Simple, unified analysis of Johnson-Lindenstrauss with applications

Yingru Li* YINGRULI@LINK.CUHK.EDU.CN
The Chinese University of Hong Kong, Shenzhen, China

Abstract

We present a simplified and unified analysis of the Johnson-Lindenstrauss (JL) lemma, a cornerstone of dimensionality reduction for managing high-dimensional data. Our approach simplifies understanding and unifies various constructions under the JL framework, including spherical, binary-coin, sparse JL, Gaussian, and sub-Gaussian models. This unification preserves the intrinsic geometry of data, essential for applications from streaming algorithms to reinforcement learning. We provide the first rigorous proof of the spherical construction's effectiveness and introduce a general class of sub-Gaussian constructions within this simplified framework. Central to our contribution is an innovative extension of the Hanson-Wright inequality to high dimensions, complete with explicit constants. By using simple yet powerful probabilistic tools and analytical techniques, such as an enhanced diagonalization process, our analysis solidifies the theoretical foundation of the JL lemma by removing an independence assumption and extends its practical applicability to contemporary algorithms.

Keywords: Dimensionality reduction, Johnson-Lindenstrauss, Hanson-Wright, Matrix factorization, Uncertainty estimation, Epistemic Neural Networks (ENN), Hypermodel

1. Introduction

In the realm of modern computational algorithms, dealing with high-dimensional data often necessitates a preliminary step of dimensionality reduction. This process is not merely a matter of convenience but a critical operation that preserves the intrinsic geometry of the data. Such dimensionality reduction techniques find widespread application across a diverse array of fields, including but not limited to streaming algorithms [20], compressed sensing [3, 5], numerical linear algebra [30], feature hashing [29], uncertainty estimation [18, 22] and reinforcement learning [17, 18]. These applications underscore the technique's versatility and its fundamental role in enhancing algorithmic efficiency.

The essence of geometry preservation within the context of dimensionality reduction can be mathematically formulated as the challenge of designing a probability distribution over matrices that effectively retains the norm of any vector within a specified error margin after transformation. Specifically, for a given vector $x \in \mathbb{R}^n$, the objective is to ensure that with probability at least $1-\delta$, the norm of x after transformation by a matrix $\Pi \in \mathbb{R}^{m \times n}$ drawn from the distribution $\mathcal{D}_{\varepsilon,\delta}$ remains ϵ-approximation of its original norm, as shown below:

$$\mathbb{P}_{\Pi \sim \mathcal{D}_{\varepsilon,\delta}} \left(\|\Pi x\|_2^2 \in \left[(1-\varepsilon)\|x\|_2^2, (1+\varepsilon)\|x\|_2^2 \right] \right) \geq 1-\delta \tag{1}$$

* The author would like to acknowledge Professor Zhi-Quan (Tom) Luo for advising this project.

A foundational result in this domain, the following Johnson-Lindenstrauss (JL) lemma, establishes a theoretical upper bound on the reduced dimension m, achievable while adhering to the above-prescribed fidelity criterion.

Lemma 1 (JL lemma [14]) *For any $0 < \varepsilon, \delta < 1/2$, there exists a distribution $\mathcal{D}_{\varepsilon,\delta}$ on $\mathbb{R}^{m\times n}$ for $m = O(\varepsilon^{-2}\log(1/\delta))$ that satisfies eq. (1).*

Recent research [13, 15] has validated the optimality of the dimension m specified by this lemma, further cementing its significance in the field of dimensionality reduction.

Initially, the constructive proof for Lemma 1 is based on the random k-dimensional subspace [7, 9, 14]. Projection to a random subspace involves computing a random rotation matrix, which requires computational-intensive orthogonalization processes. Along the decades, many alternative JL distributions $\mathcal{D}_{\varepsilon,\delta}$ were developed for the convenience of computation and storage. Indyk and Motwani [12] chooses the entries of Π as independent Gaussian random varaibles, i.e. $\Pi \sim \frac{1}{\sqrt{m}} \cdot N(0,1)^{\otimes(m\times n)}$ where the random matrix is easier and faster to generate by skipping the orthogonalization procedure. Achlioptas [1] showed the Gaussian distribution can be relaxed to a much simpler distribution only by drawing random binary coins, i.e. $\frac{1}{\sqrt{m}} \cdot \mathcal{U}(\{1,-1\})^{\otimes(m\times n)}$. Matoušek [19] generalizes such analytical techniques to i.i.d sub-Gaussian entries $\mathrm{SG}^{\otimes(m\times n)}$. To further speedup the projection on high-dimensional sparse data, a series of works on design and analyze sparse JL transform [6, 11, 16] was proposed. In sparse JL, the column vector could be expressed as entrywise multiplication $\sqrt{\frac{m}{s}}\sigma \odot \eta$ by $\sigma \sim \frac{1}{\sqrt{m}}\mathcal{U}(\{1,-1\})^{\otimes m}$ and a random vector η with only s non-zero entries. These works extends the class of JL distributions.

One alternative is the spherical construction where each column of Π is independently sampled from uniform distribution over the sphere $\mathbb{S}^{m-1}$, i.e., $\Pi \sim \mathcal{U}(\mathbb{S}^{m-1})^{\otimes n}$. Spherical construction was recently shown its superior performance in the application of incremental uncertainty estimation and reinforcement learning via hypermodel [8, 17, 18] and epistemic neural networks (ENN) [22, 23]. However, existing analysis of JL requires some notion of independence across the entries of each column vector in the random projection matrix Π while the spherical construction violates. This limitation comes from the requirement on the sum of independent random variables to facilitates the concentration analysis within the existing probabilistic analytical frameworks.

Challenge: Prove that spherical construction is a JL distribution satisfying Lemma 1.

JL dist. (w/o scaling)	$N(0,1)^{\otimes(m\times n)}$	$\mathcal{U}(\{1,-1\})^{\otimes(m\times n)}$	$\mathrm{SG}^{\otimes(m\times n)}$	SJLT	$\mathcal{U}(\mathbb{S}^{m-1})^{\otimes n}$	$\mathrm{SGV}^{\otimes n}$
[12]	✓					
[1]		✓				
[19]	✓	✓	✓			
[16]				✓		
[6]				✓		
[11]				✓		
Our work	✓	✓	✓	✓	✓	✓

Table 1: What types of constructions can be covered in the literature? SG stands for the distribution of sub-Gaussian random variables in $\mathbb{R}$. SGV stands for the distribution of sub-Gaussian random vectors in $\mathbb{R}^m$. SJLT stands for sparse JL transform introduced in [16].

We provide novel probability tools to resolve this challenge, as one of the contributions highlighted below:

- *Analysis of JL*: In Section 2, we present a unified but simple analysis of the Johnson-Lindenstrauss, encompassing spherical, binary-coin, Sparse JL (Proposition 15), Gaussian (Proposition 18) and sub-Gaussian constructions as particular instances. Proposition 8 marks the first rigorous demonstration of the spherical construction's efficacy, to the best of our knowledge. Also, with our analytical framework, we discover a new class of sub-Gaussian constructions in Definition 20, exhibiting potential useful properties. Summaries are in Table 1.

- *Technical innovations*: Our unified approach to JL analysis leverages an extension of the Hanson-Wright inequality to high dimensions, as detailed in Theorem 6. This tool is essential as it removes the requirement on independence across entries within a column vector of the projection matrix, the key to handle the spherical construction and a more general class of sub-Gaussian constructions. While the closest reference we identified is Exercise 6.2.7 in [27], our extensive review found no existing proofs of this assertion, nor does the mentioned exercise specify concrete constants, unlike our Theorem 6. Thus, our work in extending the Hanson-Wright inequality to high-dimension, complete with specific proof techniques, represents a significant advancement. Innovations include a novel approach to diagonalization step for the quadratic form.

- *Applications*: Leveraging our unified JL analysis and a covering argument, in Proposition 13, we establish a sufficient condition for reduced dimensionality within the context of covariance factorization procedures. This is inspired by the domains of uncertainty estimation and reinforcement learning. Recent neural network models, such as hypermodels [8, 17, 18] and epistemic neural networks [22, 23], leverage spherical random vectors to update a factorization matrix for incremental uncertainty estimation but lack rigorous guarantees. Our analysis justifies their effectiveness for the first time under the linear setups.

Notations. We say a random variable X is K-sub-Gaussian if $\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp\left(\lambda^2 K^2/2\right)$ for all $\lambda \in \mathbb{R}$. For random variables X in high-dimension $\mathbb{R}^m$, we say it is K-sub-Gaussian if for every fixed $v \in \mathbb{S}^{m-1}$ if the scalarized random variable $\langle v, X\rangle$ is K-sub-Gaussian.

2. Simple and Unified analysis of Johnson-Lindenstrauss

In this section, we are going to provide a simple and unified analysis for the following Johnson-Lindenstrauss constructions of random projection matrix satisfying lemma 1.

Definition 2 (Gaussian construction) *Gaussian construction of the random projection matrix* $\Pi = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ *correspond to each* $\mathbf{z}_i \sim \frac{1}{\sqrt{m}} N(0, I_m)$ *independently.*

Definition 3 (Binary-coin construction) *Binary-coin construction of the random projection matrix* $\Pi = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ *correspond to each* $\mathbf{z}_i \sim \frac{1}{\sqrt{m}} \mathcal{U}(\{1, -1\}^m)$ *independently.*

Definition 4 (s-sparse JL) *Sparse JL transform matrix* $\Pi = (\sqrt{\frac{m}{s}}\eta_1 \odot \mathbf{z}_1, \ldots, \sqrt{\frac{m}{s}}\eta_n \odot \mathbf{z}_n)$ *is a random matrix with each* $\mathbf{z}_i \sim P_{\mathbf{z}}$ *independently where* $P_{\mathbf{z}} := \frac{1}{\sqrt{m}} \mathcal{U}(\{1, -1\}^m)$ *and each* η_i *is independently and uniformly sampled from all possible* s*-hot vectors, where* s*-hot vectors is with exactly* s *non-zero entries with number* 1*. This construction is introduced by [16], also called countsketch.*

Notably, the entries $(\mathbf{z}_{i1}, \mathbf{z}_{i2}, \dots, \mathbf{z}_{im})$ within the random vector $\mathbf{z}_i$ in (1) Gaussian, (2) Binary-coin and (3) sparse JL constructions are mutually independent. However, the condition on the entry-independence is not true the next construction presented, which brings the major analytical difficulties that have not been discussed in the literature.

