\mathrm{SSVR}(\mathbf{X}, \{d_n\}, k, T, \mathrm{RMAP})$
  for $t = 1 \dots T$ **do**
    for $i = 1 \dots N$ **do** $\mathbf{\Omega}_i^{(t)} = \mathrm{RMAP}(\mathrm{d_i}, \mathrm{k})$
    end for
    $\mathbf{\Omega}^{(t)} = \mathbf{\Omega}_1^{(t)} \odot \cdots \odot \mathbf{\Omega}_N^{(t)}$
    $(\mathbf{Q}^{(t)}, \sim) = \mathrm{QR}(\mathbf{X}\mathbf{\Omega}^{(\mathrm{t})})$
    $\hat{\mathbf{X}}^{(t)} = \mathbf{Q}^{(t)}\mathbf{Q}^{(t)T}\mathbf{X}$
  end for
  $\hat{\mathbf{X}} = \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{X}}^{(t)}$
  return $\mathbf{G}$
end function

Furthermore, the extension of TRP to tensor data is also natural. To be specific, the n^{th} unfolding of a large tensor $\mathcal{X} \in \mathbb{R}^{I_1\times\cdots\times I_N}$, denoted as $\mathbf{X}^{(n)}$, has dimension $I_n \times I_{(-n)}$, where $I_{(-n)} = \prod_{i\neq n, i\in[N]} I_i$. To construct a sketch for the unfolding, we need to create a random matrix of size $I_{(-n)} \times k$. Then, our TRP becomes a natural choice to avoid the otherwise extremely expensive storage cost. For many popular tensor approximation algorithms, it is even necessary to perform sketching for every dimension of the tensor [8, 21]. In the simulation section, we perform experiments for the unfolding of the higher-order order tensor with our structured sketching algorithms (Figure 6). For more details in tensor algebra, please refer to [13].

Experimental Setup In sketching problems, considering a N-D tensor $\mathcal{X} \in \mathbb{R}^{I^N}$ with equal length along all dimensions, we want to compare the performance of the low rank approximation with different maps for its first unfolding $\mathbf{X}^{(1)} \in \mathbb{R}^{I\times I^{N-1}}$.

We construct the tensor $\mathcal{X}$ in the following way. Generate a core tensor $\mathcal{C} \in \mathbb{R}^{r^N}$, with each entry $\mathrm{Unif}([0, 1])$. Independently generate N orthogonal arm matrices by first creating $\mathbf{A}_1, \dots, \mathbf{A}_N \in \mathbb{R}^{r\times I}$ and then computing the arm matrices by $(\mathbf{Q}_n, \sim) = \mathrm{QR}(\mathbf{A}_\mathrm{n})$, for $1 \leqslant n \leqslant N$.

$$\mathcal{X} = \mathcal{C} \times_1 \mathbf{Q}_1 \cdots \times_N \mathbf{Q}_N + \sqrt{\frac{0.01 \cdot \|\mathcal{X}^\natural\|_F^2}{I^N}}\mathcal{N}(0, 1).$$

Then, we construct the mode-1 unfolding of $\mathbf{X} = \mathbf{X}^{(1)}$, which has a rank smaller than or equal to r.

In our simulation, we consider the scenarios of 2-D (900×900), 3-D ($400 \times 400 \times 400$), 4-D ($100 \times 100 \times 100 \times 100$) tensor data, with corresponding mode-1 unfolding of size 900×900, 400×160000, 100×1000000 respectively and $r = 5$. In each scenario, we compare the performance for Gaussian RP, TRP, and TRP(5) maps with varying k from 5 to 25. The TRP map in these scenarios has 2, 4, 6 components of size $30 \times k$, $20 \times k$, $10 \times k$ respectively. And the number of runs variance reduction averaging is $T = 5$. In the end, we evaluate the performance by generating the random matrix 100 times and compute the relative error $\frac{\|\mathbf{X}-\hat{\mathbf{X}}\|}{\|\mathbf{X}\|}$, and constructing a 95% confidence interval for it.

Result From Figure 6, we can observe that the relative error decreases as k increases as expected for all dimension reduction maps. The difference of the performance between the Khatri-Rao map and Gaussian map is small when $N = 2$, but increases when N increases, whereas the Khatri-Rao variance reduced method is particularly effective producing strictly better performance than the other two.

Appendix D Technical Lemmas

In this section, we list some technical lemmas we use in this paper. All of them are about tail probability of sub-Gaussian or generalized sub-exponential variables.

