Near-Optimal Sample Complexity Bounds for Circulant Binary Embedding

Samet Oymak*

February 28, 2016

Abstract

Binary embedding is the problem of mapping points from a high-dimensional space to a Hamming cube in lower dimension while preserving pairwise distances. An efficient way to accomplish this is to make use of fast embedding techniques involving Fourier transform e.g. circulant matrices. While binary embedding has been studied extensively, theoretical results on fast binary embedding are rather limited. In this work, we build upon the recent literature to obtain significantly better dependencies on the problem parameters. A set of N points can be properly embedded into the Hamming cube $\{\pm 1\}^k$ with δ distortion, by using $k \sim \delta^{-3} \log N$ samples which is optimal in the number of points N and compares well with the optimal distortion dependency δ^{-2}. Our optimal embedding result applies in the regime $\log N \lesssim n^{1/3}$. Furthermore, if the looser condition $\log N \lesssim \sqrt{n}$ holds, we show that all but an arbitrarily small fraction of the points can be optimally embedded. We believe our techniques can be useful to obtain improved guarantees for other nonlinear embedding problems.

1 Introduction

Binary embedding problem aims to map a set of points in a high-dimensional space to the Hamming cube in a lower dimension. The task is preserving the distances between the points while keeping embedding dimension as small as possible. A common approach to accomplish this task is applying a random map to the data. In particular, given a point $\boldsymbol{x} \in \mathbb{R}^n$, we first apply a linear transformation $\boldsymbol{x} \to \boldsymbol{A}\boldsymbol{x} \in \mathbb{R}^k$ and then apply the discretization $\boldsymbol{A}\boldsymbol{x} \to \text{sgn}(\boldsymbol{A}\boldsymbol{x})$. Given a set S and distortion level $\delta > 0$, we are interested in ensuring that for all $\boldsymbol{x}, \boldsymbol{y} \in S$, $\boldsymbol{A}$ satisfies

$$|k^{-1}\|\text{sgn}(\boldsymbol{A}\boldsymbol{x}), \text{sgn}(\boldsymbol{A}\boldsymbol{y})\|_H - \text{ang}(\boldsymbol{x}, \boldsymbol{y})| \leq \delta.$$

Here, $\|\cdot,\cdot\|_H$ is the Hamming distance between two $0, 1$ vectors and $\text{ang}(\cdot)$ is the angular distance which returns the minimum angle between two points normalized by π. Often we are interested in embedding a large set of points $S = \{\boldsymbol{v}_i\}_{i=1}^N$ or a continuous set such as a subspace. An important aspect of the embedding problems is the tradeoff between the number of points N and the embedding dimension m. For linear embedding, classical Johnson-Lindenstrauss (JL) Lemma guarantees that by using $k \approx \delta^{-2} \log N$ samples, N points can be embedded with δ distortion. More recently, this tradeoff attracted significant attention for the binary embedding problem. Specifically, by choosing $\boldsymbol{A}$ to be a Gaussian matrix, it can be trivially shown that one can achieve a good binary embedding under the same assumption of $k \approx \delta^{-2} \log N$. This arguments have also been extended to arbitrary (e.g. continuous) sets which are of interest for sparse estimation problems.

While the results on dense Gaussian matrices are valuable, for most applications we are interested in faster projections where embedding can be done in near-linear time. Such projections make use of fast matrix multiplications such as the Fourier Transform followed by random diagonal modulations and are broadly called Fast Johnson-Lindenstrauss Transform (FJLT). In this work, we focus on circulant embedding matrices where projection matrix $\boldsymbol{A}$ is given by $\boldsymbol{A} = \boldsymbol{R}\boldsymbol{C}_{\boldsymbol{h}}\text{diag}(\boldsymbol{r})$. Here,

*Google Inc. 1600 Amphitheatre Parkway, Mountain View, CA 94043.

- $\boldsymbol{R} \in \mathbb{R}^{k \times n}$ is the restriction operator that selects k rows out of n uniformly at random.
- $\boldsymbol{h}, \boldsymbol{r} \in \mathbb{R}^n$ are independent vectors with independent standard normal entries.
- $\boldsymbol{C_h}$ is a circulant matrix whose first row is equal to $\boldsymbol{h}^*$.
- $\mathrm{diag}(\boldsymbol{r})$ is a diagonal matrix obtained from the vector $\boldsymbol{r}$.

The theoretical results for fast binary embedding techniques are rather limited [13, 22, 23]. Related to us, very recently Yu et al. provided an analysis of circulant projections. Loosely speaking, the authors show that by using $k \sim \log^2 N$ samples, binary embedding with small distortion is possible as long as $\log N \lesssim n^{1/6}$. Another related work connected to nonlinear embedding is due to Le et al. [11]. Here, the authors speed up Kernel approximation [17] by making use of FJLT however the number of required Fourier features scale quadratically due to suboptimal concentration bounds. A natural question is whether circulant projections can achieve the optimal binary embedding guarantees. In this work, we answer this question positively. We show that using $k \sim \log N$ samples, binary embedding via circulant matrices will be successful as long as $\log N \lesssim n^{1/2}$. This shows that Fast JL Transform not only works well for linear embedding but also for highly nonlinear problems and the embedding behavior is essentially same.

Contributions: Specifically, we have two sets of results. Our first set of results consider embedding with circulant projections and the associated theorem has a dependency on the coherence of the set $\{\boldsymbol{v}_i\}_{i=1}^N$. When the points are not spiky, (i.e. small infinity norm), the optimal embedding works for a larger regime of N. For maximally incoherent sets we can allow $\log N \lesssim n^{1/2}$. Our second result is a corollary of the first one and attempts to remove the dependence on incoherence. This is done by applying an additional layer of randomness $\boldsymbol{x} \to \boldsymbol{H}\mathrm{diag}(\boldsymbol{b})\boldsymbol{v}$ where $\boldsymbol{H}$ is the Hadamard transform and $\mathrm{diag}(\boldsymbol{b})$ is a diagonal matrix with independent Rademacher diagonal entries. The overall embedding takes the form $\boldsymbol{v} \to \mathrm{sgn}(\boldsymbol{RC_h}\mathrm{diag}(\boldsymbol{r})\boldsymbol{H}\mathrm{diag}(\boldsymbol{b})\boldsymbol{v})$. Observe that all matrix multiplications are still near-linear time. This model makes no assumption on the set $\{\boldsymbol{v}_i\}_{i=1}^N$ and optimal embedding is possible as soon as $\log N \lesssim n^{1/3}$. Furthermore, if $\log N \lesssim \sqrt{n}$, fast and optimal binary embedding still succeeds for all but arbitrarily small fraction of the points.

1.1 Related Literature

Binary embedding with dense Gaussian matrices is a rather well studied problem. Guarantees for finite set of points can be obtained by applying a standard Chernoff bound. Embedding continuous sets is a more challenging problem and it is studied in a series of papers [2, 8, 13, 15, 16] with results mostly restricted to Gaussian ensemble. Much less is known for the fast embedding techniques that make use of Fourier or Hadamard transforms. We can split the existing results in this direction into two groups.

• **Fast JL embedding followed by dense (two-layer) map:** This map is given by $\boldsymbol{x} \to \mathrm{sgn}(\boldsymbol{GF}\mathrm{diag}(\boldsymbol{b})\boldsymbol{x})$ where $\boldsymbol{G} \in \mathbb{R}^{k \times k'}$ is a dense Gaussian matrix, $\boldsymbol{F} \in \mathbb{R}^{k' \times n}$ is the subsampled Discrete Fourier Transform matrix, $\boldsymbol{h}$ has independent Rademacher entries, and $k \approx k' \approx \mathcal{O}(\log N)$. This first applies a fast linear dimensionality reduction to linearly embed points to the lower dimensionality space $\mathbb{R}^{k'}$. Next, we use a dense Gaussian matrix to obtain a binary embedding guarantee. This approach is not computationally efficient as soon as $k \gtrsim \mathcal{O}(\sqrt{n})$ as dense Gaussian multiplication becomes more expensive than the Fourier transform. In [22], Yi et al. propose a related but more efficient algorithm by replacing $\boldsymbol{G}$ with a more involved procedure involving Toeplitz matrices.

The optimal embedding bound of the present paper applies in the regime $k \leq \sqrt{n}$ which shows that circulant projections perform as good as two-layer maps computationally (both require $\mathcal{O}(n \log n)$ in the regime $k \leq \mathcal{O}(\sqrt{n})$). However, the proposed approach is much simpler and easily extends to the regime $k \geq \mathcal{O}(\sqrt{n})$ in an efficient manner (albeit without proof).

• **Simply use Fast JL embedding:** We simply apply a Fast JL Transform by using a circulant matrix. The map we consider has the form $\boldsymbol{x} \to \mathrm{sgn}(\boldsymbol{RC_h}\mathrm{diag}(\boldsymbol{r}))$ where $\boldsymbol{R} \in \mathbb{R}^{k \times n}$ is the subsampling operator, $\boldsymbol{C_h}$ is a circulant matrix and $\boldsymbol{r}, \boldsymbol{h}$ are vectors with iid $\mathcal{N}(0,1)$ entries. Since circulant matrices are diagonalized by the Discrete Fourier Transform, the computational complexity of embedding is always $\mathcal{O}(n \log n)$ independent of the sample size $k \leq n$. Yu et al. [23] very recently provided an analysis of this map with rigorous

sample complexity bounds. While their result has significant dependency on the set geometry, under best circumstances, they show that $k \gtrsim \log^2 N$ samples are sufficient for successful embedding as long as $\log N \lesssim n^{1/6}$. As an example of geometric dependence, the results of [23] depends on the maximal correlation of the point set $\sup_{i\neq j} |\boldsymbol{v}_i^* \boldsymbol{v}_j|$ and becomes arbitrarily suboptimal as this number approaches 1. In particular, they cannot allow points that are close to each other. There are also several works on the applications of fast binary projections in large scale image retrieval and hashing algorithms [6, 21, 24].

2 Main results

Suppose we are given N unit vectors in $\mathbb{R}^n$ namely $\{\boldsymbol{v}_i\}_{i=1}^N$. Our task is mapping this points to a low-dimensional Hamming cube in $\mathbb{R}^k$ while preserving the distances. We are interested in ensuring that for all $1 \leq i, j \leq N$, $\boldsymbol{A}$ satisfies

$$|k^{-1}\|\mathrm{sgn}(\boldsymbol{A}\boldsymbol{v}_i), \mathrm{sgn}(\boldsymbol{A}\boldsymbol{v}_j)\|_H - \mathrm{ang}(\boldsymbol{v}_i, \boldsymbol{v}_j)| \leq \delta.$$

As a geometric feature, we shall make use of the coherence of the set which is defined as

$$\rho_{cross} = \max\{ \sup_{1\leq i\leq N} \|\boldsymbol{v}_i\|_{\ell_\infty}, \quad \sup_{1\leq i\neq j\leq N} \frac{\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_{\ell_\infty}}{\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_{\ell_2}} \}.$$

For our results to work, we make the following assumptions on N, k, n and the coherence parameter.