Definition 5 (Spherical construction) *Spherical construction of the random projection matrix $\Pi = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ corresponds to each $\mathbf{z}_i \sim \mathcal{U}(\mathbb{S}^{m-1})$ independently.*

Before stating our main result for Johnson-Lindenstrauss, we introduce the underlying new probability tool that enables the analysis of spherical construction.

Theorem 6 (High-dimensional Hanson-Wright inequality) *Let $X_1, \dots, X_n$ be independent, mean zero random vectors in $\mathbb{R}^m$, each X_i is K_i-subGaussian. Let $K = \max_i K_i$. Let $A = (a_{ij})$ be an $n \times n$ matrix. For any $t \geq 0$, we have*

$$\mathbb{P}\left(|\sum_{i,j:i\neq j}^{n} a_{ij}\langle X_i, X_j\rangle| \geq t \right) \leq 2\exp\left(-\min\left\{ \frac{t^2}{64mK^4\|A\|_F^2}, \frac{t}{8K^2\|A\|_2} \right\} \right).$$

Remark 7 *This is an high-dimension extension of famous Hanson-Wright inequality [10, 24, 31]. The Theorem 6 with exact constant is new in the literature, which maybe of independent interest. Our proof technique generalizes from [24] with new treatments on the diagnolization. The proof of Theorem 6 can be found in Section E. An extension of Theorem 6 to $\sum_{i,j=1}^{n} a_{ij}\langle X_i, X_j\rangle$ with non-negative diagonal is in Theorem 25.*

Now, we are ready to provide the unified analysis on Johnson-Lindenstrauss, a simple and direct application of Theorem 6.

Proposition 8 (Binary-coin; Spherical) *The Binary-coin and Spherical construction of the random projection matrix $\Pi \in \mathbb{R}^{m\times n}$ in definitions 3 and 5 with $m \geq 64\varepsilon^{-2}\log(2/\delta)$ satisfy Lemma 1.*

Proof From examples 1 and 2 as will be discussed in Section C, we know that the random variables sampled from $\mathcal{U}(\mathbb{S}^{m-1})$ or $\frac{1}{\sqrt{m}}\mathcal{U}(\{1,-1\}^m)$ are $\frac{1}{\sqrt{m}}$-sub-Gaussian with mean-zero and unit-norm. Let $x \in \mathbb{R}^d$ be the vector to be projected. By the construction of Π,

$$\|\Pi x\|^2 - \|x\|^2 = \underbrace{\sum_{1\leq i\neq j\leq n} x_i x_j\langle \mathbf{z}_i, \mathbf{z}_j\rangle}_{\text{off-diagonal}} + \underbrace{\sum_{i=1}^{n} x_i^2(\|\mathbf{z}_i\|^2 - 1)}_{\text{diagonal}} \tag{2}$$

As by the condition on unit norm, the diagonal term is zero. We apply Theorem 6 with $A = xx^\top$ and $t = \varepsilon\|x\|^2$. Since $K = 1/\sqrt{m}$ and $\|A\|_F = \sqrt{\mathrm{tr}(xx^\top xx^\top)} = \|x\|^2$, $\|A\|_2 = \|x\|^2$, then

$$\begin{aligned}\mathbb{P}\left(|\sum_{1\leq i\neq j\leq n} x_i x_j\langle \mathbf{z}_i, \mathbf{z}_j\rangle| \geq \varepsilon\|x\|^2 \right) &\leq 2\exp\left(-\min\left\{ \frac{\varepsilon^2\|x\|^4}{64K^4m\|A\|_F^2}, \frac{\varepsilon\|x\|^2}{8\sqrt{2}K^2\|A\|_2} \right\} \right) \\ &\leq 2\exp\left(-m\min\left\{ \varepsilon^2/64, \varepsilon/8\sqrt{2} \right\} \right).\end{aligned}$$

This implies that to get the RHS upper bound by δ, we need $m \geq 64\varepsilon^{-2}\log(2/\delta)$ for any fixed $\varepsilon \in (0, 1)$. ■

Remark 9 *This proposition is a unified analysis for (1) Spherical construction from random vectors in example 1 (2) Binary coin construction from random vectors in example 2. For classical Gaussian construction where* $\mathbf{z}_i \sim N(0, (1/m)I_m)$ *which does not satisfy unit-norm assumption, the diagonal term in eq.* (2) *is non-zero and needs additional treatments. As analyzed latter in Proposition 18 within the same framework, the requirement for dimension* $m = 8(1 + 2\sqrt{2})^2\varepsilon^{-2}\log(2/\delta)$ *in the Gaussian construction is larger than the one for Spherical construction. This observation may explain the practical superiority of Spherical construction.*

Remark 10 *Reduction of JL to the classical Hanson-Wright [10, 24, 31] has been exploited in [6, 16, 21], e.g. section 5.1 in [21]. However, as mentioned in section 1, their analytical assumption on the entry-wise independence, required by the reduction to classical Hanson-Wright, is violated in the spherical construction. Therefore, our high-dimensional extension of Hanson-Wright is crucial for the new unified analysis of JL, accommodating the spherical construction.*

3. Conclusion

This study marks a pivotal advancement in dimensionality reduction research by offering a simple and unified framework for the Johnson-Lindenstrauss lemma. Our streamlined approach not only makes the lemma more accessible but also broadens its application across various data-intensive fields, including a pioneering validation of spherical construction for uncertainty estimation and reinforcement learning. The simplification of the theoretical underpinnings, alongside the unification of multiple constructions under a single analytical lens, represents a significant contribution to both the academic and practical realms.

Through the extension of the Hanson-Wright inequality, providing precise constants for high-dimensional scenarios, and the introduction of novel probabilistic and analytical methods, we reinforce the JL lemma's indispensable role in navigating the complexities of high-dimensional data. This work underscores the power of simple, unified analyses in driving forward the understanding and application of fundamental concepts in computational algorithms and beyond, highlighting the direct pathway for future extensions and adaptations of random projection and Johnson-Lindentrauss.

Acknowledgments

The author would like to thank Professor Zhi-Quan (Tom) Luo for advising this project and Jiancong Xiao for helpful comments on the manuscript.

Appendix A. Application in Uncertainty Estimation

Folklore suggests scalable and incremental uncertainty estimation through hypermodels [8, 17, 18] and epistemic neural networks (ENN) [22, 23], yet no rigorous guarantees exist. These works consider settings where feature vectors $x_t \in \mathbb{R}^d$ for $t = 1, \ldots, T$ appear in a streaming fashion. This data stream assumption is grounded in reinforcement learning, where an agent interacts with environments and receives new observations sequentially.

Li et al. [18] summarize the closed-form incremental algorithm in linear setups, where it incrementally updates an $\mathbb{R}^{d\times M}$ matrix $\mathbf{A}$ using the sequences $(x_t)_{t\geq 1}$ and $(\mathbf{z}_t)_{t\geq 1}$, resulting in a matrix at time T given by

$$\mathbf{A} = \mathbf{\Sigma}\left(\mathbf{\Sigma}_0^{-1/2}\mathbf{Z}_0 + \frac{1}{\sigma}\sum_{t=1}^{T} x_t \mathbf{z}_t^\top\right), \tag{3}$$

where (1) $\mathbf{Z}_0 \in \mathbb{R}^{d\times M}$ and $\mathbf{z}_t \in \mathbb{R}^M$ are algorithm-generated random matrix and random vectors, and (2) $\mathbf{\Sigma} = \left(\mathbf{\Sigma}_0^{-1} + \frac{1}{\sigma^2}\sum_{t=1}^{T} x_t x_t^\top\right)$ is the posterior covariance matrix. Here, $\mathbf{\Sigma}_0 \in \mathbb{R}^{d\times d}$ is the prior covariance matrix and σ is the standard deviation of the response noise in the linear-Gaussian model.

Dwaracherla et al. [8], Li et al. [18], Osband et al. [22] typically generate these random vectors using spherical distribution and state that the goal is to ensure the matrix $\mathbf{A}$ is an approximate factorization of the posterior covariance matrix $\mathbf{\Sigma}$, i.e.,

$$\mathbf{A}\mathbf{A}^\top \approx \mathbf{\Sigma}. \tag{4}$$

Li et al. [18] provide an argument in expectation, i.e., $\mathbb{E}[\mathbf{A}\mathbf{A}^\top] = \mathbf{\Sigma}$, and Osband et al. [22] provide an argument of asymptotic convergence, i.e., $\mathbf{A}\mathbf{A}^\top \xrightarrow{a.s.} \mathbf{\Sigma}$ when $M \to \infty$. These statements do not justify the usefulness of hypermodels or ENN for uncertainty estimation. A high-probability non-asymptotic characterization of the approximation in eqs. (3) and (4) is necessary for rigorous justification of their usefulness. Unfortunately, such results are not known in the literature.

We now provide the first analysis using our proposed unified probability tool in proposition 8. First, we state the standard covering argument on the sphere and the argument on computing the norm on the covering set.

Lemma 11 (Covering number of a sphere) *There exists a set $\mathcal{C}_\varepsilon \subset \mathbb{S}^{d-1}$ with $|\mathcal{C}_\varepsilon| \leq (1 + 2/\varepsilon)^d$ such that for all $x \in \mathbb{S}^{d-1}$ there exists a $y \in \mathcal{C}_\varepsilon$ with $\|x - y\|_2 \leq \varepsilon$.*

Lemma 12 (Computing spectral norm on a covering set) *Let $\boldsymbol{A}$ be a symmetric $d \times d$ matrix, and let $\mathcal{C}_\varepsilon$ be an ε-covering of $\mathbb{S}^{d-1}$ for some $\varepsilon \in (0, 1)$. Then,*

$$\|\boldsymbol{A}\| = \sup_{x\in\mathbb{S}^{d-1}} |x^\top \boldsymbol{A} x| \leq (1 - 2\varepsilon)^{-1} \sup_{x\in\mathcal{C}_\varepsilon} |x^\top \boldsymbol{A} x|.$$

Now we state the result in covariance matrix factorization with the specific goal of approximating the quadratic form

$$(1 - \varepsilon)x^\top \mathbf{\Sigma} x \leq x^\top \mathbf{A}\mathbf{A}^\top x \leq (1 + \varepsilon)x^\top \mathbf{\Sigma} x, \quad \forall x \in \mathcal{X}, \tag{5}$$

where $\mathcal{X}$ might be some set of interest in applications, e.g., the action space in bandit problems or the state-action joint space in reinforcement learning. Notice that the approximation in eq. (4), i.e., $(1-\varepsilon)\mathbf{\Sigma} \preceq \mathbf{A}\mathbf{A}^\top \preceq (1+\varepsilon)\mathbf{\Sigma}$, reduces to eq. (5) when the set $\mathcal{X}$ is a compact set, e.g., $\{x \in \mathbb{R}^d : \|x\| = 1\}$.