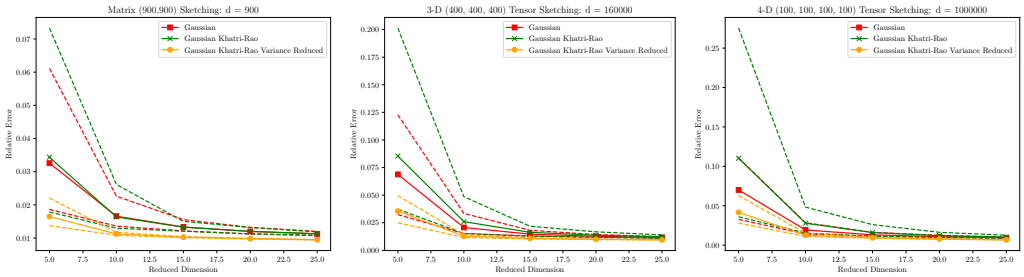

Figure 6: Relative Error for the low-rank tensor unfolding approximation: *we compare the relative errors for low-rank tensor approximation with different input size: 2-D (*900×900*), 3-D (*$400 \times 400 \times 400$*), 4-D (*$100 \times 100 \times 100 \times 100$*). In each setting, we compare the performance of Gaussian RP, TRP, and TRP(5). The dashed line stands for the 95% confidence interval.*

Definition D.1. *A random variable x is called sub-Gaussian if $\mathbb{E}|x|^p = \mathcal{O}(p^{p/2})$ when $p \to \infty$. With this, we define sub-Gaussian norm for x (less than infinity) as*

$$\|x\|_{\varphi_2} = \sup_{p \geqslant 1} p^{-1/2} (\mathbb{E}|x|^p)^{1/p}. \tag{D.1}$$

Note that for Bernoulli random variable, i.e., $\{-1, 1\}$ with prob. $\{\frac{1}{2}, \frac{1}{2}\}$, $\varphi_2 = 1$; any bounded random variable with absolute value less than $M > 0$ has $\varphi_2 \leqslant M$. For standard Gaussian random variable, $\varphi_2 = 1$.

Lemma D.1. *(Hanson-Wright Inequality) Let $\mathbf{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$ be a random vector with independent components X_i which satisfies $\mathbb{E}\mathbf{x}_i = 0$ and $\varphi_2(x_1) \leqslant K$. Let A be an $n \times n$ matrix. Then, for every $\eta \geqslant 0$, there exists a general constant c s.t.*

$$\mathbb{P}\left(|\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbb{E}\mathbf{x}^\top \mathbf{A}\mathbf{x}| \geqslant \eta\right) \leqslant 2\exp\left[-c\min\left\{\frac{\eta}{K^2\|A\|}, \frac{\eta^2}{\|A\|_F^2 K^4}\right\}\right].$$

Proof. Please refer to [19] □

Lemma D.2. *Let $\mathbf{x}$ be a random variable whose tail probability satisfies for every $\eta \geqslant 0$, there exists a constant c_1 s.t.*

$$\mathbb{P}\left(|x| \geqslant \eta\right) \leqslant 2\exp\left[-c_1 \min\left(\eta, \eta^2\right)\right].$$

Then for any $k \geqslant 1$, x satisfies generalized sub-exponential moment condition A.1 with $\alpha = 1$, i.e.,

$$\mathbb{E}|x|^k \leqslant (Ck)^k,$$

where $C = 1 + \frac{1}{c_1}$.

Proof.

$$\begin{aligned}
\mathbb{E}|x|^k &= \int_0^1 k x^{k-1} 2\exp[-c_1 x^2]dx + \int_1^\infty k x^{k-1} 2\exp[-c_1 x]dx \\
&\leqslant 1 + + \frac{1}{c_1^k}\int_0^\infty k y^{k-1} 2\exp[-y]dy \\
&= 1 + \frac{1}{c_1^k} k\Gamma(k-1) \leqslant \left[1 + \frac{1}{c_1^k}\right] k^k.
\end{aligned} \tag{D.2}$$

Noticing $\left[1 + \frac{1}{c_1^k}\right]^{1/k} \leqslant 1 + \frac{1}{c_1}$, we finish the proof. □

Lemma D.3. *For a random vector $\mathbf{x}$ with each element independent and identically distributed with mean zero and variance 1, suppose each element of $\mathbf{x}$ satisfies generalized sub-exponential moment condition as in* (D.2)*, that there exists a general constant C s.t. $\mathbb{E}|x_1|^k \leqslant (Ck)^{\alpha k}$. Then for any matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, there exists a general constant c_1*

$$\mathbb{P}\left(\left|\mathbf{x}^\top \mathbf{A}\mathbf{x} - \mathbb{E}\mathbf{x}^\top \mathbf{A}\mathbf{x}\right| \geqslant \eta\right) \leqslant C\exp\left(-c_1\left[\frac{\eta}{\|A\|_F}\right]^{1/(2(1+\alpha))}\right).$$

Proof. The proof is directly from Lemma 8.3 in [7] and we change the statement on generalized sub-exponential R.V. directly to the statement on the moment condition. □