Condition 2.1 *There exists sufficiently large nonnegative constants c_1, c_2, c_3[1], such that*

1. $k > c_1 \delta^{-3} \log N$.
2. $c_2 \delta k \rho_{cross} \log n < 1$.
3. $\delta \geq c_3 k \rho_{cross}$.

Observe that in the maximally incoherent case ($\rho_{cross} = \mathcal{O}(n^{-1/2})$), we can pick $\delta = o(1)$, $k = \mathcal{O}((\log n)^{-1} n^{1/2})$ and $\log N = \mathcal{O}(\delta^{-3} k)$. Hence, our optimal embedding result applies up to $\mathcal{O}(\sqrt{n})$ as the embedding dimension. Our main result is on fast binary embedding of finite set of points with near-optimal embedding dimensions and is stated in the next theorem.

Theorem 2.2 *Let $\boldsymbol{A} = \boldsymbol{R}\boldsymbol{C}_{\boldsymbol{h}} diag(\boldsymbol{r}) \in \mathbb{R}^k$ be a circulant projection as described above. Under the assumptions of Condition 2.1, with probability $1 - \exp(-c_4 \delta^3 k)$, for all $\boldsymbol{x}, \boldsymbol{y} \in \{\boldsymbol{v}_i\}_{i=1}^N$, we have that*

$$|k^{-1}\|sgn(\boldsymbol{A}\boldsymbol{x}), sgn(\boldsymbol{A}\boldsymbol{y})\|_H - ang(\boldsymbol{x}, \boldsymbol{y})| \leq \delta.$$

This result applies to arbitrary set of points; however, it depends on the incoherence of the set ρ_{cross}. One can get rid of this dependency by applying an additional layer of randomization. In particular, let $\boldsymbol{H}$ be a Hadamard matrix of size n and let $\boldsymbol{b} \in \mathbb{R}^n$ be a vector with independent Rademacher entries. If n is not a power of 2, we can simply zero-pad the vectors. Consider the map

$$\boldsymbol{A}_{\boldsymbol{H}} = \boldsymbol{A}\boldsymbol{H}\mathrm{diag}(\boldsymbol{b}) = \boldsymbol{R}\boldsymbol{C}_{\boldsymbol{h}}\mathrm{diag}(\boldsymbol{r})\boldsymbol{H}\mathrm{diag}(\boldsymbol{b}).$$

For this map, we have the following result that is incoherence-free.

Theorem 2.3 *Consider the binary embedding via the operator $\boldsymbol{x} \to sgn(\boldsymbol{A}_{\boldsymbol{H}}\boldsymbol{x})$. There exists universal constants $c, C > 0$ such that following holds. Suppose*

$$\log N \leq c\delta^2 (\log n)^{-1} n^{1/3}.$$

Then, with probability $1 - \exp(-c \log N)$, the point set $\boldsymbol{w}_i = \boldsymbol{H} diag(\boldsymbol{b})\boldsymbol{v}_i$ obeys the incoherence condition with $\rho_{cross} = C\delta(\log n)^{-1/2} n^{-1/3}$. Consequently, as soon as $k \geq c_1 \delta^{-3} \log N$, with probability $1 - \exp(-c_4 \delta^3 k)$,

$$|k^{-1}\|sgn(\boldsymbol{A}_{\boldsymbol{H}}\boldsymbol{x}), sgn(\boldsymbol{A}_{\boldsymbol{H}}\boldsymbol{y})\|_H - ang(\boldsymbol{x}, \boldsymbol{y})| \leq \delta.$$

[1] $c, C, \{c_i, C_i\}_{i\geq 0}, c', C'$ will be used to denote absolute constants that may vary from line to line.

Proof This result follows from the fact that the set of points obtained by the map $\boldsymbol{v}_i \to \boldsymbol{H}\boldsymbol{b}\boldsymbol{v}_i$ has desirable geometric features (small ρ_{cross}) with high probability. In particular, combine Theorem 2.2 with Lemma B.2. ■

Finally, the next result shows that one can optimally embed most of the points as long as $\log N \lesssim \mathcal{O}(\sqrt{n})$.

Theorem 2.4 *Consider the binary embedding via the operator $\boldsymbol{x} \to sgn(\boldsymbol{A}_{\boldsymbol{H}}\boldsymbol{x})$. There exists universal constants $c, C > 0$ such that following holds. Suppose*

$$\log N \le c\delta^3 (\log n)^{-2} n^{1/2}.$$

Then, with probability $1 - n^{-2}$ (over $\boldsymbol{H}$), there exists $S_{good} \subseteq \{\boldsymbol{v}_i\}_{i=1}^N$ such that

$$|S_{good}| \ge (1 - c_5 n^{-2})N, \quad \text{and} \quad \text{for all } \boldsymbol{v} \in S_{good} : \|\boldsymbol{H}\, diag(\boldsymbol{b})\boldsymbol{v}\|_{\ell_\infty} \le \rho_{cross}$$

where $\rho_{cross} = C\sqrt{\log n / n}$. Consequently, as soon as $k \ge c_1 \delta^{-3} \log N$, with probability $1 - n^{-2} - \exp(-c_4 \delta^3 k)$, all $\boldsymbol{x}, \boldsymbol{y}$ chosen from S_{good} obeys

$$|k^{-1} \|sgn(\boldsymbol{A}_{\boldsymbol{H}}\boldsymbol{x}), sgn(\boldsymbol{A}_{\boldsymbol{H}}\boldsymbol{y})\|_H - ang(\boldsymbol{x}, \boldsymbol{y})| \le \delta.$$

Proof This result follows from the fact that all but a small fraction of the set of points obtained by the map $\boldsymbol{v}_i \to \boldsymbol{H}\boldsymbol{b}\boldsymbol{v}_i$ has desirable geometric features (small ρ_{cross}) with high probability. In particular, combine Theorem 2.2 with Lemma B.3. Pick $p = n^{-2}$ in Lemma B.3. ■

3 Conclusions and Open Problems

In this work, we showed that fast binary embedding with near optimal dimensions are possible. In particular, our embedding bounds are consistent with the state of the art results for linear embedding, indicating that fast binary embedding is feasible under identical conditions to fast linear embedding such as [1,10,14]. This is the first such result for fast binary embedding and significantly improves over related literature (e.g. [11,23]). We believe the tools developed in this paper broadly applies to nonlinear embedding tasks. For instance, our argument may be used to improve the concentration estimates of Fastfood features [11] which is a popular fast kernel approximation technique. Our embedding result holds for finite set of points and it is of interest to extend this work to continuous sets. A weakness of our result is the fact that the embedding dimension scales up to $\mathcal{O}(\sqrt{n})$ which limits the number of points to $\log N \lesssim \mathcal{O}(\sqrt{n})$. This work opens up several research directions.

- **Fast embedding in linear regime:** Does fast binary embedding work with embedding dimension n? In other words, can we pick $k \sim \mathcal{O}(n)$ to embed $N \sim \exp(\mathcal{O}(k))$ points? If not, is there a fundamental bottleneck at $k \sim \mathcal{O}(\sqrt{n})$?
- **Practical considerations:** Our result on circulant embedding $\boldsymbol{C}_{\boldsymbol{h}}\text{diag}(\boldsymbol{r})$ requires $\boldsymbol{h}$ and $\boldsymbol{r}$ to have Gaussian entries. We believe $\boldsymbol{r}$ can have Rademacher entries without impacting the performance. It would possibly improve the performance as the operator $\boldsymbol{v} \to \text{diag}(\boldsymbol{r})\boldsymbol{v}$ preserves the inner products when $\boldsymbol{r}$ is Rademacher. Furthermore, it is not clear whether the incoherence assumption in Theorem 2.2 is necessary. Numerical results of prior work [23,24] indicates that the map $\boldsymbol{v} \to \text{sign}(\boldsymbol{C}_{\boldsymbol{h}}\text{diag}(\boldsymbol{r})\boldsymbol{v})$ works well which suggests that we may not need additional randomization via Hadamard transform. This would allow us to discard one layer of the embedding, namely, $\boldsymbol{v} \to \boldsymbol{H}\text{diag}(\boldsymbol{b})\boldsymbol{v}$.
- **General nonlinear embedding:** With a minor modification of our analysis, it is possible to obtain fast embedding bounds for a more general model $f(\boldsymbol{A}\boldsymbol{x})$ where f is a function that apply pointwise. The important use cases would be to replace $\text{sgn}(\cdot)$ function with a general function such as quantization, ReLU, sigmoid etc [5,7]. It would also be of interest to investigate quadratic samples arising in phase retrieval [4,9].

- **Embedding of continuous sets:** Our current results apply to finite set of points however it is of interest to embed continuous sets such as subspaces or sparse and low-rank manifolds. While this problem is studied for dense Gaussian embedding matrices, we believe similar results can be obtained for fast embedding matrices by building on this work and [23].

The rest of the paper is dedicated to the proof of our main result Theorem 2.2. Before going into technical details, we introduce the necessary notation. Given a vector $\boldsymbol{x} \in \mathbb{R}^n$, let $s_i(\boldsymbol{x})$ be the vector obtained by shifting entries of $\boldsymbol{x}$ by i position, i.e. $s_i(\boldsymbol{x})$. In particular, jth entry of $s_i(x)$ is same as $(i+j)$th entry of $\boldsymbol{x}$ modulo n. $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ returns minimum and maximum singular values of a matrix respectively. $\|\cdot\|$ denotes the spectral norm of a matrix and is same as $\sigma_{\max}(\cdot)$. $\mathrm{diag}(\cdot)$ returns a diagonal matrix from a vector input or returns the vector of diagonal entries of a matrix. $c, C, \{c_i, C_i\}_{i\geq 0}, c', C'$ will be used to denote absolute constants. For nonzero $\boldsymbol{x}$, $\bar{\boldsymbol{x}} = \boldsymbol{x}/\|\boldsymbol{x}\|_{\ell_2}$ and $\bar{0} = 0$. Throughout this work, Hadamard and Discrete Fourier Transform matrices are normalized to be unitary. A standard Gaussian vector obeys the distribution $\mathcal{N}(0, \mathbf{I})$. S denotes a subset of $\{1, 2, \ldots, n\}$ obtained by picking k-elements uniformly at random without replacement. Define direct coherence to be $\rho_{direct} = \sup_{1\leq i\leq N} \|\boldsymbol{v}_i\|_{\ell_\infty}$. Let $\theta > 0$ be the smallest angle between these points namely $\min_{i\neq j} \mathrm{ang}(\boldsymbol{v}_i, \boldsymbol{v}_j)$. It is trivial to show that the cross coherence can be bounded as $\rho_{cross} \leq 2\sin(\theta)^{-1}\rho_{direct}$. ℓth entry of a vector of size n is same as "$\ell \pmod n$"th entry of the vector.