Proposition 13 *Equation (5) holds with probability at least $1-\delta$ for the compact set $\mathcal{X} := \{x \in \mathbb{R}^d : \|x\| = 1\}$ if $M \geq 64\varepsilon^{-2}(d\log 9 + \log(2/\delta))$; for a finite set $\mathcal{X}$, if $M \geq 64\varepsilon^{-2}\log(2|\mathcal{X}|/\delta)$.*

Proof Let us denote the random matrix as

$$\mathbf{Z}^\top = (\mathbf{Z}_0^\top, \mathbf{z}_1, \ldots, \mathbf{z}_T) \in \mathbb{R}^{M\times(d+T)},$$

and the data matrix as

$$\boldsymbol{X} = (\mathbf{\Sigma}_0^{-1/2}, x_1/\sigma, \ldots, x_T/\sigma)^\top \in \mathbb{R}^{(d+T)\times d}.$$

Notice the inverse posterior covariance matrix is $\mathbf{\Sigma}^{-1} = \mathbf{\Sigma}_0^{-1} + (1/\sigma^2)\sum_{t=1}^T x_t x_t^\top = \boldsymbol{X}^\top\boldsymbol{X}$. Then, we can represent

$$\mathbf{A} = \mathbf{\Sigma}\left(\mathbf{\Sigma}_0^{-1/2}\mathbf{Z}_0 + \frac{1}{\sigma}\sum_{t=1}^T x_t\mathbf{z}_t^\top\right) = \mathbf{\Sigma}\boldsymbol{X}^\top\mathbf{Z}.$$

Then $\mathbf{A}\mathbf{A}^\top = \mathbf{\Sigma}\boldsymbol{X}^\top\mathbf{Z}\mathbf{Z}^\top\boldsymbol{X}\mathbf{\Sigma}$ and $\mathbf{\Sigma} = \mathbf{\Sigma}\boldsymbol{X}^\top\boldsymbol{X}\mathbf{\Sigma}$. The (ε,δ)-approximation goal in eq. (5) reduces to a random projection argument with projection matrix $\mathbf{Z}^\top \in \mathbb{R}^{M\times(d+T)}$ and the vector $\boldsymbol{X}\mathbf{\Sigma}x$ to be projected:

$$(1-\varepsilon)\|\boldsymbol{X}\mathbf{\Sigma}x\|^2 \leq \|\mathbf{Z}^\top\boldsymbol{X}\mathbf{\Sigma}x\|^2 \leq (1+\varepsilon)\|\boldsymbol{X}\mathbf{\Sigma}x\|^2, \quad \forall x \in \mathcal{X}. \tag{6}$$

For the compact set $\mathcal{X} = \mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\| = 1\}$, by standard covering argument in lemma 12 and proposition 8, eq. (6) holds with probability $1-\delta$ when $M \geq 64\varepsilon^{-2}(d\log 9 + \log(2/\delta))$. For a finite set $\mathcal{X}$, direct application of the union bound with proposition 8 yields the result. ■

Appendix B. Sparse JL and General sub-Gaussian Constructions

B.1. Sparse JL transform

We also present a generalization of theorem 6 that will be helpful to analyze sparse JL transform.

Theorem 14 (Generalized High-dimensional Hanson-Wright) *Let $b_1, \ldots, b_n$ be fixed vectors in $\mathbb{R}^m$ where b_{ik} is the k-th entry of the vector b_i. Let $X_1, \ldots, X_n$ be independent, mean zero random vectors in $\mathbb{R}^m$, each X_i is K_i-subGaussian. Let $K = \max_i K_i$. Let $A = (a_{ij})$ be an $n \times n$ matrix. For any $t \geq 0$, we have*

$$\mathbb{P}\left(|\sum_{i,j:i\neq j}^n a_{ij}\langle b_i \odot X_i, b_j \odot X_j\rangle| \geq t\right) \leq 2e^{-\min\left\{\frac{t^2}{64K^4\sum_{k=1}^m \|A_k^b\|_F^2}, \frac{t}{8K^2\max_k \|A_k^b\|_2}\right\}}.$$

where A_k^b is a matrix with entries $A_k^b(i,j) = a_{ij}b_{ik}b_{jk}$ for each $(k,i,j) \in [m]\times[n]\times[n]$.

Theorem 14 extends Theorem 6 in a way that each random vector X_i is entry-wise scaled by corresponding b_i for $i \in [n]$. When $b_1 = b_2 = \cdots = b_n = \mathbf{1}$ is all-one vector, it reduces to Theorem 6. The proof of Theorem 14 is similar to Theorem 6 and is deferred to Section F. Now we are ready to include the sparse JL construction into our unified analytical framework.

Proposition 15 *The sparse JL construction in definition 4 with* $m \simeq \varepsilon^{-2}\log(1/\delta)$ *and* $s \simeq \varepsilon^{-1}\log(1/\delta)$ *satisfies Lemma 1.*

Proof From example 2, we know that $\mathbf{z}_i \sim P_{\mathbf{z}} = \frac{1}{\sqrt{m}}\mathcal{U}(\{1,-1\}^m)$ is a $\frac{1}{\sqrt{m}}$-sub-Gaussian random vector with mean zero and unit-norm. Let $x \in \mathbb{R}^d$ be the vector to be projected. By the construction of Π,

$$\|\Pi x\|^2 - \|x\|^2 = \underbrace{\sum_{1\le i\neq j\le n} \frac{m}{s} x_i x_j \langle \eta_i \odot \mathbf{z}_i, \eta_j \odot \mathbf{z}_j\rangle}_{\text{off-diagonal}} + \underbrace{\sum_{i=1}^n x_i^2 (\frac{m}{s}\|\eta_i \odot \mathbf{z}_i\|^2 - 1)}_{\text{diagonal}} \tag{7}$$

By the sparse JL construction in definition 4, the diagonal term in eq. (7) is zero. W.L.O.G, we assume that $\|x\|^2 = 1$. We could apply Theorem 14 conditioned on $(\eta_i)_i$ with $A = (m/s)xx^\top$, $(b_i) = (\eta_i)$ and $t = \varepsilon$. The constructed matrix in the Theorem 14 will be $A_k^b = \frac{m}{s}(x \odot \eta^k)(x \odot \eta^k)^\top$ where $\eta^k = (\eta_{1k}, \eta_{2k}, \ldots, \eta_{nk})$. Indeed, $\|A_k^b\|_F = \sum_{ij}(m/s)^2 x_i^2 x_j^2 \eta_{ik}\eta_{jk}$ and $\|A_k^b\|_2 = (m/s)\|(x \odot \eta^k)\|_2^2 \le (m/s)$. Since $K = 1/\sqrt{m}$, Theorem 14 yields,

$$\mathbb{P}\left(\text{off-diagonal} \ge \varepsilon \mid (\eta_i)_{i=1}^n\right) \le 2\exp\left(-\frac{\varepsilon^2}{64(1/s^2)\sum_{k=1}^m \sum_{i,j} x_i^2 x_j^2 \eta_{ik}\eta_{jk}}\right) + 2\exp\left(-\frac{\varepsilon}{8\sqrt{2}(1/s)}\right).$$

With a translation of tail bound to moment bound in lemma 17,

$$\underbrace{\left(\mathbb{E}\left[|\text{off-diagonal}|^p \mid (\eta_i)_{i=1}^n\right]\right)^{1/p}}_{(a)} \lesssim \frac{\sqrt{p}}{s}\sqrt{\sum_{ij} x_i^2 x_j^2 \sum_{k=1}^m \eta_{ik}\eta_{jk}} + \frac{p}{s}. \tag{8}$$

Then by the tower property and eq. (8)

$$\begin{aligned}
\left(\mathbb{E}\left[|\text{off-diagonal}|^p\right]\right)^{1/p} &= \left(\mathbb{E}\left[(a)^p\right]\right)^{1/p} \\
&\lesssim \left(\mathbb{E}\left(\frac{\sqrt{p}}{s}\sqrt{\sum_{ij} x_i^2 x_j^2 \sum_{k=1}^m \eta_{ik}\eta_{jk}} + \frac{p}{s}\right)^p\right)^{1/p} \\
&\le \frac{\sqrt{p}}{s}\underbrace{\left(\mathbb{E}\left(\sqrt{\sum_{ij} x_i^2 x_j^2 \sum_{k=1}^m \eta_{ik}\eta_{jk}}\right)^p\right)^{1/p}}_{(b)} + \frac{p}{s},
\end{aligned} \tag{9}$$

where the last inequality is by triangular inequality of L_p-norm. The term (b) can be bounded as follows when $p \simeq s^2/m$,

$$(b) \overset{(1)}{\le} \sqrt{\sum_{ij} x_i^2 x_j^2 \left(\mathbb{E}\left(\sum_{k=1}^m \eta_{ik}\eta_{jk}\right)^p\right)^{1/p}} \overset{(2)}{\lesssim} \sqrt{\sum_{ij} x_i^2 x_j^2 p} = \sqrt{p}, \tag{10}$$

where (1) is by Jensen's inequality; (2) follows by lemma 16 as $\left(\mathbb{E}\left(\sum_{k=1}^{m}\eta_{ik}\eta_{jk}\right)^p\right)^{1/p} \lesssim \sqrt{s^2/m}\cdot\sqrt{p}+p \simeq p$ when $p \simeq s^2/m$; and the last equality follows the assumption $\|x\|^2 = 1$, resulting $\sum_i x_i^2 \sum_j x_j^2 = 1 \cdot 1$. Therefore, plugging the upper bound in eq. (10) with $p \simeq s^2/m$ into eq. (9),

$$(\mathbb{E}\left[|\text{off-diagonal}|^p\right])^{1/p} \lesssim \sqrt{\frac{p}{m}} + \frac{p}{s} \simeq \frac{p}{s} \simeq \frac{s}{m}$$

Then by Markov's inequality and the settings of $p \simeq s^2/m, s \simeq \varepsilon m, m \simeq \varepsilon^{-2}\log(1/\delta)$,

$$\begin{aligned}\mathbb{P}\left(\left|\|\Pi x\|_2^2 - 1\right| > \varepsilon\right) &= \mathbb{P}\left(|\text{off-diagonal}| > \varepsilon\right) < \varepsilon^{-p}\cdot\mathbb{E}\left[|\text{off-diagonal}|^p\right] \\ &< \varepsilon^{-p}\cdot(\frac{s}{m})^p \cdot C^p < C^{\log(1/\delta)} < \delta,\end{aligned}$$

where C is some constant as a result of configuration in p, m, s for the purpose. ■

Lemma 16 *For $\eta_i, i = 1, \ldots, n$ defined in definition 4, the p-th moment of $\sum_{k=1}^{m}\eta_{ik}\eta_{jk}$ is bounded*

$$\left(\mathbb{E}\left(\sum_{k=1}^{m}\eta_{ik}\eta_{jk}\right)^p\right)^{1/p} \lesssim \sqrt{s^2/m}\cdot\sqrt{p}+p$$

Proof Suppose the event I is that $\eta_{i,a_1}, \ldots, \eta_{i,a_s}$ are all 1 , where $a_1 < a_2 < \ldots < a_s$. Note that conditioned on event I, the sum $\sum_{k=1}^{m}\eta_{ik}\eta_{jk}$ can be written as $\sum_{k=1}^{s} Y_k$, where Y_k is an indicator random variable for the event that $\eta_{j,a_k} = 1$. The $(Y_k)_{k=1}^s$ are not independent, but for any integer $p \geq 1$ their p th moment is upper bounded by the case that the $(Y_k)_{k=1}^s$ are independent Bernoulli each of expectation (s/m) (this can be seen by simply expanding $(\sum_{k=1}^{s} Y_k)^p$ then comparing with the independent Bernoulli case monomial by monomial in the expansion as shown in [6]). Finally, via the moment version of the Bernstein inequality, we obtain

$$\left(\mathbb{E}\left(\sum_{k=1}^{s} Y_k\right)^p\right)^{1/p} \lesssim \sqrt{s\frac{s}{m}\left(1-\frac{s}{m}\right)}\cdot\sqrt{p}+p \leq \sqrt{\frac{s^2}{m}}\cdot\sqrt{p}+p.$$

The lemma follows from taking the expectation over the event I and the tower property of expectation,

$$\mathbb{E}\left[\left(\sum_{k=1}^{m}\eta_{ik}\eta_{jk}\right)^p\right] = \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{k=1}^{m}\eta_{ik}\eta_{jk}\right)^p \middle| I\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\left(\sum_{k=1}^{s} Y_k\right)^p \middle| I\right]\right].$$

■

Lemma 17 (Theorem 2.3 in [4]) *Let Z be a scalar random variable. The following statements are equivalent. (a) There exist $\sigma, K > 0$ s.t. $\forall p \geq 1, \|Z\|_p \leq C_a(\sigma\sqrt{p}+Kp)$. (b) There exist $\sigma, K > 0$ s.t. $\forall \lambda > 0, \mathbb{P}(|Z| > \lambda) \leq C_b\left(e^{-C_b'\lambda^2/\sigma^2} + e^{-C_b'\lambda/K}\right)$. The constants C_a, C_b and C_b' change by at most some absolute constant factor.*

B.2. General sub-Gaussian construction without unit-norm

In this section, we consider the cases where the diagonal term in the decomposition (eq. (2)) is non-zero. We need additional conditions to guarantee Lemma 1, a two-sided probability bound. Before diving into the general treatment of sub-Gaussian setups, let us first look at the classical Gaussian construction in definition 2 where the column vector does not satisfy the unit-norm condition and we could get some intuition on more general case.

Proposition 18 (Gaussian) *The Gaussian construction of the random projection matrix $\Pi \in \mathbb{R}^{m\times n}$ in definition 2 with $m \geq 8(1+2\sqrt{2})^2\varepsilon^{-2}\log(2/\delta)$ satisfy Lemma 1.*

Remark 19 *The required dimension $m = 8(1+2\sqrt{2})^2\varepsilon^{-2}\log(2/\delta)$ in the Gaussian construction to guarantee lemma 1 is larger than the one $m = 64\varepsilon^{-2}\log(2/\delta)$ in spherical and binary coin construction as shown in proposition 8. Since we analyze these constructions within the same analytical framework, the smaller m in Spherical construction may explain its practical superiority.*

Proof The random variables sampled from $N(0, \frac{1}{m}I_m)$ are $\frac{1}{\sqrt{m}}$-sub-Gaussian with mean-zero. The off-diagonal term as decomposed in eq. (2) can be dealt as the same in proposition 8 via theorem 6. However, the diagonal term is non-zero in Gaussian construction. Notice that, the diagonal term $\sum_{i=1}^n x_i^2(\|\mathbf{z}_i\|^2 - 1)$, is essentially a weighted sum of i.i.d. χ^2_m random variables. Let $Z_{ij} \sim N(0,1)$ for all $(i,j) \in [n]\times[m]$.

$$\mathbb{E}\left[\exp(\lambda \sum_{i=1}^n x_i^2(\|\mathbf{z}_i\|^2 - 1))\right] = \mathbb{E}\left[\exp\left(\sum_{i=1}^n\sum_{j=1}^m \frac{\lambda x_i^2}{m}(Z_{ij}^2 - 1)\right)\right]. \tag{11}$$

As $\max_i \lambda x_i^2/m \leq 1/2$, the moment generating function of the diagonal terms will become

$$\prod_{i=1}^n\prod_{j=1}^m \frac{\exp(-\lambda x_i^2/m)}{\sqrt{1-2\lambda x_i^2/m}} \leq \exp\left(m\cdot\frac{2\lambda^2}{m^2}\sum_i x_i^4\right), \quad \forall|\lambda| < \frac{m}{4\max_i x_i^2}, \tag{12}$$

where the last inequality is due to $\frac{\exp(-x)}{\sqrt{1-2x}} \leq \exp 2x^2$ for $|x| < 1/4$. Notice $\max_i x_i^2 = \|x\|_\infty^2$. Finally, we have,

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n x_i^2(\|\mathbf{z}_i\|^2 - 1) \geq t\|x\|^2\right) &\leq \inf_{|\lambda|<\frac{m}{4\|x\|_\infty^2}} \exp(-\lambda t + 2\lambda^2\sum_i x_i^4/m) \\ &= \exp\left(-m\cdot\min\left\{\frac{t^2}{8\sum_i x_i^4}, \frac{t}{8\|x\|_\infty^2}\right\}\right).\end{aligned}$$

As we need to deal with diagonal term separately with the off-diagonal term in eq. (2), say let $\varepsilon = \varepsilon_1 + \varepsilon_2$,

$$\begin{aligned}\mathbb{P}(|\|\Pi x\|^2 - \|x\|^2| \geq \varepsilon\|x\|^2) &\leq \mathbb{P}(|\text{off-diagonal}| \geq \varepsilon_1\|x\|^2) + \mathbb{P}(|\text{diagonal}| \geq \varepsilon_2\|x\|^2) \\ &\leq 2\exp\left(-m\cdot\min\left\{\frac{\varepsilon_1^2}{64}, \frac{\varepsilon_1}{8\sqrt{2}}\right\}\right) + 2\exp\left(-m\cdot\min\left\{\frac{\varepsilon_2^2}{8}, \frac{\varepsilon_2}{8}\right\}\right),\end{aligned}$$

where the last inequality is true due to the fact $\|x\|_\infty^2 \le \|x\|^2$ and $\sum_i x_i^4 < \|x\|^4$. Let $\varepsilon_1 = \frac{2\sqrt{2}}{1+2\sqrt{2}}\varepsilon$ and $\varepsilon_2 = \frac{1}{1+2\sqrt{2}}\varepsilon$, we conclude in the Gaussian construction of Π

$$\mathbb{P}(|\|\Pi x\|^2 - \|x\|^2| \ge \varepsilon\|x\|^2) \le 4\exp\left(-\frac{m\varepsilon^2}{8(1+2\sqrt{2})^2}\right).$$

To guarantee Lemma 1, we require $m \ge 8(1+2\sqrt{2})^2\varepsilon^{-2}\log(4/\delta)$. ∎

In general, we cannot expect a lower tail bound for the squared norm of sub-Gaussian random variables in high dimension. Since lemma 1 is a two-sided tail bound, we make the following Bernstein-type assumption on the squared norm, in addition to the mean-zero independent sub-Gaussian condition.

Definition 20 (Sub-Gaussian construction with Bernstein condition) *Sub-Gaussian construction of the random projection matrix $\Pi = (\mathbf{z}_1, \dots, \mathbf{z}_n)$ has each column $\mathbf{z}_i$ being independent $\sqrt{1/m}$-sub-Gaussian random variable in $\mathbb{R}^m$ with mean zero. Additionally, there exists a universal constant $C > 0$ such that*

$$\mathbb{E}\left|\|\mathbf{z}_i\|^2 - \mathbb{E}\|\mathbf{z}_i\|^2\right|^k \le Ck!\left(\frac{1}{m}\right)^{\frac{k-2}{2}} \qquad \forall k = 3, 4, \dots$$

Remark 21 *Gaussian construction in definition 2 is a special case of the sub-Gaussian construction in definition 20 as χ_m^2 satisfies the Bernstein condition. Meanwhile, the sub-Gaussian construction in definition 20 generalize the spherical and binary-coin constructions. As we do not assume the random vector in each column has fixed norm, this also relax the analytical assumption of the Theorem 5.58 in [26] for extreme singular value of random matrix with independent sub-Gaussian columns.*

Remark 22 *Sub-Gaussian construction in definition 20 requires the same order of m as in Gaussian construction to guarantee lemma 1. The proof is a direct application of the Composition property of sub-Exponential random variables [27, 28].*

Appendix C. Typical sub-Gaussian distributions

In this section, we examine the properties of typical distribution for construction random projection matrix. Specifically, we examine sub-Gaussian condition of two high-dimensional distributions: (1) Uniform distribution over the unit sphere, and (2) Uniform distribution over the scaled cube. Before diving to the details, we first introduce a useful lemma on centered MGF for Beta distribution with a tight sub-Gaussian constant.