4 Controlling the Conditioning of the Projection Matrix

To simplify our notation, we shall assume that $N \geq n$. $n > N$ case can be recovered by setting $n = N$ in our main result. Let $\{r_i\}_{i=1}^k$ be distinct numbers selected from the set $\{1, 2, \ldots, n\}$ uniformly at random.

Definition 4.1 (Random shift vectors) *Let $\boldsymbol{x} \in \mathcal{S}^{n-1}$ and let $\boldsymbol{r} \in \mathbb{R}^n$ be a standard Gaussian vector. Random shift vectors of $\boldsymbol{x}$ are a set of random vectors $\{\boldsymbol{X}_i\}_{i=1}^n$ such that $\boldsymbol{X}_i = s_i(\mathrm{diag}(\boldsymbol{r})\boldsymbol{x})$ for $0 \leq i \leq n-1$. Define $\boldsymbol{Y}_i$ in the identical manner given vector $\boldsymbol{y}$ for the same choice of $\boldsymbol{r}$.*

The following theorem summarizes the main result of this section by providing a spectral norm bound on subsampled random shift vectors.

Theorem 4.2 *Pick unit vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ satisfying $\boldsymbol{x}^*\boldsymbol{y} = 0$ and $\|\boldsymbol{x}\|_{\ell_\infty}, \|\boldsymbol{y}\|_{\ell_\infty} \leq \rho$. Form a matrix $\boldsymbol{M} \in \mathbb{R}^{n\times 2k}$ by picking the same k vectors $\{\boldsymbol{X}_{r_i}\}_{i=1}^k, \{\boldsymbol{Y}_{r_i}\}_{i=1}^k$ from each of $\{\boldsymbol{X}_i\}_{i=1}^n$ and $\{\boldsymbol{Y}_i\}_{i=1}^n$ without replacement uniformly at random and then stacking next to each other. With probability $1 - 2\exp(-\delta^2 k)$ (over $\boldsymbol{r}$ and selection of $\{\boldsymbol{X}_{r_i}\}_{i=1}^k$'s), we have that*

$$\sigma_{\max}(\boldsymbol{M}^*\boldsymbol{M} - \mathbf{I}) \leq C\delta k\rho \log n. \tag{4.1}$$

Corollary 4.3 *Let $\boldsymbol{x}, \boldsymbol{y}$ be unit vectors obeying $ang(\boldsymbol{x}, \boldsymbol{y}) = \theta$ and $\|\boldsymbol{x}\|_{\ell_\infty}, \|\boldsymbol{y}\|_{\ell_\infty} \leq \rho$. Form a matrix $\boldsymbol{M} \in \mathbb{R}^{n\times 2k}$ by picking the same k vectors $\{\boldsymbol{X}_{r_i}\}_{i=1}^k, \{\boldsymbol{Y}_{r_i}\}_{i=1}^k$ from each of $\{\boldsymbol{X}_i\}_{i=1}^n$ and $\{\boldsymbol{Y}_i\}_{i=1}^n$ without replacement uniformly at random and then stacking next to each other. With probability $1 - 6\exp(-\delta^2 k)$ (over $\boldsymbol{r}$ and selection of $\{\boldsymbol{X}_{r_i}\}_{i=1}^k$'s), we have that*

$$\sigma_{\max}(\boldsymbol{M}^*\boldsymbol{M} - \mathbf{I}_\theta) \leq C\delta k\rho \log n. \tag{4.2}$$

where $\mathbf{I}_\theta \in \mathbb{R}^{2k\times 2k}$ is given by the matrix $\begin{bmatrix} \mathbf{I}_k & \cos(\theta)\mathbf{I}_k \\ \cos(\theta)\mathbf{I}_k & \mathbf{I}_k \end{bmatrix}$.

Proof Proof is based on Theorem 4.2. Consider the decomposition $\boldsymbol{y} = \cos(\theta)\boldsymbol{x} + \sin(\theta)\boldsymbol{y}'$ where $\boldsymbol{x}^*\boldsymbol{y}' = 0$. Denote the k chosen columns $\{\boldsymbol{X}_{r_i}\}_{i=1}^k$ by the matrix $\boldsymbol{X} \in \mathbb{R}^{n\times k}$ and the corresponding matrix to $\{\boldsymbol{Y}_{r_i}\}_{i=1}^k$ by $\boldsymbol{Y} \in \mathbb{R}^{n\times k}$ and set $\boldsymbol{Y}' = \sin(\theta)^{-1}(\boldsymbol{Y} - \cos(\theta)\boldsymbol{X})$. Now observe that

$$\boldsymbol{X}^*\boldsymbol{Y} = \cos(\theta)\boldsymbol{X}^*\boldsymbol{X} + \sin(\theta)\boldsymbol{X}^*\boldsymbol{Y}'.$$

Using Theorem 4.2 on $\boldsymbol{X}^*\boldsymbol{X}$, we know that for an absolute constant $c_1 > 0$, with probability $1 - 2\exp(-\delta^2 k)$

$$\|\boldsymbol{X}^*\boldsymbol{X} - \mathbf{I}\| \le c_1 k\rho \log n.$$

Next we apply Theorem 4.2 to the matrix $[\boldsymbol{X}\ \boldsymbol{Y}']$. From Lemma A.4, $\boldsymbol{y}'$ obeys $\|\boldsymbol{y}'\|_{\ell_\infty} \le \frac{2\rho}{\sin(\theta)}$. Consequently, we have the spectral norm estimate

$$\|\boldsymbol{X}^*\boldsymbol{Y}'\| \le \|[\boldsymbol{X}\ \boldsymbol{Y}']^*[\boldsymbol{X}\ \boldsymbol{Y}']\| \le \sin(\theta)^{-1} c_1 \delta k\rho \log n.$$

Combining these, and using triangle inequality, we obtain

$$\|\boldsymbol{X}^*\boldsymbol{Y} - \cos(\theta)\mathbf{I}\| \le \sin(\theta)\|\boldsymbol{X}^*\boldsymbol{Y}'\| + \cos(\theta)\|\boldsymbol{X}^*\boldsymbol{X} - \mathbf{I}\| \le 2c_1\delta k\rho\log n + \cos(\theta)c_1\delta k\rho\log n \le 3c_1\delta k\rho\log n.$$

Finally, we need to estimate the remaining submatrices. In particular, direct applications of Theorem 4.2 yields

$$\max\{\|\boldsymbol{X}^*\boldsymbol{X} - \mathbf{I}\|, \|\boldsymbol{Y}^*\boldsymbol{Y} - \mathbf{I}\|\} \le c_1\delta k\rho\log n.$$

Combining these estimates and representing $\boldsymbol{M}$ as 4 $k\times k$ submatrix involving $\boldsymbol{X}^*\boldsymbol{X}, \boldsymbol{Y}^*\boldsymbol{Y}, \boldsymbol{X}^*\boldsymbol{Y}, \boldsymbol{Y}^*\boldsymbol{X}$, we find

$$\|\boldsymbol{M}^*\boldsymbol{M} - \mathbf{I}_\theta\| \le \|\boldsymbol{Y}^*\boldsymbol{Y} - \mathbf{I}\| + \|\boldsymbol{X}^*\boldsymbol{X} - \mathbf{I}\| + 2\|\boldsymbol{X}^*\boldsymbol{Y} - \cos(\theta)\mathbf{I}\| \le 8c_1\delta k\rho\log n.$$

Using a union bound, this final event happens with probability $1 - 6\exp(-\delta^2 k)$. ∎

4.1 Proof of Theorem 4.2

Theorem 4.4 (Hanson-Wright Theorem [18]) *Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{g} \in \mathbb{R}^n$ be a standard Gaussian vector. There exists a constant $c > 0$ such that*

$$\mathbb{P}(|\boldsymbol{g}^*\boldsymbol{A}\boldsymbol{g} - \mathbb{E}[\boldsymbol{g}^*\boldsymbol{A}\boldsymbol{g}]| \ge t) \le 2\exp\left(-c\left\{\frac{t^2}{\|\boldsymbol{A}\|_F^2}, \frac{t}{\|\boldsymbol{A}\|}\right\}\right).$$

The following lemma follows as a corollary of Hanson-Wright Theorem.

Lemma 4.5 *Let $\boldsymbol{x}, \boldsymbol{y}$ be two unit vectors where $\|\boldsymbol{x}\|_{\ell_\infty}, \|\boldsymbol{y}\|_{\ell_\infty} \le \rho$ and $\boldsymbol{x}^*\boldsymbol{y} = 0$. Let $\boldsymbol{g}$ be a standard Gaussian vector and $\boldsymbol{G} = diag(\boldsymbol{g})$. Then, the followings hold*

For all $1 \le i \ne j \le n$

$$\max\{\mathbb{P}(|s_i(\boldsymbol{G}\boldsymbol{x})^* s_j(\boldsymbol{G}\boldsymbol{y})| > t), \mathbb{P}(|s_i(\boldsymbol{G}\boldsymbol{x})^* s_j(\boldsymbol{G}\boldsymbol{x})| > t), \mathbb{P}(|\,\|\boldsymbol{G}\boldsymbol{x}\|_{\ell_2}^2 - 1| > t)\} \le 2\exp(-c\min\{\frac{t}{\rho^2}, \frac{t^2}{\rho^2}\}).$$

If $\|\boldsymbol{x}\|_{\ell_\infty} \le \rho, \|\boldsymbol{y}\|_{\ell_\infty} \le \rho'$ we additionally have $\mathbb{P}(|s_i(\boldsymbol{G}\boldsymbol{x})^* s_j(\boldsymbol{G}\boldsymbol{y})| > t) \le 2\exp(-c\min\{\frac{t}{\rho\rho'}, \frac{t^2}{\rho^2}\})$.

Proof The proofs are based on Hanson-Wright Theorem. $s_i(\boldsymbol{G}\boldsymbol{x})^* s_j(\boldsymbol{G}\boldsymbol{y})$ can be viewed as $\boldsymbol{g}^*\boldsymbol{M}\boldsymbol{g}$ where $\boldsymbol{M}$ is a weighted permutation matrix whose $(\ell + i, \ell + j)$th entry is of the form $\boldsymbol{x}_{\ell+i}\boldsymbol{y}_{\ell+j}$ for $1 \le \ell \le n$ and whose remaining entries are 0. Consequently, this matrix has maximum spectral norm ρ^2 and maximum Frobenius norm of ρ. $\mathbb{E}[s_i(\boldsymbol{G}\boldsymbol{x})^* s_j(\boldsymbol{G}\boldsymbol{y})]$ is clearly 0 when $i \ne j$ and for $i = j$ it is equal to $\boldsymbol{x}^*\boldsymbol{y} = 0$. Hence, Hanson-Wright yields the desired bound.

For the second and third relations, identical argument applies. We additionally use the fact that $\mathbb{E}[\|\boldsymbol{G}\boldsymbol{x}\|_{\ell_2}^2] = 1$. The last statement follows by modifying the spectral norm estimate from ρ^2 to $\rho\rho'$. ∎

Theorem 4.6 *Pick unit vectors satisfying $\boldsymbol{x}^*\boldsymbol{y} = 0$, $\|\boldsymbol{x}\|_{\ell_\infty}, \|\boldsymbol{y}\|_{\ell_\infty} \le \rho$, and $\rho < c_0 \max\{\delta, (\log n)^{-1/2}\}$ for a sufficiently small constant $c_0 > 0$. Define the matrix $\boldsymbol{A} = [\boldsymbol{X}_1\ \ldots\ \boldsymbol{X}_n\ \boldsymbol{Y}_1\ \ldots\ \boldsymbol{Y}_n]$. Form the matrix $\boldsymbol{M} \in \mathbb{R}^{n\times 2k}$ by picking the same k columns from $\{\boldsymbol{X}_i\}_{i=1}^n$ and $\{\boldsymbol{Y}_i\}_{i=1}^n$ uniformly at random. Denote the indices of these columns (i.e. support set) by S which is a subset of $\{1, 2, \ldots, n\}$. Let $\Phi \in \mathbb{R}^{n\times 2n}$ be the matrix obtained by normalizing the columns of $\boldsymbol{A}$. Let $\boldsymbol{P}$ be the matrix obtained by normalizing the columns of $\boldsymbol{M}$ which is a submatrix of Φ. With probability $1 - 3e^{-\delta^2 k}$ over the generation of S, we have that*

$$\mathbb{E}[\sigma_{\max}(\boldsymbol{M}^*\boldsymbol{M} - \mathbf{I})] \le ck\rho\log n. \tag{4.3}$$

Proof We first calculate the coherence of the matrix Φ which is defined as $\mu(\Phi) = \sup_{i\neq j} |\phi_i^* \phi_j|$ where ϕ_i is the ith column of Φ for $1 \le i \le 2n$.

Lemma 4.7 $\mathbb{P}(\mu(\Phi) \le c_1\sqrt{\log n}\rho) \ge 1 - 4n^{-3}$ *where the probability is over* $\boldsymbol{r}$ *and* $c_1 > 0$ *is an absolute constant.*

Proof For some absolute constant $c_1 = 8c_2^{-1/2}, c_2 > 0$ we have the followings. Applying Lemma 4.5, for all terms, we have that $|\,\|\boldsymbol{X}_i\|_{\ell_2}^2 - 1|, |\boldsymbol{X}_i^*\boldsymbol{X}_j|, |\boldsymbol{X}_i^*\boldsymbol{Y}_j|, |\boldsymbol{X}_i^*\boldsymbol{Y}_i| \le \gamma$ with probability $1-4n^2\exp(-c_2\min\{\gamma\rho^{-2}, \gamma^2\rho^{-2}\})$. Hence, picking $\gamma = c_1\sqrt{\log n}\rho/2$, we can guarantee that $\sup_{i\neq j}|\phi_i^*\phi_j|$ is small for all i, j pairs with probability $1 - 4n^{-3}$ after normalizing the columns by their lengths $\|\boldsymbol{X}_i\|_{\ell_2}, \|\boldsymbol{Y}_i\|_{\ell_2}$. ∎

Next, Lemma 4.9 shows that the spectral norm of Φ can be bounded as $\|\Phi\| \le c_3\rho\sqrt{n\log n}$ with probability $1 - n^{-3}$ as well. Assume $c_3 > c_1$ without losing generality.

Let us call the event that "$\mu(\Phi) \le c_3\sqrt{\log n}\rho$ and $|\,\|\boldsymbol{X}_i\|_{\ell_2}^2 - 1| \le c_3\sqrt{\log n}\rho$ and $\|\Phi\| \le c_3\rho\sqrt{n\log n}$" as E which is an event over $\boldsymbol{r}$ with probability at least $1-5n^{-3}$. Split $\mathbb{P}$ into $\boldsymbol{X}$ and $\boldsymbol{Y}$ parts namely $\boldsymbol{P} = [\boldsymbol{P}_{\boldsymbol{X}}\ \boldsymbol{P}_{\boldsymbol{Y}}]$ where $\boldsymbol{P}_{\boldsymbol{X}}, \boldsymbol{P}_{\boldsymbol{Y}} \in \mathbb{R}^{n\times k}$. Conditioned on E, applying Theorem C.4 with $u = 2\delta\sqrt{k}\log k$, with probability $1 - e^{-\delta^2 k}$ over the choice of support S, we have that

$$\mathbb{E}_{\boldsymbol{r}|E}[\sigma_{\max}(\boldsymbol{P}_{\boldsymbol{X}}^*\boldsymbol{P}_{\boldsymbol{X}} - \mathbf{I})] \le c_4 k\rho\log n \qquad \mathbb{E}_{\boldsymbol{r}|E}[\sigma_{\max}(\boldsymbol{P}_{\boldsymbol{Y}}^*\boldsymbol{P}_{\boldsymbol{Y}} - \mathbf{I})] \le c_4 k\rho\log n$$

since we can bound (C.4) as follows

$$\begin{aligned}\mu(\Phi)\sqrt{k}u + \frac{k}{n}\|\Phi\|^2 &\le c_3(\rho\sqrt{\log n}\sqrt{k}\sqrt{k}\sqrt{\log k}\delta + \frac{k}{n}(\rho^2 n\log n)) \\ &\le c_3(k\delta\rho\log n + k\rho^2\log n) \le 2c_3 k\delta\rho\log n.\end{aligned}$$

where we used the fact that $\delta \ge \rho$. Similarly, applying Theorem C.7, we estimate the cross term as

$$\mathbb{E}_{\boldsymbol{r}|E}[\sigma_{\max}(\boldsymbol{P}_{\boldsymbol{X}}^*\boldsymbol{P}_{\boldsymbol{Y}})] \le c_4 k\rho\log n$$

This yields

$$\mathbb{E}_{\boldsymbol{r}|E}[\sigma_{\max}(\boldsymbol{P}^*\boldsymbol{P} - \mathbf{I})] \le 4c_4 k\rho\log n$$

with probability $1 - 3e^{-\delta^2 k}$. Next, observe that

$$\boldsymbol{P} = \boldsymbol{M}\mathrm{diag}(\alpha)$$

where $\alpha \in \mathbb{R}^{2k}$ is a vector whose entries lie between $\sqrt{1 \pm c_3\sqrt{\log n}\rho}$. Let $c_5 = \max\{2c_3, 8c_4\}$. Consequently, applying Lemma B.4

$$\sigma_{\max}(\boldsymbol{M}^*\boldsymbol{M} - \mathbf{I}) \le c_5 k\rho\log n. \tag{4.4}$$

For the complementary event $\bar{E}$, independent of the support S we will use a simpler estimate namely

$$\sigma_{\max}(\boldsymbol{M}^*\boldsymbol{M} - \mathbf{I}) \le \sigma_{\max}(\boldsymbol{M}^*\boldsymbol{M}) \le \|\boldsymbol{M}\|_F^2 = k(\|\boldsymbol{X}_1\|_{\ell_2}^2 + \|\boldsymbol{Y}_1\|_{\ell_2}^2) \le 2k\rho^2\|\boldsymbol{r}\|_{\ell_2}^2.$$

For this case, applying Lemma B.1 with $p = 5n^{-3}$ yields that

$$\mathbb{E}[\|\boldsymbol{r}\|_{\ell_2}^2 |\bar{E}]\mathbb{P}(\bar{E}) \le c_6 n^{-2} \tag{4.5}$$

Combining the estimates over E (4.4) and $\bar{E}$ (4.5), we find that with the desired probability over S $(1-e^{-\delta^2 k})$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{r}}[\sigma_{\max}(\boldsymbol{M}^*\boldsymbol{M} - \mathbf{I})] &= \mathbb{E}_{\boldsymbol{r}}[\sigma_{\max}(\boldsymbol{M}^*\boldsymbol{M} - \mathbf{I})|E]\mathbb{P}(E) \\ &\quad + \mathbb{E}_{\boldsymbol{r}}[\sigma_{\max}(\boldsymbol{M}^*\boldsymbol{M} - \mathbf{I})|\bar{E}]\mathbb{P}(\bar{E}) \\ &\le c_5 k\rho\log n + c_6 n^{-2} \approx c_5 k\rho\log n\end{aligned}$$

where we used the fact that $k\rho\log n \ge n^{-2}$. ∎

4.2 Probabilistic bounds on the singular values

Lemma 4.8 *Let $\boldsymbol{R} \in \mathbb{R}^{n\times k}$ be a matrix obtained by picking k elements from $\{\boldsymbol{X}_i\}_{i=1}^n \subset \mathbb{R}^n$ and stacking them next to each other. The maximum and minimum singular values of $\boldsymbol{R}$ are $\sqrt{k}\,\|\boldsymbol{x}\|_{\ell_\infty}$ Lipschitz function of $\boldsymbol{r}$.*

Proof We view $\boldsymbol{R}$ as a random matrix obtained from the vector $\boldsymbol{r}$. Given an alternative vector $\hat{\boldsymbol{r}}$, construct $\hat{\boldsymbol{R}}$ from circular shifts of the vector $\operatorname{diag}(\hat{\boldsymbol{r}})\boldsymbol{x}$ in an identical manner to $\boldsymbol{R}$ (i.e. form $\{\hat{\boldsymbol{X}}_i\}_{i=1}^n$ and pick the same k elements). Applying Lemma A.1, we have that

$$\|\hat{\boldsymbol{R}} - \boldsymbol{R}\|_F = \sqrt{k}\,\|\operatorname{diag}(\hat{\boldsymbol{r}})\boldsymbol{x} - \operatorname{diag}(\boldsymbol{r})\boldsymbol{x}\|_{\ell_2} \le \sqrt{k}\,\|\boldsymbol{x}\|_{\ell_\infty}\,\|\boldsymbol{r} - \hat{\boldsymbol{r}}\|_{\ell_2}$$

which is the desired conclusion. ∎

Lemma 4.9 *Let $\boldsymbol{X} = [\boldsymbol{X}_1 \ \dots \ \boldsymbol{X}_n]$ where $\boldsymbol{X}_i$ are obtained by circular shifts of $\operatorname{diag}(\boldsymbol{r})\boldsymbol{x}$. There exists an absolute constant $c > 0$ such that*

$$\mathbb{P}(\|\boldsymbol{X}\| \ge c\rho\sqrt{n\log n}) \le c_1 n^{-3}. \tag{4.6}$$

Next, consider the matrix Φ of Theorem 4.6. Assuming $\rho < c'(\log n)^{-1/2}$, Φ obeys the following similar bound

$$\mathbb{P}(\|\Phi\| \ge 4c\rho\sqrt{n\log n}) \le (2c_1 + 4)n^{-3}.$$

Proof Let $g \sim \mathcal{N}(0,1)$. From Stirling's approximation, we have that

$$\mathbb{E}[g^{2d}] = (2d)!! = \frac{(2d)!}{2^d d!} \le \frac{e(2d)^{2d+1/2}\exp(-2d)}{2^d\sqrt{2\pi}d^{d+1/2}\exp(-d)} \le c_1(2/e)^d d^d.$$