Lemma 23 (MGF of Beta distribution) *For any $\alpha, \beta \in \mathbb{R}_+$ with $\alpha \ge \beta$. Random variable $X \sim \mathrm{Beta}(\alpha, \beta)$ has variance $\mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$ and the centered MGF*

$$\mathbb{E}\left[\exp(\lambda(X - \mathbb{E}[X]))\right] \le \exp\left(\frac{\lambda^2 \mathrm{Var}(X)}{2}\right).$$

Remark 24 *The constant in lemma 23 is new in the literature and seems to be tight as it already achieve the same constant in the MGF of Gaussian distribution with variance* $\mathrm{Var}\,(X)$.

Proof For $X \sim \mathrm{Beta}(\alpha, \beta)$, Skorski [25] gives a novel order-2-recurrence for central moments.

$$\begin{aligned}\mathbb{E}\left[(X-\mathbb{E}[X])^p\right] =& \frac{(p-1)(\beta-\alpha)}{(\alpha+\beta)(\alpha+\beta+p-1)} \cdot \mathbb{E}\left[(X-\mathbb{E}[X])^{p-1}\right] \\ &+ \frac{(p-1)\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+p-1)} \cdot \mathbb{E}\left[(X-\mathbb{E}[X])^{p-2}\right]\end{aligned}$$

Let $m_p := \frac{\mathbb{E}[(X-\mathbb{E}[X])^p]}{p!}$, When $\alpha \geq \beta$, it follows that m_p is non-negative when p is even, and negative otherwise. Thus, for even p,

$$m_p \leq \frac{1}{p} \cdot \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+p-1)} m_{p-2} \leq \frac{\mathrm{Var}\,(X)}{p} \cdot m_{p-2}.$$

Repeating this $p/2$ times and combining with $m_p \leqslant 0$ for odd p, we obtain

$$m_p \leqslant \begin{cases} \frac{\mathrm{Var}(X)^{\frac{p}{2}}}{p!!} & p \text{ even} \\ 0 & p \text{ odd} \end{cases}.$$

Using $p!! = 2^{p/2}(p/2)!$ for even p, for $t \geqslant 0$ we obtain

$$\mathbb{E}\left[\exp(\lambda[X-\mathbb{E}\left[X\right]])\right] \leqslant 1 + \sum_{p=2}^{+\infty} m_p \lambda^p = 1 + \sum_{p=1}^{+\infty} (\lambda^2 \mathrm{Var}\,(X)/2)^p/p! = \exp\left(\frac{\lambda^2 \mathrm{Var}\,(X)}{2}\right)$$

■

Example 1 (Uniform distribution over m-dimensional sphere $\mathcal{U}(\mathbb{S}^{m-1})$) *Unit-norm condition is trivial to verify. Given a random vector* $\mathbf{z} \sim \mathcal{U}(\mathbb{S}^{m-1})$*, for any* $v \in \mathbb{S}^{m-1}$*, we have*

$$\langle \mathbf{z}, v \rangle \sim 2\,\mathrm{Beta}\left(\frac{m-1}{2}, \frac{m-1}{2}\right) - 1.$$

Thus, by lemma 23, we confirm that the random variable $\mathbf{z} \in \mathbb{R}^m$ *is* $\frac{1}{\sqrt{m}}$*-sub-Gaussian.*

Example 2 (Uniform distribution over scaled m-dimensional cube) *The random variable* $\mathbf{z} \sim \frac{1}{\sqrt{m}} \cdot \mathcal{U}(\{1,-1\}^m)$ *is* $\frac{1}{m}$*-sub-Gaussian and with unit-norm. This is because we could sample the random vector* $\mathbf{z}$ *by sample each entry independently from* $\mathrm{z}_i \sim \frac{1}{\sqrt{m}}\mathcal{U}(\{1,-1\})$ *for* $i \in [m]$*. Then, for any* $v \in \mathbb{S}^{m-1}$*, by independence,*

$$\mathbb{E}\left[\exp(\lambda\langle v, \mathbf{z}\rangle)\right] = \prod_{i=1}^{m} \mathbb{E}\left[\exp(\lambda v_i \mathrm{z}_i)\right] \leq \prod_{i=1}^{m} \exp(\lambda^2 v_i^2/2m) = \exp(\lambda^2 \sum_i v_i^2/2m).$$

The inequality is due to MGF of rademacher distribution (e.g. Example 2.3 in [28]).

Appendix D. Non-negative Diagonal Extension for High-dimensional Hanson-Wright

Theorem 25 (High-dimensional Hanson-Wright with non-negative diagonal) *Let $X_1, \ldots, X_n$ be independent, mean zero random vectors in $\mathbb{R}^m$, each X_i is K_i-subGaussian. Let $K = \max_i K_i$. Let $A = (a_{ij})$ be an $n \times n$ matrix such that $a_{ii} \geq 0$. There exists a universal constant $C > 0$ such that for any $t \geq 0$, we have*

$$\mathbb{P}\left(|\sum_{i,j=1}^{n} a_{ij}\langle X_i, X_j\rangle| \geq t\right) \leq \exp\left(-C \min\left\{\frac{t^2}{mK^4\|A\|_F^2}, \frac{t}{K^2\|A\|_2}\right\}\right).$$

Proof Decompose $\sum_{1\leq i,j\leq n} a_{ij}\left\langle X_i, X_j\right\rangle = \sum_{i=1}^n a_{ii}\left\|X_i\right\|^2 + S$, where $S = \sum_{1\leq i\neq j\leq n} a_{ij}\left\langle X_i, X_j\right\rangle$. In view of the off-diagonal sum bound for S in Theorem 6, it suffices to show the following inequality for the diagonal sum: for any $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} a_{ii}\left\|X_i\right\|^2 \geq m\sum_{i=1}^{n} a_{ii}K_i^2 + t\right) \leq \exp\left[-C\min\left(\frac{t^2}{mK^4\sum_{i=1}^n a_{ii}^2}, \frac{t}{K^2\max_{1\leq i\leq n} a_{ii}}\right)\right] \tag{13}$$

since $\sum_{i=1}^n a_{ii}^2 \leq \|A\|_F^2$ and $\bar{a} := \max_{1\leq i\leq n} a_{ii} \leq \|A\|_2$. By Markov's inequality and Lemma 28, we have for any $\lambda > 0$ and $t > 0$,

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^{n} a_{ii}\left(\left\|X_i\right\|^2 - mK_i^2\right) \geq t\right) &\leq e^{-\lambda t}\prod_{i=1}^{n}\mathbb{E}\left[e^{\lambda a_{ii}\left(\|X_i\|^2 - mK_i^2\right)}\right] \\ &\leq e^{-\lambda t}\prod_{i=1}^{n} e^{2\lambda^2 a_{ii}^2 mK_i^4} \\ &\leq \exp\left(-\lambda t + 2\lambda^2 m\left(\sum_{i=1}^{n} a_{ii}^2\right)K^4\right)\end{aligned}$$

holds for all $0 \leq \lambda < \left(4K^2\bar{a}\right)^{-1}$. Choosing

$$\lambda = \frac{t}{4\left(\sum_{i=1}^n a_{ii}^2\right)mK^4} \wedge \frac{1}{8\bar{a}K^2\|\Gamma\|_2},$$

we get eq. (13). ■

Lemma 26 (Gaussianization for squared norm of a σ-sub-Gaussian random variable in $\mathbb{R}^n$) *Let X be a random variable in $\mathbb{R}^n$ such that $\mathbb{E}[X] = 0$ and $\mathbb{E}[e^{z^\top X}] \leq \exp(\sigma^2\|z\|^2/2)$ for all $z \in \mathbb{R}^n$. Let $Z \sim N(0, \sigma^2 I)$. Then,*

$$\mathbb{E}\left[\exp\frac{t\|X\|_2^2}{2}\right] \leq \mathbb{E}\left[\exp\frac{t\|Z\|_2^2}{2}\right], \quad \forall 0 \leq t < \sigma^{-2}.$$

Proof The case for $t = 0$ is obvious. Consider $t \in (0, \sigma^{-2})$. Observe that

$$\begin{aligned} A &:= \frac{1}{(2\pi)^{n/2}\sigma^n} \int_{\mathbb{R}^n} \exp\left(-\frac{\|z\|^2}{2t}\right) \mathbb{E}\left[\exp z^\top X\right] dz \\ &\overset{(1)}{=} \mathbb{E}\left[\frac{1}{(2\pi)^{n/2}\sigma^n} \int_{\mathbb{R}^n} \exp\left(-\frac{\|z - tX\|_2^2}{2t}\right) dz \exp\left(\frac{t\|X\|_2^2}{2}\right)\right] \\ &\overset{(2)}{=} \mathbb{E}\left[\exp\left(\frac{t\|X\|_2^2}{2}\right)\right] \frac{1}{(2\pi)^{n/2}\sigma^n} \int_{\mathbb{R}^n} \exp\left(-\frac{\|z\|_2^2}{2t}\right) dz \\ &\overset{(3)}{=} \mathbb{E}\left[\exp\left(\frac{t\|X\|_2^2}{2}\right)\right] \frac{1}{t^{-n/2}\sigma^n}, \end{aligned}$$

where (1) follows from Fubini's theorem, (2) from the translational invariance of the Gaussian density integral, and (3) from that the integration of the standard Gaussian distribution $N(0, I_n)$ equals to one (requires $t > 0$). Thus, we get

$$\mathbb{E}\left[\exp\left(\frac{t\|X\|_2^2}{2}\right)\right] = t^{-n/2}\sigma^n A.$$

Since $\mathbb{E}\left[\exp z^T X\right] \le \exp(\sigma^2\|z\|^2/2)$ for all $z \in \mathbb{R}^n$, we have for $t \in \left(0, \sigma^{-2}\right)$,

$$\begin{aligned} A &\le \frac{1}{(2\pi)^{n/2}\sigma^n} \int_{\mathbb{R}^n} e^{-\frac{\|z\|^2}{2t}} e^{\frac{\sigma^2\|z\|^2}{2}} dz \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \int_{\mathbb{R}^n} e^{-\frac{1}{2}\left(t^{-1} - \sigma^2\right)\|z\|^2} dz \\ &= \frac{1}{\sigma^n(t^{-1} - \sigma^2)^{n/2}}. \end{aligned}$$

Then we have

$$\mathbb{E}\left[e^{\frac{t\|X\|_2^2}{2}}\right] \le \frac{t^{-n/2}\sigma^n}{\sigma^n(t^{-1} - \sigma^2)^{n/2}} = \frac{1}{(1 - \sigma^2 t)^{n/2}} \quad \forall 0 \le t < \sigma^{-2}.$$