Pick a complex standard normal $g' = g_1 + ig_2$ where $g_1, g_2 \sim \mathcal{N}(0,1)$. Comparing the moments of g' to g

$$\mathbb{E}[|g'|^{2d}] = \mathbb{E}[(g_1^2 + g_2^2)^d] \le \mathbb{E}[(g_1^2 + g_1^2)^d] \le c_1(4/e)^d d^d.$$

Suppose h is a real random variable obeying $\mathbb{E}[h^{2d}] \le n\,\mathbb{E}[|g'|^{2d}] \le c_1 n(4/e)^d d^d$ for all $d \ge 1$. Then, using Markov inequality

$$\mathbb{P}(|h| \ge t) \le t^{-2d} c_1 n(4/e)^d d^d.$$

Pick $t = \sqrt{2d}$ and $d = (C\log n)/2$ to find that

$$\mathbb{P}(|h| \ge t) \le t^{-2d} c_1 n(2/e)^d d^d \le c_1 n(2/e)^{(C\log n)/2} \le c_1 n^{-3}$$

by picking $C > 0$ to be a large enough constant. This gives

$$\mathbb{P}(|h| \ge \sqrt{C\log n}) \le c_1 n^{-3}. \tag{4.7}$$

The remaining discussion will analyze the spectral norm of $\boldsymbol{X}$ to make use of (4.7). Observe that the random variable $\operatorname{tr}((\boldsymbol{X}^*\boldsymbol{X})^d) \ge \|\boldsymbol{X}\|^{2d}$. Now form the complex circulant matrix $\boldsymbol{X}'$ by stacking circular shifts $\{s_i(\operatorname{diag}(\boldsymbol{g})\boldsymbol{x}')\}_{i=1}^n$ next to each other where are entries of the vector $\boldsymbol{x}'$ are equal to ρ i.e. $\boldsymbol{x}' = [\rho \ \dots \ \rho]^*$ and $\boldsymbol{g}$ is a vector of independent random variables where each entry is distributed as g'.

Singular values of $\rho^{-1}\boldsymbol{X}'$ are trivial. In particular, singular values are absolute values of the eigenvalues and eigenvalues are independent complex Gaussian random variables whose imaginary and real parts have variance $\sqrt{n}$. Next, we relate $\boldsymbol{X}'$ to $\boldsymbol{X}$. Let $\boldsymbol{X}'_{real}$ denote the real part of the matrix $\boldsymbol{X}'$. First observe that, the following deterministic relation holds for all $d \ge 1$

$$\operatorname{trace}(\boldsymbol{X}'^*\boldsymbol{X}')^d \ge \operatorname{trace}((\boldsymbol{X}'_{real})^*\boldsymbol{X}'_{real})^d.$$

On the other hand, $\mathbb{E}\operatorname{trace}((\boldsymbol{X}'_{real})^*\boldsymbol{X}'_{real})^d \ge \mathbb{E}\operatorname{trace}(\boldsymbol{X}^*\boldsymbol{X})^d$. This follows form the fact that when the traces are expanded term by term, each individual nonzero term of the left-hand side dominates that of the

right-hand side as entries of $\boldsymbol{x}'$ are at least as large as that of $\boldsymbol{x}$ (in absolute value). Finally observe that $\operatorname{trace}(\boldsymbol{X}'^*\boldsymbol{X}')^d \sim \sum_{i=1}^n g_i^{2d}$ where g_i are independent standard complex with variance $2\rho^2 n$. Consequently, setting $h = \|\boldsymbol{X}\|$, we find that

$$\mathbb{E}[h^{2d}] = \mathbb{E}[\|\boldsymbol{X}\|^{2d}] \le \mathbb{E}\operatorname{trace}(\boldsymbol{X}^*\boldsymbol{X})^d \le \mathbb{E}\operatorname{trace}(\boldsymbol{X}'^*\boldsymbol{X})^d = \mathbb{E}[\sum_{i=1}^n g_i^{2d}].$$

To conclude with the proof of the first statement, apply the estimate (4.7) by normalizing both sides by $\rho\sqrt{n}$ and obtain the advertised result (4.6).

To obtain the result on Φ, recalling Theorem 4.6, we write $\Phi\operatorname{diag}(\alpha) = [\boldsymbol{X}\ \boldsymbol{Y}]$ where $\boldsymbol{X}, \boldsymbol{Y} = [\boldsymbol{Y}_1\ \ldots\ \boldsymbol{Y}_n]$ have spectral norm at most $c\rho\sqrt{n\log n}$ and α is a diagonal (length) normalization matrix whose entries are at least $1/2$ with probability $1 - 4n^{-3}$ as soon as $\rho < c'(\log n)^{-1/2}$ for sufficiently small constant $c' > 0$. ∎

4.2.1 Finalizing the Proof of Theorem 4.2

Lemma 4.10 *Consider the setup in Theorem 4.2 and set* $\boldsymbol{M} = [\boldsymbol{X}_{r_1}\ \ldots\ \boldsymbol{X}_{r_k}\ \boldsymbol{Y}_{r_1}\ \ldots\ \boldsymbol{Y}_{r_k}]$. *There exists a constant* $c_1 > 0$ *such that with probability* $1 - 4\exp(-\delta^2 k)$ *over the generation of* $\{r_i\}_{i=1}^k$ *and modulation* $\boldsymbol{r}$, *we have that*

$$\|\boldsymbol{M}^*\boldsymbol{M} - \mathbf{I}\| \le c_1\rho\delta k\log n.$$

Proof From Theorem 4.6, we know that with probability $1 - 3e^{-\delta^2 k}$ over support S

$$\mathbb{E}[\|\boldsymbol{M}^*\boldsymbol{M} - \mathbf{I}\|] \le c\delta k\rho\log n.$$

On the other hand, $\sigma_{\min}(\boldsymbol{M})$ and $\sigma_{\max}(\boldsymbol{M})$ are $\sqrt{k}\,\|\boldsymbol{x}\|_{\ell_\infty}$ Lipschitz functions of $\boldsymbol{r}$. Consequently, conditioned on S, applying Lemma A.2, we have that

$$\mathbb{P}(|\|\boldsymbol{M}\| - \mathbb{E}[\|\boldsymbol{M}\|]| \le \delta\sqrt{2k}\,\|\boldsymbol{x}\|_{\ell_\infty}) \le 2\exp(-\delta^2 k).$$

Combining the expectation and deviation estimates, with probability $1 - 2\exp(-\delta^2 k)$, we obtain that

$$\|\boldsymbol{M}\| \le \sqrt{1 + c\delta k\rho\log n} + \sqrt{2}\delta k\rho \le 1 + (c + \sqrt{2})\delta k\rho\log n.$$

The exact same argument applies to the minimum singular value $\sigma_{\min}(\boldsymbol{M})$ which gives

$$\sigma_{\min}(\boldsymbol{M}) \ge \sqrt{1 - c\delta k\rho\log n} - \sqrt{2}\delta k\rho \ge 1 - (c + \sqrt{2})\delta k\rho\log n$$

allowing us to conclude with the desired result. ∎

5 On orthogonal decomposition of Gaussian circulant pairs

Let $\boldsymbol{x}, \boldsymbol{y}$ be two unit vectors chosen from $\{\boldsymbol{v}_i\}_{i=1}^N$. Form $\{\boldsymbol{X}_{r_i}, \boldsymbol{Y}_{r_i}\}_{i=1}^k$ via uniform sampling of Gaussian circular rotations $\{\boldsymbol{X}_i, \boldsymbol{Y}_i\}_{i=1}^n$. Decompose $\boldsymbol{X}_{r_i} = \boldsymbol{X}'_{r_i} + \boldsymbol{p}_i$, $\boldsymbol{Y}_{r_i} = \boldsymbol{Y}'_{r_i} + \boldsymbol{p}'_i$ where $\boldsymbol{p}_i, \boldsymbol{p}'_i$ are the projections of $\boldsymbol{X}_{r_i}, \boldsymbol{Y}_{r_i}$ onto the span of $\{\boldsymbol{X}'_{r_j}, \boldsymbol{Y}'_{r_j}\}_{j=1}^{i-1}$. Observe that this has a similar flavor to QR decomposition.

Lemma 5.1 *Let* S_i *be the subspace spanned by* $\{\boldsymbol{X}_{r_j}, \boldsymbol{Y}_{r_j}\}_{j=1}^{i-1}$. *With probability* $1 - 4\exp(-\delta^2 k)$, *we have that for all* $1 \le i \le k$

$$\max\{\|\mathcal{P}_{S_i}(\boldsymbol{X}_{r_i})\|_{\ell_2} = \|\boldsymbol{p}_i\|_{\ell_2},\ \ \|\mathcal{P}_{S_i}(\boldsymbol{Y}_{r_i})\|_{\ell_2} = \|\boldsymbol{p}'_i\|_{\ell_2}\} \le c_1\delta k\rho_{direct}. \tag{5.1}$$

Proof Our proof is in similar spirit to Lemma 12 of [23]. The main difference is that we apply an additional orthogonalization procedure that reduces dependency on the correlation $|\boldsymbol{x}^*\boldsymbol{y}|$ and improves our estimates. To start analysis, let us focus on $\boldsymbol{X}_{r_i}$ only. First observe that

$$\text{span}([\boldsymbol{X}_{r_1} \ \dots \ \boldsymbol{X}_{r_i} \ \boldsymbol{Y}_{r_1} \ \dots \ \boldsymbol{Y}_{r_i}]) = \text{span}([\boldsymbol{X}'_{r_1} \ \dots \ \boldsymbol{X}'_{r_i} \ \boldsymbol{Y}'_{r_1} \ \dots \ \boldsymbol{Y}'_{r_i}]).$$

Next observe that

$$\text{span}([\boldsymbol{X}_{r_1} \ \dots \ \boldsymbol{X}_{r_i} \ \boldsymbol{Y}_{r_1} \ \dots \ \boldsymbol{Y}_{r_i}]) = \text{span}([\boldsymbol{X}_{r_1} \ \dots \ \boldsymbol{X}_{r_i} \ \boldsymbol{Y}^{\perp}_{r_1} \ \dots \ \boldsymbol{Y}^{\perp}_{r_i}])$$

where $\boldsymbol{Y}^{\perp}{}_{r_1}$ is obtained by the Gaussian circular rotations of $\boldsymbol{y}^{\perp} = \frac{\boldsymbol{y}-\cos(\theta)\boldsymbol{x}}{\|\boldsymbol{y}-\cos(\theta)\boldsymbol{x}\|_{\ell_2}}$ where θ is the angle between $\boldsymbol{x}$ and $\boldsymbol{y}$. Consequently, we can focus on understanding the projection of $\boldsymbol{X}_{r_i}$ onto the column span of $\boldsymbol{M}_i = [\boldsymbol{X}_{r_1} \ \dots \ \boldsymbol{X}_{r_{i-1}} \ \boldsymbol{Y}^{\perp}{}_{r_1} \ \dots \ \boldsymbol{Y}^{\perp}{}_{r_{i-1}}]$. Let $\boldsymbol{M}_i$ have singular value decomposition $\boldsymbol{U}_L \Sigma \boldsymbol{U}_R^*$ where $\Sigma \in \mathbb{R}^{2(i-1)\times 2(i-1)}$. Consider the vector

$$\boldsymbol{q}_i = \boldsymbol{M}_i^* \boldsymbol{X}_{r_i} \in \mathbb{R}^{2(i-1)}$$

From Lemma 4.5, we know that each entry of $\boldsymbol{q}_i$ is less than $c_2 \max\{\rho_{direct}\rho_{cross}\delta^2 k, \rho_{direct}\delta\sqrt{k}\} \leq c_2\delta\rho_{direct}\sqrt{k}$ for all $1 \leq i \leq k$ with probability $1-\exp(-\delta^2 k)$ where we used the fact that $\delta^2 k \geq c' \log N \geq c' \log n$. On the other hand, using Theorem 4.2, with the same probability all matrices $\{\boldsymbol{M}_i\}_{i=1}^k$ satisfy

$$\|\boldsymbol{M}_i^*\boldsymbol{M}_i - \mathbf{I}\| \leq c_3\rho_{cross} k \log n \implies \sigma_{\min}(\boldsymbol{M}_i) = \sigma_{\min}(\Sigma) \geq 1 - c_3\rho_{cross}\delta k \log n \geq 1/2.$$

Consequently, the projection can be bounded as

$$\|\mathcal{P}_{S_i}(\boldsymbol{X}_{r_i})\|_{\ell_2} = \|\boldsymbol{U}_L^*\boldsymbol{X}_{r_i}\|_{\ell_2} \leq \sigma_{\min}^{-1}(\Sigma)\,\|\Sigma\boldsymbol{U}_L^*\boldsymbol{X}_{r_i}\|_{\ell_2} = \sigma_{\min}^{-1}(\Sigma)\,\|\boldsymbol{M}_i^*\boldsymbol{X}_{r_i}\|_{\ell_2}.$$

This implies that, with $1-2\exp(-\delta^2 k)$ probability, $\|\mathcal{P}_{S_i}(\boldsymbol{X}_{r_i})\|_{\ell_2} \leq c_2\sqrt{2k \times (\delta\rho_{direct}\sqrt{k})^2} = 2c_2\delta k\rho_{direct}$. The identical argument applies to $\boldsymbol{Y}_{r_i}$. ∎

Lemma 5.2 *Consider the matrix $\boldsymbol{P} \in \mathbb{R}^{n\times 2k}$ obtained by concatenating $\boldsymbol{p}_i, \boldsymbol{p}'_i$ for $1 \leq i \leq k$. Under initial assumptions, we have that $\|\boldsymbol{P}\| \leq 7$ with probability $1-8\exp(-\delta^2 k)$.*

Proof Consider the matrix $\boldsymbol{M} = [\boldsymbol{X}_{r_1} \ \dots \ \boldsymbol{X}_{r_k} \ \boldsymbol{Y}_{r_1} \ \dots \ \boldsymbol{Y}_{r_k}]$. From Corollary 4.3, we know that $\|\boldsymbol{M}\| \leq 3$ with probability $1-6\exp(-\delta^2 k)$. On the other hand, using Gaussian concentration, each column of $\boldsymbol{M}$ obeys

$$\mathbb{P}(\|\boldsymbol{X}_{r_i}\|_{\ell_2} \leq 2) \leq 1 - \exp(-0.5\rho^{-2}).$$

Using our initial assumption $\delta \geq c' k\rho$ (see Condition 2.1), this holds for all columns with probability $1 - k\exp(-0.5\rho^{-2}) \geq 1-\exp(-\delta^2 k)$. Given this, observe that $\boldsymbol{X}'_{r_i}$ is perpendicular to $\{\boldsymbol{X}'_{r_j}\}_{j\neq i}$ and $\|\boldsymbol{X}'_{r_i}\|_{\ell_2} \leq \|\boldsymbol{X}_{r_i}\|_{\ell_2} \leq 2$. This ensures that

$$\|[\boldsymbol{X}'_{r_1} \ \dots \ \boldsymbol{X}'_{r_k}]\| \leq \max_{1\leq i\leq k} \|\boldsymbol{X}_{r_k}\|_{\ell_2} \leq 2.$$

The same argument applies to $\boldsymbol{Y}'_{r_i}$ ensuring $\boldsymbol{M}' = [\boldsymbol{X}'_{r_1} \ \dots \ \boldsymbol{X}'_{r_k} \ \boldsymbol{Y}'_{r_1} \ \dots \ \boldsymbol{Y}'_{r_k}]$ has spectral norm of at most 4. Consequently $\|\boldsymbol{P}\| = \|\boldsymbol{M} - \boldsymbol{M}'\| \leq \|\boldsymbol{M}\| + \|\boldsymbol{M}'\| \leq 7$. ∎

Lemma 5.3 *The matrix $\boldsymbol{P} = [\boldsymbol{p}_1 \ \dots \ \boldsymbol{p}_k \ \boldsymbol{p}'_1 \ \dots \ \boldsymbol{p}'_k]$ obeys the following bounds with probability $1-12\exp(-c\delta^2 k)$.*

- *Each column $\boldsymbol{p}_i, \boldsymbol{p}'_i$ of $\boldsymbol{P}$ satisfies $\|\boldsymbol{p}_i\|_{\ell_2}, \|\boldsymbol{p}'_i\|_{\ell_2} \leq C\delta\rho k$ for all $1 \leq i \leq k$.*
- *Spectral norm of $\boldsymbol{P}$ satisfies $\|\boldsymbol{P}\| \leq 7$.*

Proof The proof follows directly by making use of Lemmas 5.2 and 5.1. ∎

6 Final perturbation analysis

We are in a position to prove our main result Theorem 2.2.

Proof The proof is based on perturbation analysis, namely to what extent structured samples deviate from Gaussian-like behavior. We break the analysis in two parts, namely over $\boldsymbol{r}$ and over $\boldsymbol{h}$.

• **Upper bounds on the perturbation due to $\boldsymbol{r}$:**

Recall that $\boldsymbol{C} = \boldsymbol{C}_{\boldsymbol{h}}$ is the circulant part of the embedding operator where $\boldsymbol{h} \sim \mathcal{N}(0, \mathbf{I})$ is its first row and ith row is equal to $s_i(\boldsymbol{h})$ for $1 \le i \le n$. Given any two points $\boldsymbol{x}, \boldsymbol{y}$, chosen from $\{\boldsymbol{v}_i\}_{i=1}^N$ consider the vectors $\boldsymbol{x}' = \mathrm{diag}(\boldsymbol{r})\boldsymbol{x} = \boldsymbol{X}_1$ and $\boldsymbol{y}' = \mathrm{diag}(\boldsymbol{r})\boldsymbol{y} = \boldsymbol{Y}_1$. Now, observe that the ith entry of $\boldsymbol{C}\boldsymbol{x}'$ is equal to

$$s_i(\boldsymbol{h})^* \boldsymbol{x}' = \boldsymbol{h}^* s_{n-i}(\boldsymbol{x}') = \boldsymbol{h}^* \boldsymbol{X}_{n-i}.$$

Similarly the ith entry of $\boldsymbol{C}\boldsymbol{y}'$ is equal to $\boldsymbol{h}^* \boldsymbol{Y}_{n-i}$. Consequently, for a random subsampling $\boldsymbol{RC} \in \mathbb{R}^{k \times n}$ of $\boldsymbol{C}$, we have that

$$\boldsymbol{RC}\boldsymbol{x}' = \boldsymbol{M}_{\boldsymbol{x}} \boldsymbol{h}, \;\; \boldsymbol{RC}\boldsymbol{y}' = \boldsymbol{M}_{\boldsymbol{y}} \boldsymbol{h},$$

where $\boldsymbol{M}_{\boldsymbol{x}} = [\boldsymbol{X}_{r_1} \; \ldots \; \boldsymbol{X}_{r_k}]^*$ and $\boldsymbol{M}_{\boldsymbol{y}} = [\boldsymbol{Y}_{r_1} \; \ldots \; \boldsymbol{Y}_{r_k}]^*$ and $\{r_i\}_{i=1}^k$ are randomly selected coordinates. Next, for each $1 \le i \le k$, we decompose $\boldsymbol{X}_{r_i}, \boldsymbol{Y}_{r_i}$ as described in Section 5.

$$\boldsymbol{X}_{r_i} = \boldsymbol{X}'_{r_i} + \boldsymbol{p}_i, \;\; \boldsymbol{Y}_{r_i} = \boldsymbol{Y}'_{r_i} + \boldsymbol{p}'_i.$$

Since $\boldsymbol{h}$ is a standard Gaussian vector, by construction, $\boldsymbol{h}^* \boldsymbol{X}'_{r_i}$ and $\boldsymbol{h}^* \boldsymbol{Y}'_{r_i}$ is independent of $\{\boldsymbol{h}^* \boldsymbol{X}'_{r_j}, \boldsymbol{h}^* \boldsymbol{Y}'_{r_j}\}_{j \neq i}$. To proceed, let us estimate the angle between $\boldsymbol{X}'_{r_i}, \boldsymbol{Y}'_{r_i}$ probabilistically.

Firstly, $\|\boldsymbol{X}_{r_i}\|_{\ell_2}^2, \|\boldsymbol{Y}_{r_i}\|_{\ell_2}^2$ lies between $1 \pm c_1 \delta\rho\sqrt{k}$ with probability $1 - \exp(-\delta^2 k)$. Next, with the same probability $|\boldsymbol{X}_{r_i}^* \boldsymbol{Y}_{r_i} - \boldsymbol{x}^* \boldsymbol{y}| \le c_1 \delta\rho\sqrt{k}$. Together these imply that $|\bar{\boldsymbol{X}}_{r_i}^* \bar{\boldsymbol{Y}}_{r_i} - \boldsymbol{x}^* \boldsymbol{y}| \le 4 c_1 \delta\rho\sqrt{k}$ where $\bar{\boldsymbol{a}} = \boldsymbol{a}/\|\boldsymbol{a}\|_{\ell_2}$. Making use of Lemma A.5, we can conclude that

$$|\mathrm{ang}(\boldsymbol{X}_{r_i}, \boldsymbol{Y}_{r_i}) - \mathrm{ang}(\boldsymbol{x}, \boldsymbol{y})| \le c_2 \sqrt{\rho\delta\sqrt{k}} := \Delta_{\boldsymbol{r}}. \tag{6.1}$$

In particular, since $\boldsymbol{X}_{r_i}, \boldsymbol{Y}_{r_i}$ are circulant rotations of $\boldsymbol{X}_{r_1}, \boldsymbol{Y}_{r_1}$ the angle between is exactly same i.e. $\mathrm{ang}(\boldsymbol{X}_{r_i}, \boldsymbol{Y}_{r_i}) = \mathrm{ang}(\boldsymbol{X}_{r_j}, \boldsymbol{Y}_{r_j})$ for $1 \le i, j \le k$.