On the other hand, for $Z \sim N(0, \sigma^2 I_n)$, similar calculations show that

$$\begin{aligned} \mathbb{E}\left[e^{\frac{s\|Z\|_2^2}{2}}\right] &= \frac{1}{(2\pi)^{n/2}\sigma^n} \int_{\mathbb{R}^n} e^{-\frac{1}{2}\sigma^{-2}\|z\|^2} e^{\frac{s}{2}\|z\|^2} dz \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \int_{\mathbb{R}^n} e^{-\frac{1}{2}(\sigma^{-2} - s)\|z\|^2} dz \\ &= \frac{1}{(1 - \sigma^2 s)^{n/2}} \quad \forall s < \sigma^{-2}. \end{aligned}$$

■

Remark 27 *Lemma 26 is true only for the upper tail as it requires $t \ge 0$. Without imposing additional assumptions, we cannot expect a lower tail bound for sub-Gaussian random variables as discussed in [2].*

Lemma 28 (Upper bound for MGF of squared norm of a σ-sub-Gaussian random variable in $\mathbb{R}^n$) *In the setting of lemma 26, we have*

$$\mathbb{E}\left[\exp\left(\frac{t}{2}\left(\|X\|_2^2 - n\sigma^2\right)\right)\right] \le \exp\left(\frac{t^2}{2}(n\sigma^4)\right) \quad \forall 0 \le t < (2\sigma^2)^{-1}. \tag{14}$$

Consequently, we have for any $u > 0$,

$$\mathbb{P}\left(\|X\|_2^2 - n\sigma^2 \ge u\right) \le \exp\left[-\frac{1}{8}\min\left(\frac{u^2}{n\sigma^4}, \frac{u}{\sigma^2}\right)\right]. \tag{15}$$

Proof Let $Z \sim N(0, \sigma^2 I_n)$. By the calculations in lemma 26, we have for all $t < \sigma^{-2}$,

$$\mathbb{E}\left[e^{\frac{t}{2}\left(\|Z\|_2^2 - n\sigma^2\right)}\right] = \frac{e^{-\frac{t}{2}n\sigma^2}}{(1-\sigma^2 t)^{n/2}} = \left(\frac{e^{-t\sigma^2/2}}{\sqrt{1-\sigma^2 t}}\right)^n,$$

Using the inequality

$$\frac{e^{-t}}{\sqrt{1-2t}} \le e^{2t^2} \quad \forall |t| < 1/4,$$

we have

$$\mathbb{E}\left[e^{\frac{t}{2}\left(\|Z\|_2^2 - n\sigma^2\right)}\right] \le \exp(-t^2\sigma^4/2) \quad \forall |t| < (2\sigma^2)^{-1}.$$

Combining the last inequality with lemma 26, we get eq. (14).

By Markov's inequality, we have for any $u > 0$ and $0 \le t < \left(2\sigma^2\right)^{-1}$,

$$\mathbb{P}\left(\|X\|_2^2 - n\sigma^2 \ge u\right) \le e^{-\frac{tu}{2} + \frac{t^2\sigma^4}{2}}.$$

Choosing $t = t^* := \frac{u}{2n\sigma^4} \wedge \frac{1}{2\sigma^2}$, we get

$$\mathbb{P}\left(\|X\|_2^2 - n\sigma^2 \ge u\right) \le \exp\left(-\frac{ut^*}{4}\right) = \exp\left[-\frac{1}{8}\min\left(\frac{u^2}{n\sigma^4}, \frac{u}{\sigma^2}\right)\right].$$

■

Appendix E. Proof of High-dimensional Hanson-Wright in Theorem 6

Proof We prove the one-side inequality and the other side is similar by replacing A with $-A$. Let

$$S = \sum_{i,j:i\neq j}^{n} a_{ij}\langle X_i, X_j\rangle. \tag{16}$$

Step 1: decoupling. Let $\iota_1, \dots, \iota_d \in \{0, 1\}$ be symmetric Bernoulli random variables, (i.e., $\mathbb{P}(\iota_i = 0) = \mathbb{P}(\iota_i = 1) = 1/2$) that are independent of $X_1, \dots, X_n$. Since

$$\mathbb{E}\left[\iota_i(1-\iota_i)\right] = \begin{cases} 0, & i = j, \\ 1/4, & i \neq j, \end{cases}$$

we have $S = 4\mathbb{E}_\iota[S_\iota]$, where

$$S_\iota = \sum_{i,j=1}^{n} \iota_i(1-\iota_j)a_{ij}\langle X_i, X_j\rangle$$

and the expectation $\mathbb{E}_\iota[\cdot]$ is the expectation taken with respect to the random variables ι_i. By Jensen's inequality and $\exp(\lambda x)$ is a convex function in x for any $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}[\exp(\lambda S)] \le \mathbb{E}_{X,\iota}[\exp(4\lambda S_\iota)].$$

Let $\Lambda_\iota = \{i \in [d] : \iota_i = 1\}$. Then we write

$$S_\iota = \sum_{i\in\Lambda_\iota}\sum_{j\in\Lambda_\iota^c} a_{ij}\langle X_i, X_j\rangle = \sum_{j\in\Lambda_\iota^c}\langle \sum_{i\in\Lambda_\iota} a_{ij}X_i, X_j\rangle.$$

Taking expectation over $(X_j)_{j\in\Lambda_\iota^c}$ (i.e., conditioning on $(\iota_i)_{i=1,\dots,d}$ and $(X_i)_{i\in\Lambda_\iota}$), it follows that

$$\mathbb{E}_{(X_j)_{j\in\Lambda_\iota^c}}[\exp(4\lambda S_\iota)] = \prod_{j\in\Lambda_\iota^c}\mathbb{E}_{(X_j)_{j\in\Lambda_\iota^c}}\left[e^{4\lambda\langle\sum_{i\in\Lambda_\iota} a_{ij}X_i, X_j\rangle}\right]$$

by the independence among $(X_j)_{j\in\Lambda_\iota}$. By the assumption that X_j are independent sub-Gaussian with mean zero, we have

$$\mathbb{E}_{(X_j)_{j\in\Lambda_\iota^c}}[\exp(4\lambda S_\iota)] \le \exp\left(\sum_{j\in\Lambda_\iota^c} 8\lambda^2 K_j^2 \|\sum_{i\in\Lambda_\iota} a_{ij}X_i\|^2\right) =: \exp\left(8\lambda^2\sigma_\iota^2\right).$$

Thus we get

$$\mathbb{E}_X[\exp(4\lambda S_\iota)] \le \mathbb{E}_X\left[\exp(8\lambda^2\sigma_\iota^2)\right].$$

Step 2: reduction to Gaussian random variables. For $j = 1,\dots,n$, let g_j be independent $N\left(0, 16K_j^2\boldsymbol{I}\right)$ random variables in $\mathbb{R}^m$ that are independent of $X_1,\dots,X_n$ and $\iota_1,\dots,\iota_n$. Define

$$T := \sum_{j\in\Lambda_\iota^c}\langle g_j, \sum_{i\in\Lambda_\iota} a_{ij}X_i\rangle.$$

Then, by the definition of Gaussian random variables in $\mathbb{R}^m$, we have

$$\begin{aligned}\mathbb{E}_g[\exp(\lambda T)] &= \prod_{j\in\Lambda_\iota^c}\mathbb{E}_g\left[e^{\langle g_j, \lambda\sum_{i\in\Lambda_\iota} a_{ij}X_i\rangle}\right]\\ &= \exp\left(8\lambda^2\sum_{j\in\Lambda_\iota^c} K_j^2\|\sum_{i\in\Lambda_\iota} a_{ij}X_i\|^2\right) = \exp\left(8\lambda^2\sigma_\iota^2\right)\end{aligned}$$

So it follows that

$$\mathbb{E}_X[\exp(4\lambda S_\iota)] \le \mathbb{E}_{X,g}[\exp(\lambda T)].$$

Since $T = \sum_{i\in\Lambda_\iota} \langle \sum_{j\in\Lambda_\iota^c} a_{ij} g_j, X_i\rangle$, by the assumption that X_i are independent sub-Gaussian with mean zero, we have

$$\mathbb{E}_{(X_i)_{i\in\Lambda_\iota}} [\exp(\lambda T)] \le \exp\left(\frac{\lambda^2}{2}\sum_{i\in\Lambda_\iota} K_i^2 \| \sum_{j\in\Lambda_\iota^c} a_{ij} g_j\|^2\right),$$

which implies that

$$\mathbb{E}_X [\exp(4\lambda S_\iota)] \le \mathbb{E}_g \left[\exp\left(\lambda^2 \tau_\iota^2/2\right)\right] \tag{17}$$

where $\tau_\iota^2 = \sum_{i\in\Lambda_\iota} K_i^2 \| \sum_{j\in\Lambda_\iota^c} a_{ij} g_j\|^2$. Note that τ_ι^2 is a random variable that depends on $(\iota_i)_{i=1}^d$ and $(g_j)_{j=1}^n$.

Step 3: diagonalization. We have $g_j = \sum_{k=1}^m \langle g_j, e_k\rangle e_k$ and

$$\begin{aligned}
\tau_\iota^2 &= \sum_{i\in\Lambda_\iota} K_i^2 \left\| \sum_{j\in\Lambda_\iota^c} a_{ij} g_j \right\|^2 = \sum_{i\in\Lambda_\iota} K_i^2 \left\| \sum_{k=1}^m \left(\sum_{j\in\Lambda_\iota^c} a_{ij}\langle g_j, e_k\rangle\right) e_k \right\|^2 \\
&= \sum_{k=1}^m \sum_{i\in\Lambda_\iota} \left(\sum_{j\in\Lambda_\iota^c} K_i a_{ij} \langle g_j, e_k\rangle\right)^2 \\
&= \sum_{k=1}^m \|P_\iota \tilde{A}(I-P_\iota) G_k\|^2
\end{aligned}$$

where the last second step follows from Parseval's identity. $G_{jk} := \langle g_j, e_k\rangle, j = 1,\dots,n$, are independent $N\left(0, 16K_j^2\right)$ random variables. $G_k = (G_{1k},\dots,G_{nk})^\top \in \mathbb{R}^n$. $\widetilde{A} = (\tilde{a}_{ij})_{i,j=1}^n$ with $\tilde{a}_{ij} = K_i a_{ij}$. Let $P_\iota \in \mathbb{R}^{n\times n}$ be the restriction matrix such that $P_{\iota,ii} = 1$ if $i \in \Lambda_\iota$ and $P_{\iota,ij} = 0$ otherwise.