With these, we can state the following result that summarizes the properties of the perturbation. Below we additionally used the fact that $\log N \le \delta^2 k/4$.

Lemma 6.1 *$\{\boldsymbol{X}_{r_i}, \boldsymbol{Y}_{r_i}\}$ satisfies the following with probability $1 - 12\exp(-\delta^2 k/2)$ for all $\boldsymbol{x}, \boldsymbol{y}$ pairs chosen from $\{\boldsymbol{v}_i\}_{i=1}^N$ where the probability is over $\boldsymbol{r}$ and support S.*

- *$|ang(\boldsymbol{X}_{r_i}, \boldsymbol{Y}_{r_i}) - ang(\boldsymbol{x}, \boldsymbol{y})| \le \Delta_{\boldsymbol{r}}$.*
- *$\{\boldsymbol{X}'_{r_i}, \boldsymbol{Y}'_{r_i}\}_{i=1}^k$ are orthogonal pairs and for all i, $\boldsymbol{X}_{r_i} - \boldsymbol{X}'_{r_i} = \boldsymbol{p}_i$, $\boldsymbol{Y}_{r_i} - \boldsymbol{Y}'_{r_i} = \boldsymbol{p}'_i$ where $\boldsymbol{p}_i, \boldsymbol{p}'_i$ obey*

$$\|\boldsymbol{p}_i\|_{\ell_2} \le C\rho\delta k, \;\; \|[\boldsymbol{p}_1 \; \ldots \; \boldsymbol{p}_k]\| \le 7.$$

What remains is to characterize the effect of perturbation error on the binary embedding distortion. Let $\theta_i = \mathrm{ang}(\boldsymbol{X}'_{r_i}, \boldsymbol{Y}'_{r_i})$. Applying Lemma A.5 again, we know for a fact that (by picking $c_2 > 0$ to be a large enough constant)

$$\mathrm{ang}(\boldsymbol{X}'_{r_i}, \boldsymbol{X}_{r_i}) \le c_2 \delta\rho k/2, \;\; \mathrm{ang}(\boldsymbol{Y}'_{r_i}, \boldsymbol{Y}_{r_i}) \le c_2 \rho\delta k/2.$$

Together, these ensure that

$$|\theta_i - \mathrm{ang}(\boldsymbol{x}, \boldsymbol{y})| = |\mathrm{ang}(\boldsymbol{X}'_{r_i}, \boldsymbol{Y}'_{r_i}) - \mathrm{ang}(\boldsymbol{x}, \boldsymbol{y})| \le c_2 \delta\rho k + |\mathrm{ang}(\boldsymbol{X}_{r_i}, \boldsymbol{Y}_{r_i}) - \mathrm{ang}(\boldsymbol{x}, \boldsymbol{y})| \le c_2 \delta\rho k + \Delta_{\boldsymbol{r}} := \Delta'_{\boldsymbol{r}}.$$

$\Delta'_{\boldsymbol{r}}$ will be the source of embedding distortion due to $\boldsymbol{r}$ and our initial assumptions will guarantee that it is small. Next section develops estimates for the remaining source of the perturbation which is connected to $\boldsymbol{h}$.

• **Probabilistic analysis of the perturbation due to $\boldsymbol{h}$:**
Pick $\delta_{buff} > 0$. For the rest of the discussion probabilities will be over $\boldsymbol{h}$. Let us define the events

$$E_i = (\boldsymbol{h}^* \boldsymbol{X}'_{r_i} > \delta_{buff} \text{ and } \boldsymbol{h}^* \boldsymbol{Y}'_{r_i} < -\delta_{buff}) \text{ or } (\boldsymbol{h}^* \boldsymbol{X}'_{r_i} < -\delta_{buff} \text{ and } \boldsymbol{h}^* \boldsymbol{Y}'_{r_i} > \delta_{buff}),$$
$$\bar{E}_i = (\boldsymbol{h}^* \boldsymbol{X}'_{r_i} > \delta_{buff} \text{ and } \boldsymbol{h}^* \boldsymbol{Y}'_{r_i} > \delta_{buff}) \text{ or } (\boldsymbol{h}^* \boldsymbol{X}'_{r_i} < -\delta_{buff} \text{ and } \boldsymbol{h}^* \boldsymbol{Y}'_{r_i} < -\delta_{buff}).$$

E_i and $\bar{E}_i$ are the robust versions of the events $\operatorname{sgn}(\boldsymbol{h}^* \boldsymbol{X}'_{r_i}) \neq \operatorname{sgn}(\boldsymbol{h}^* \boldsymbol{Y}'_{r_i})$ and $\operatorname{sgn}(\boldsymbol{h}^* \boldsymbol{X}'_{r_i}) = \operatorname{sgn}(\boldsymbol{h}^* \boldsymbol{Y}'_{r_i})$ respectively.

Without losing generality, let us consider the event E_i. Recall that with probability $1 - \exp(-\delta^2 k)$, for all $1 \leq i \leq k$, we can guarantee that $0.75 \leq \|\boldsymbol{X}_{r_i}\|_{\ell_2}^2 \leq 2$ and $\|\boldsymbol{p}_i\|_{\ell_2}^2 \leq 0.25$. Hence, conditioned on $\boldsymbol{r}$, $\boldsymbol{h}^* \boldsymbol{X}'_{r_i}$ (and $\boldsymbol{h}^* \boldsymbol{X}'_{r_i}$) is a Gaussian random variable with variance between 0.5 to 2. Also, observe that $\mathbb{P}(\operatorname{sgn}(\boldsymbol{h}^* \boldsymbol{X}'_{r_i}) \neq \operatorname{sgn}(\boldsymbol{h}^* \boldsymbol{Y}'_{r_i})) = \theta_i$. Consequently, letting $\theta = \operatorname{ang}(\boldsymbol{x}, \boldsymbol{y})$, from small ball probability of Gaussians, we have that

$$\mathbb{P}(E_i) \geq \operatorname{ang}(\theta_i) - c_3 \delta_{buff} \geq \operatorname{ang}(\theta) - c_3(\delta_{buff} + \Delta'_{\boldsymbol{r}}).$$

Let $E = \sum_{i=1}^k 1_{E_i}$. Consequently, applying a standard Chernoff bound, we find that

$$\mathbb{P}(E \geq k(\operatorname{ang}(\theta) - c_3(\delta_{buff} + \Delta'_{\boldsymbol{r}}) - \delta_{buff})) := \mathbb{P}(E \geq k(\operatorname{ang}(\theta) - c_4(\delta_{buff} + \Delta'_{\boldsymbol{r}}))) \geq \exp(-2\delta_{buff}^2 k) \tag{6.2}$$

where the probability is over $\boldsymbol{h}$.

Next, we consider the impact of perturbations $\{\boldsymbol{p}_i, \boldsymbol{p}'_i\}_{i=1}^k$. Using the facts that $\|\boldsymbol{P}\|_F^2 \leq C^2 \rho^2 \delta^2 k^3, \|\boldsymbol{P}\| \leq 7$ and applying Lemma A.3, we have that

$$\mathbb{P}(\|[\boldsymbol{p}_1 \ \boldsymbol{p}_2 \ \dots \ \boldsymbol{p}_k]^* \boldsymbol{h}\|_{\ell_2}^2 \leq \rho^2 \delta^2 k^3 + t) \geq 1 - \exp(-c_5 \min\{\frac{t^2}{50 C^2 \rho^2 \delta^2 k^3}, \frac{t}{50}\}).$$

To proceed, pick $t = \varepsilon_{buff} \delta_{buff}^2 k$ to obtain that with probability $1 - \exp(-c_6 \min\{\frac{\varepsilon_{buff}^2 \delta_{buff}^4}{\rho^2 \delta^2 k}, \varepsilon_{buff} \delta_{buff}^2 k\})$, perturbation obeys

$$\|\boldsymbol{P}^* \boldsymbol{h}\|_{\ell_2}^2 \leq \rho^2 \delta^2 k^3 + \varepsilon_{buff} \delta_{buff}^2 k. \tag{6.3}$$

The same bound applies to the perturbation over $\boldsymbol{y}$ namely $\boldsymbol{P}' = [\boldsymbol{p}'_1 \ \dots \ \boldsymbol{p}'_k]$. Now for any $1 \leq i \leq k$, observe that $\operatorname{sgn}(\boldsymbol{X}_{r_i}) \neq \operatorname{sgn}(\boldsymbol{Y}_{r_i})$ whenever

$$\text{i) } E_i \text{ holds} \quad \text{and} \quad \text{ii) } \max\{|\boldsymbol{h}^*(\boldsymbol{X}_{r_i} - \boldsymbol{X}'_{r_i})|, |\boldsymbol{h}^*(\boldsymbol{Y}_{r_i} - \boldsymbol{Y}'_{r_i})|\} < \delta_{buff}. \tag{6.4}$$

We know that E_i holds on at least $k(\operatorname{ang}(\theta) - c_4(\delta_{buff} + \Delta'_{\boldsymbol{r}}))$ coordinates. Next, we can upper bound the number of coordinates for which (6.4) does not hold. Using the estimate (6.3), this number is given by

$$\frac{\|\boldsymbol{P}^* \boldsymbol{h}\|_{\ell_2}^2 + \|\boldsymbol{P}'^* \boldsymbol{h}\|_{\ell_2}^2}{\delta_{buff}^2} \leq k \varepsilon_{buff} + \delta_{buff}^{-2} \rho^2 \delta^2 k^3. \tag{6.5}$$

With the estimates (6.5) and (6.2), we find that for all pairs $\boldsymbol{x}, \boldsymbol{y}$ with probability

$$1 - N^2 \exp(-c_6 \min\{\frac{\varepsilon_{buff}^2 \delta_{buff}^4}{\rho^2 \delta^2 k}, \varepsilon_{buff} \delta_{buff}^2 k\}) - \exp(-\delta^2 k/2)$$

we have that

$$k^{-1} \|\boldsymbol{RAx}, \boldsymbol{RAy}\|_H \geq \operatorname{ang}(\boldsymbol{x}, \boldsymbol{y}) - [\varepsilon_{buff} + \delta_{buff}^{-2} \delta^2 \rho^2 k^2 + c_3(\delta_{buff} + \Delta'_{\boldsymbol{r}})].$$

The identical (symmetric) argument allows us to obtain the upper bound on the Hamming distance to conclude that

$$|k^{-1} \|\boldsymbol{RAx}, \boldsymbol{RAy}\|_H - \operatorname{ang}(\boldsymbol{x}, \boldsymbol{y})| \leq \varepsilon_{buff} + \delta_{buff}^{-2} \delta^2 \rho^2 k^2 + c_3(\delta_{buff} + \Delta'_{\boldsymbol{r}}).$$

With these bounds, we find that binary embedding with $c_{final} \delta$ distortion succeeds with probability $1 - 2\exp(-c_6 \delta^3 k/2)$ under the following conditions:

- $\varepsilon_{buff} \le \delta$,
- $\delta_{buff} \le \delta$,
- $\rho^2 k^2 \le \delta_{buff}^2 \delta^{-1}$,
- Via $\Delta_{\boldsymbol{r}}'$: $\rho\delta\sqrt{k} \le \delta^2$ i.e. $\rho\sqrt{k} \le \delta$,
- Via $\Delta_{\boldsymbol{r}}'$: $\rho\delta k \le \delta$.