Define normal random variables $Z_k = (Z_{1k},\dots,Z_{nk})^\top \sim N(0, I)$ for each $k = 1,\dots,m$. Then we have $G_k \overset{D}{=} \Gamma^{1/2} Z_k$ where $\Gamma = 16\,\mathrm{diag}(K_1^2,\dots,K_n^2)$.

Let $\tilde{A}_\iota := P_\iota \tilde{A}(I-P_\iota)$. Then by the rotational invariance of Gaussian distributions, we have

$$\sum_{k=1}^m \|\tilde{A}_\iota G_k\|^2 \overset{D}{=} \sum_{k=1}^m \|\tilde{A}_\iota \Gamma^{1/2} Z_k\|^2 \overset{D}{=} \sum_{k=1}^m \sum_{j=1}^n s_j^2 Z_{jk}^2$$

where $s_j^2, j = 1,2,\dots,n$ are the eigenvalues of $\Gamma^{1/2}\tilde{A}_\iota^\top \tilde{A}_\iota \Gamma^{1/2}$.

Step 4: bound the eigenvalues. It follows that

$$\max_{j\in[n]} s_j^2 = \|\tilde{A}_\iota \Gamma^{1/2}\|_2^2 \le 16K^4 \|A\|_2^2.$$

In addition, we also have

$$\sum_{j=1}^n s_j^2 = \mathrm{tr}(\Gamma^{1/2}\tilde{A}_\iota^\top \tilde{A}_\iota \Gamma^{1/2}) \le 16K^4 \|A\|_F^2$$

and $\sum_{k=1}^m \sum_{j=1}^n s_j^2 \leq 16mK^4\|A\|_F^2$. Invoking eq. (17), we get

$$\mathbb{E}_X\left[\exp\left(4\lambda S_\iota\right)\right] \leq \prod_{k=1}^m \prod_{j=1}^n \mathbb{E}_Z\left[\exp\left(\lambda^2 s_j^2 Z_{jk}^2/2\right)\right]$$

Since Z_{jk}^2 are i.i.d. χ_1^2 random variables with the moment generating function $\mathbb{E}\left[\exp\left(tZ_{jk}^2\right)\right] = (1-2t)^{-1/2}$ for $t < 1/2$, we have

$$\mathbb{E}_X\left[\exp\left(4\lambda S_\iota\right)\right] \leq \prod_{k=1}^m \prod_{j=1}^n \frac{1}{\sqrt{1-\lambda^2 s_j^2}} \quad \text{ if } \max_j \lambda^2 s_j^2 < 1.$$

Using $(1-z)^{-1/2} \leq \exp(z)$ for $z \in [0, 1/2]$, we get that if $\lambda^2 \max_j s_j^2 \leq 1/2$, i.e., $32K^4\|A\|_2^2\lambda^2 < 1$, then

$$\mathbb{E}_X\left[\exp\left(4\lambda S_\iota\right)\right] \leq \exp\left(\lambda^2 \sum_{k=1}^m \sum_{j=1}^n s_j^2\right) \leq \exp\left(16\lambda^2 mK^4\|A\|_F^2\right).$$

Note that the last inequality is uniform in ι. Taking expectation with respect to δ, we obtain that

$$\mathbb{E}_X\left[\exp\left(\lambda S\right)\right] \leq \mathbb{E}_{X,\iota}\left[\exp\left(4\lambda S_\iota\right)\right] \leq \exp\left(16\lambda^2 mK^4\|A\|_F^2\right)$$

whenever $|\lambda| < (4\sqrt{2}K^2\|A\|_2)^{-1}$.

Step 5: Conclusion. Now we have

$$\mathbb{P}(S \geq t) \leq \exp\left(-\lambda t + 16\lambda^2 mK^4\|A\|_F^2\right) \quad \text{ for } |\lambda| \leq \left(4\sqrt{2}K^2\|A\|_2\right)^{-1}$$

Optimizing in λ, we deduce that there exists a universal constant $C > 0$ such that

$$\mathbb{P}(S \geq t) \leq \exp\left[-\min\left(\frac{t^2}{64mK^4\|A\|_F^2}, \frac{t}{8\sqrt{2}K^2\|A\|_2}\right)\right].$$

■

Appendix F. Proof of Generalized High-dimensional Hanson-Wright in Theorem 14

Proof We prove the one-side inequality and the other side is similar by replacing A with $-A$. Let

$$S = \sum_{i,j:i\neq j}^n a_{ij}\langle b_i \odot X_i, b_j \odot X_j\rangle. \tag{18}$$

Step 1: decoupling. Let $\iota_1, \ldots, \iota_d \in \{0, 1\}$ be symmetric Bernoulli random variables, (i.e., $\mathbb{P}(\iota_i = 0) = \mathbb{P}(\iota_i = 1) = 1/2$) that are independent of $X_1, \ldots, X_n$. Since

$$\mathbb{E}\left[\iota_i(1-\iota_i)\right] = \begin{cases} 0, & i = j, \\ 1/4, & i \neq j, \end{cases}$$

we have $S = 4\mathbb{E}_\iota [S_\iota]$, where

$$S_\iota = \sum_{i,j=1}^{n} \iota_i(1-\iota_j)a_{ij}\langle b_i \odot X_i, b_j \odot X_j\rangle$$

and the expectation $\mathbb{E}_\iota [\cdot]$ is the expectation taken with respect to the random variables ι_i. By Jensen's inequality, we have

$$\mathbb{E}[\exp(\lambda S)] \le \mathbb{E}_{X,\iota}[\exp(4\lambda S_\iota)].$$

Let $\Lambda_\iota = \{i \in [d] : \iota_i = 1\}$. Then we write

$$S_\iota = \sum_{i\in\Lambda_\iota}\sum_{j\in\Lambda_\iota^c} a_{ij}\langle b_i \odot X_i, b_j \odot X_j\rangle = \sum_{j\in\Lambda_\iota^c}\langle \sum_{i\in\Lambda_\iota} a_{ij} b_i \odot b_j \odot X_i, X_j\rangle.$$

Taking expectation over $(X_j)_{j\in\Lambda_\iota^c}$ (i.e., conditioning on $(\iota_i)_{i=1,\dots,d}$ and $(X_i)_{i\in\Lambda_\iota}$), it follows that

$$\mathbb{E}_{(X_j)_{j\in\Lambda_\iota^c}}[\exp 4\lambda S_\iota] = \prod_{j\in\Lambda_\iota^c}\mathbb{E}_{(X_j)_{j\in\Lambda_\iota^c}}\left[e^{\lambda\langle\sum_{i\in\Lambda_\iota} a_{ij}b_i\odot b_j\odot X_i, X_j\rangle}\right]$$

by the independence among $(X_j)_{j\in\Lambda_\iota}$. By the assumption that X_j are independent sub-Gaussian with mean zero, we have

$$\mathbb{E}_{(X_j)_{j\in\Lambda_\iota^c}}[\exp 4\lambda S_\iota] \le \exp\left(\sum_{j\in\Lambda_\iota^c} 8\lambda^2 K_j^2 \|\sum_{i\in\Lambda_\iota} a_{ij} b_i \odot b_j \odot X_i\|^2\right) =: \exp\left(8\lambda^2\sigma_\iota^2\right).$$

Thus we get

$$\mathbb{E}_X[\exp(4\lambda S_\iota)] \le \mathbb{E}_X\left[\exp(8\lambda^2\sigma_\iota^2)\right].$$

Step 2: reduction to Gaussian random variables. For $j = 1,\dots,n$, let g_j be independent $N\left(0, 16K_j^2\boldsymbol{I}\right)$ random variables in $\mathbb{R}^m$ that are independent of $X_1,\dots,X_n$ and $\iota_1,\dots,\iota_n$. Define

$$T := \sum_{j\in\Lambda_\iota^c}\langle g_j, \sum_{i\in\Lambda_\iota} a_{ij} b_i \odot b_j \odot X_i\rangle.$$

Then, by the definition of Gaussian random variables in $\mathbb{R}^m$, we have

$$\begin{aligned}\mathbb{E}_g[\exp(\lambda T)] &= \prod_{j\in\Lambda_\iota^c}\mathbb{E}_g\left[e^{\langle g_j, \lambda\sum_{i\in\Lambda_\iota} a_{ij}b_i\odot b_j\odot X_i\rangle}\right] \\ &= \exp\left(8\lambda^2\sum_{j\in\Lambda_\iota^c} K_j^2\|\sum_{i\in\Lambda_\iota} a_{ij} b_i\odot b_j\odot X_i\|^2\right) = \exp\left(8\lambda^2\sigma_\iota^2\right)\end{aligned}$$

So it follows that

$$\mathbb{E}_X[\exp(4\lambda S_\iota)] \le \mathbb{E}_{X,g}[\exp(\lambda T)].$$

Since $T = \sum_{i\in\Lambda_\iota}\langle\sum_{j\in\Lambda_\iota^c} a_{ij}b_i\odot b_j\odot g_j, X_i\rangle$, by the assumption that X_i are independent sub-Gaussian with mean zero, we have

$$\mathbb{E}_{(X_i)_{i\in\Lambda_\iota}}\left[\exp\left(\lambda T\right)\right] \le \exp\left(\frac{\lambda^2}{2}\sum_{i\in\Lambda_\iota} K_i^2\|\sum_{j\in\Lambda_\iota^c} a_{ij}b_i\odot b_j\odot g_j\|^2\right),$$

which implies that

$$\mathbb{E}_X\left[\exp\left(4\lambda S_\iota\right)\right] \le \mathbb{E}_g\left[\exp\left(\lambda^2\tau_\iota^2/2\right)\right] \tag{19}$$

where $\tau_\iota^2 = \sum_{i\in\Lambda_\iota} K_i^2\|\sum_{j\in\Lambda_\iota^c} a_{ij}b_i\odot b_j\odot g_j\|^2$. Note that τ_ι^2 is a random variable that depends on $(\iota_i)_{i=1}^d$ and $(g_j)_{j=1}^n$.

Step 3: diagonalization. We have $g_j = \sum_{k=1}^m \langle g_j, e_k\rangle e_k$ and

$$\begin{aligned}\tau_\iota^2 &= \sum_{i\in\Lambda_\iota} K_i^2\left\|\sum_{j\in\Lambda_\iota^c} a_{ij}b_i\odot b_j\odot g_j\right\|^2 = \sum_{i\in\Lambda_\iota} K_i^2\left\|\sum_{k=1}^m\left(\sum_{j\in\Lambda_\iota^c} a_{ij}\langle b_i\odot b_j\odot g_j, e_k\rangle\right)e_k\right\|^2\\ &= \sum_{k=1}^m\sum_{i\in\Lambda_\iota}\left(\sum_{j\in\Lambda_\iota^c} K_i a_{ij}b_{ik}b_{jk}\langle g_j, e_k\rangle\right)^2\\ &= \sum_{k=1}^m \|P_\iota\tilde{A}(I-P_\iota)G_k\|^2\end{aligned}$$

where the last second step follows from Parseval's identity. $G_{jk} := \langle g_j, e_k\rangle, j = 1,\dots,n$, are independent $N\left(0, 16K_j^2\right)$ random variables. $G_k = (G_{1k},\dots,G_{nk})^\top \in \mathbb{R}^n$. $\widetilde{A}_k = (\tilde{a}_{ij}b_{ik}b_{jk})_{i,j=1}^n$ with $\tilde{a}_{ij} = K_i a_{ij}$. Let $P_\iota \in \mathbb{R}^{n\times n}$ be the restriction matrix such that $P_{\iota,ii} = 1$ if $i\in\Lambda_\iota$ and $P_{\iota,ij} = 0$ otherwise.

Define normal random variables $Z_k = (Z_{1k},\dots,Z_{nk})^\top \sim N(0,I)$ for each $k = 1,\dots,m$. Then we have $G_k \overset{D}{=} \Gamma^{1/2}Z_k$ where $\Gamma = 16\,\mathrm{diag}(K_1^2,\dots,K_n^2)$.

Let $\tilde{A}_{\iota,k} := P_\iota\tilde{A}_k(I-P_\iota)$. Then by the rotational invariance of Gaussian distributions, we have

$$\sum_{k=1}^m\|\tilde{A}_{\iota,k}G_k\|^2 \overset{D}{=} \sum_{k=1}^m\|\tilde{A}_{\iota,k}\Gamma^{1/2}Z_k\|^2 \overset{D}{=} \sum_{k=1}^m\sum_{j=1}^n s_{j,k}^2 Z_{jk}^2$$

where $s_{jk}^2, j = 1,2,\dots,n$ are the eigenvalues of $\Gamma^{1/2}\tilde{A}_{\iota,k}^\top\tilde{A}_{\iota,k}\Gamma^{1/2}$ for each $k = 1,\dots,m$.

Step 4: bound the eigenvalues. It follows that

$$\max_{j\in[n]} s_{j,k}^2 = \|\tilde{A}_{\iota,k}\Gamma^{1/2}\|_2^2 \le 16K^4\|A_k^b\|_2^2.$$

In addition, we also have

$$\sum_{j=1}^n s_{jk}^2 = \mathrm{tr}(\Gamma^{1/2}\tilde{A}_{\iota,k}^\top\tilde{A}_{\iota,k}\Gamma^{1/2}) \le 16K^4\|A_k^b\|_F^2$$

and $\sum_{k=1}^m \sum_{j=1}^n s_{jk}^2 \le 16K^4 \sum_{k=1}^m \|A_k^b\|_F^2$. Invoking eq. (19), we get

$$\mathbb{E}_X\left[e^{4\lambda S_\iota}\right] \le \prod_{k=1}^m \prod_{j=1}^n \mathbb{E}_Z\left[\exp\left(\lambda^2 s_{jk}^2 Z_{jk}^2/2\right)\right]$$

Since Z_{jk}^2 are i.i.d. χ_1^2 random variables with the moment generating function $\mathbb{E}\left[e^{tZ_{jk}^2}\right] = (1 - 2t)^{-1/2}$ for $t < 1/2$, we have

$$\mathbb{E}_X\left[e^{4\lambda S_\iota}\right] \le \prod_{k=1}^m \prod_{j=1}^n \frac{1}{\sqrt{1-\lambda^2 s_{jk}^2}} \qquad \text{if } \max_{j,k} \lambda^2 s_{jk}^2 < 1.$$

Using $(1-z)^{-1/2} \le e^z$ for $z \in [0, 1/2]$, we get that if $\lambda^2 \max_{j,k} s_{jk}^2 \le 1/2$, i.e., $32K^4 \max_k \|A_k^b\|_2^2 \lambda^2 < 1$, then

$$\mathbb{E}_X\left[e^{4\lambda S_\iota}\right] \le \exp\left(\lambda^2 \sum_{k=1}^m \sum_{j=1}^n s_{jk}^2\right) \le \exp\left(16\lambda^2 K^4 \sum_{k=1}^m \|A_k^b\|_F^2\right).$$

Note that the last inequality is uniform in ι. Taking expectation with respect to δ, we obtain that

$$\mathbb{E}_X\left[e^{\lambda S}\right] \le \mathbb{E}_{X,\iota}\left[e^{4\lambda S_\iota}\right] \le \exp\left(16\lambda^2 K^4 \sum_{k=1}^m \|A_k^b\|_F^2\right)$$

whenever $|\lambda| < (4\sqrt{2}K^2 \max_k \|A_k^b\|_2)^{-1}$.

Step 5: Conclusion. Now we have

$$\mathbb{P}(S \ge t) \le \exp\left(-\lambda t + 16\lambda^2 K^4 \sum_{k=1}^m \|A_k^b\|_F^2\right) \qquad \text{for } |\lambda| \le \left(4\sqrt{2}K^2 \max_k \|A_k^b\|_2\right)^{-1}.$$

Optimizing in λ, we deduce that there exists a universal constant $C > 0$ such that

$$\mathbb{P}(S \ge t) \le \exp\left[-\min\left(\frac{t^2}{64K^4 \sum_{k=1}^m \|A_k^b\|_F^2}, \frac{t}{8\sqrt{2}K^2 \max_k \|A_k^b\|_2}\right)\right].$$

■

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] Radoslaw Adamczak. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20(none):1 – 13, 2015. doi: 10.1214/ECP.v20-3829. URL https://doi.org/10.1214/ECP.v20-3829.

[3] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive approximation*, 28:253–263, 2008.

[4] Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer, 2003.

[5] Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, 52(12):5406–5425, 2006.

[6] Michael B Cohen, TS Jayram, and Jelani Nelson. Simple analyses of the sparse johnson-lindenstrauss transform. In *1st Symposium on Simplicity in Algorithms (SOSA 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[7] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[8] Vikranth Dwaracherla, Xiuyuan Lu, Morteza Ibrahimi, Ian Osband, Zheng Wen, and Benjamin Van Roy. Hypermodels for exploration. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryx6WgStPB.

[9] P Frankl and H Maehara. The johnson-lindenstrauss lemma and the sphericity of some graphs. *Journal of Combinatorial Theory, Series B*, 44(3):355–362, 1988. ISSN 0095-8956. doi: https://doi.org/10.1016/0095-8956(88)90043-3. URL https://www.sciencedirect.com/science/article/pii/0095895688900433.

[10] D. L. Hanson and F. T. Wright. A Bound on Tail Probabilities for Quadratic Forms in Independent Random Variables. *The Annals of Mathematical Statistics*, 42(3):1079 – 1083, 1971. doi: 10.1214/aoms/1177693335. URL https://doi.org/10.1214/aoms/1177693335.

[11] Mikael Møller Høgsgaard, Lion Kamma, Kasper Green Larsen, Jelani Nelson, and Chris Schwiegelshohn. Sparse dimensionality reduction revisited, 2023.

[12] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998.

[13] Thathachar S Jayram and David P Woodruff. Optimal bounds for johnson-lindenstrauss transforms and streaming problems with subconstant error. *ACM Transactions on Algorithms (TALG)*, 9(3):1–17, 2013.

[14] William B Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. In *Conference on Modern Analysis and Probability*, volume 26, pages 189–206. American Mathematical Society, 1984.

[15] Daniel Kane, Raghu Meka, and Jelani Nelson. Almost optimal explicit johnson-lindenstrauss families. In Leslie Ann Goldberg, Klaus Jansen, R. Ravi, and José D. P. Rolim, editors, *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 628–639, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-22935-0.

[16] Daniel M Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):1–23, 2014.

[17] Yingru Li, Jiawei Xu, Lei Han, and Zhi-Quan Luo. HyperAgent: A Simple, Scalable, Efficient and Provable Reinforcement Learning Framework for Complex Environments. In *Proceedings of the 41th International Conference on Machine Learning*, Proceedings of Machine Learning Research, 2024.

[18] Ziniu Li, Yingru Li, Yushun Zhang, Tong Zhang, and Zhi-Quan Luo. HyperDQN: A randomized exploration method for deep reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=X0nrKAXu7g-.

[19] Jiří Matoušek. On variants of the johnson–lindenstrauss lemma. *Random Structures & Algorithms*, 33(2):142–156, 2008.

[20] Shanmugavelayutham Muthukrishnan et al. Data streams: Algorithms and applications. *Foundations and Trends® in Theoretical Computer Science*, 1(2):117–236, 2005.

[21] Jelani Nelson. Sketching algorithms. Lecture Notes for "Sketching Algorithms"' Course, 2020. URL https://www.sketchingbigdata.org/fall20/lec/notes.pdf. Accessed: Nov. 1, 2023.

[22] Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=dZqcC1qCmB.

[23] Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Approximate thompson sampling via epistemic neural networks. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI '23. JMLR.org, 2023.

[24] Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18(none):1 – 9, 2013. doi: 10.1214/ECP.v18-2865. URL https://doi.org/10.1214/ECP.v18-2865.

[25] Maciej Skorski. Bernstein-type bounds for beta distribution. *Modern Stochastics: Theory and Applications*, 10(2):211–228, 2023.

[26] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and GittaEditors Kutyniok, editors, *Compressed Sensing: Theory and Applications*, page 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006.

[27] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

[28] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

[29] Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 1113–1120, 2009.

[30] 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.

[31] F. T. Wright. A Bound on Tail Probabilities for Quadratic Forms in Independent Random Variables Whose Distributions are not Necessarily Symmetric. *The Annals of Probability*, 1 (6):1068 – 1070, 1973. doi: 10.1214/aop/1176996815. URL https://doi.org/10.1214/aop/1176996815.