To satisfy these, pick $\delta_{buff} = \varepsilon_{buff} = \delta$. Furthermore, our initial assumptions (Condition 2.1) guarantee that $\delta \ge C_0 \rho k \ge C_0 \max\{\rho^2 k^2, \rho\sqrt{k}\}$ for a sufficiently large constant $C_0 > 0$ which yields a total distortion proportional to δ. Finally, the probability of success is

$$1 - N^2 \exp(-c_6 \min\{\frac{\delta^4}{\rho^2 k}, \delta^3 k\}) - \exp(-\delta^2 k/2).$$

Observing $\rho^2 k \le \delta^2 / C_0^2$ ($C_0 > 1$ is sufficient) and using the initial assumption $C_1 \log N \le \delta^3 k$ for a sufficiently large constant $C_1 > 0$ we can conclude. In particular pick $C_1 > 4/c_6$. With these, we ensured that the total distortion is $c_{final}\delta$ for an absolute constant $c_{final} > 0$ with the desired probability. Rescaling δ as $\delta \to c_{final}^{-1}\delta$, we conclude with the advertised result in Theorem 2.2. ■

References

[1] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast johnson-lindenstrauss transform. In *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pages 557–563. ACM, 2006.

[2] Petros T Boufounos and Richard G Baraniuk. 1-bit compressive sensing. In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, pages 16–21. IEEE, 2008.

[3] Jean Bourgain and Lior Tzafriri. Invertibility of ?large?submatrices with applications to the geometry of banach spaces and harmonic analysis. *Israel journal of mathematics*, 57(2):137–224, 1987.

[4] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *Information Theory, IEEE Transactions on*, 61(4):1985–2007, 2015.

[5] Raja Giryes, Guillermo Sapiro, and Alex M Bronstein. Deep neural networks with random gaussian weights: A universal classification strategy? *arXiv preprint arXiv:1504.08291*, 2015.

[6] Yunchao Gong and Svetlana Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 817–824. IEEE, 2011.

[7] Laurent Jacques. A quantized johnson–lindenstrauss lemma: The finding of buffon?s needle. *Information Theory, IEEE Transactions on*, 61(9):5012–5027, 2015.

[8] Laurent Jacques, Jason N Laska, Petros T Boufounos, and Richard G Baraniuk. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *Information Theory, IEEE Transactions on*, 59(4):2082–2102, 2013.

[9] Kishore Jaganathan, Samet Oymak, and Babak Hassibi. Sparse phase retrieval: Convex algorithms and limitations. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 1022–1026. IEEE, 2013.

[10] Felix Krahmer and Rachel Ward. New and improved johnson-lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281, 2011.

[11] Quoc Le, Tamás Sarlós, and Alex Smola. Fastfood-approximating kernel expansions in loglinear time. *ICML*, 2013.

[12] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[13] Samet Oymak and Ben Recht. Near-optimal bounds for binary embeddings of arbitrary sets. *arXiv preprint arXiv:1512.04433*, 2015.

[14] Samet Oymak, Benjamin Recht, and Mahdi Soltanolkotabi. Isometric sketching of any set via the restricted isometry property. *arXiv preprint arXiv:1506.03521*, 2015.

[15] Yaniv Plan and Roman Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *Information Theory, IEEE Transactions on*, 59(1):482–494, 2013.

[16] Yaniv Plan and Roman Vershynin. Dimension reduction by random hyperplane tessellations. *Discrete & Computational Geometry*, 51(2):438–461, 2014.

[17] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2007.

[18] Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab*, 18(0), 2013.

[19] Joel A Tropp. On the conditioning of random subdictionaries. *Applied and Computational Harmonic Analysis*, 25(1):1–24, 2008.

[20] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[21] Qifan Wang, Bin Shen, Shumiao Wang, Liang Li, and Luo Si. Binary codes embedding for fast image tagging with incomplete labels. In *Computer Vision–ECCV 2014*, pages 425–439. Springer, 2014.

[22] Xinyang Yi, Constantine Caramanis, and Eric Price. Binary embedding: Fundamental limits and fast algorithm. *arXiv preprint arXiv:1502.05746*, 2015.

[23] Felix X Yu, Aditya Bhaskara, Sanjiv Kumar, Yunchao Gong, and Shih-Fu Chang. On binary embedding using circulant matrices. *arXiv preprint arXiv:1511.06480*, 2015.

[24] Felix X Yu, Sanjiv Kumar, Yunchao Gong, and Shih-Fu Chang. Circulant binary embedding. *arXiv preprint arXiv:1405.3162*, 2014.

A Standard results

Lemma A.1 *Given vectors $\boldsymbol{v}, \boldsymbol{u}$, $f(\boldsymbol{v}) = \|diag(\boldsymbol{u})\boldsymbol{v}\|_{\ell_2}$ is $\|\boldsymbol{u}\|_{\ell_\infty}$ Lipschitz function.*

Lemma A.2 (Lipschitz concentration of Gaussians) *If $f : \mathbb{R}^n \to \mathbb{R}$ is an L-Lipschitz function, for a standard Gaussian vector $\boldsymbol{g}$ $\mathbb{P}(|f(\boldsymbol{g}) - \mathbb{E}[f(\boldsymbol{g})]| > t) \leq 2\exp(-t^2/(2L^2))$.*

Lemma A.3 *Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k \in \mathbb{R}^n$ be vectors satisfying $\|\boldsymbol{v}_i\|_{\ell_2} \leq \ell$. Let $\boldsymbol{V} = [\boldsymbol{v}_1 \ \ldots \ \boldsymbol{v}_k]^*$ and $\boldsymbol{g} \sim \mathcal{N}(0, \mathbf{I}_n)$. Then, we have that*

$$\mathbb{P}(\|\boldsymbol{V}\boldsymbol{g}\|_{\ell_2}^2 \geq \|\boldsymbol{V}\|_F^2 + t) \leq \exp(-c\min\{\frac{t^2}{\|\boldsymbol{V}\|_F^2\|\boldsymbol{V}\|^2}, \frac{t}{\|\boldsymbol{V}\|^2}\}).$$

Proof Let $\boldsymbol{V}$ have singular value decomposition $\boldsymbol{U}_L\Sigma\boldsymbol{U}_R^*$ where $\Sigma \in \mathbb{R}^{k\times k}$. $\boldsymbol{U}_R^*\boldsymbol{g} \sim \boldsymbol{g}$ and $\boldsymbol{U}_L$ does not affect the ℓ_2 norm. Hence $\|\boldsymbol{U}_L\Sigma\boldsymbol{U}_R^*\|_{\ell_2}^2 \sim \|\Sigma\boldsymbol{g}\|_{\ell_2}^2$ which is a weighted sum of subexponentials where weights are at most $\|\boldsymbol{V}\|^2$. Denoting the ith weight by $w_i = \sigma_i(\boldsymbol{V})^2$ we have that

$$\sum_{i=1}^k w_i = \|\boldsymbol{V}\|_F^2, \quad \sup_{1\leq i\leq w_i} \leq \|\boldsymbol{V}\|^2.$$

Subject to these constraints, we are interested in finding $\sum_{i=1}^k w_i^2$. Observe that if $a > b > c$ we have that $(a+c)^2 + (b-c)^2 > a^2 + b^2$. Consequently, without losing generality, we can assume that nonzero singular values are as large as possible, namely $\|\boldsymbol{V}\|$ so that there are $\frac{\|\boldsymbol{V}\|_F^2}{\|\boldsymbol{V}\|^2}$ nonzero values equal to $\|\boldsymbol{V}\|$.

$$\sum_{i=1}^k w_i^2 \leq \frac{\|\boldsymbol{V}\|_F^2}{\|\boldsymbol{V}\|^2}\|\boldsymbol{V}\|^4 = \|\boldsymbol{V}\|_F^2\|\boldsymbol{V}\|^2.$$

With this bound, Proposition 5.16 of [20] yields the desired result. ■

Lemma A.4 *Let $\boldsymbol{a}, \boldsymbol{b}$ be two unit vectors obeying $\max\{\|\boldsymbol{a}\|_{\ell_\infty}, \|\boldsymbol{b}\|_{\ell_\infty}\} \leq \rho$. Let θ be the angle in between. Let $\boldsymbol{b}' = \boldsymbol{b} - \boldsymbol{a}\boldsymbol{a}^*\boldsymbol{b}$. We have that $\|\bar{\boldsymbol{b}}'\|_{\ell_\infty} \leq \frac{2\rho}{\sin(\theta)}$.*

Proof Clearly $\|\boldsymbol{b}'\|_{\ell_\infty} \leq \|\boldsymbol{b}\|_{\ell_\infty} + \|\boldsymbol{a}^*\boldsymbol{b}\boldsymbol{a}\|_{\ell_\infty} \leq 2\rho$. On the other hand since the angle between is θ, $\|\boldsymbol{b}'\|_{\ell_2} = \sin(\theta)$. ■

Lemma A.5 *Let $\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{y}, \boldsymbol{y}'$ be unit length vectors satisfying $|\boldsymbol{x}^*\boldsymbol{y} - \boldsymbol{x}'^*\boldsymbol{y}'| \leq \alpha$. We have that $|ang(\boldsymbol{x}, \boldsymbol{y}) - ang(\boldsymbol{x}', \boldsymbol{y}')| \leq 5\sqrt{\alpha}$. We also have that for a unit vector $\boldsymbol{x}$ and a perturbation $\boldsymbol{v}$, $ang(\boldsymbol{x}, \boldsymbol{x} + \boldsymbol{v}) \leq 5\|\boldsymbol{v}\|_{\ell_2}$.*

Proof Without losing generality, let $\theta = \mathrm{ang}(\boldsymbol{x}, \boldsymbol{y})$ and $\theta' = \mathrm{ang}(\boldsymbol{x}', \boldsymbol{y})$ where $0 \leq \theta \leq \theta' < \pi$. We are given that

$$\cos(\theta) - \cos(\theta') = \int_\theta^{\theta'} \sin(x)dx \leq \alpha.$$

Using the fact that $\sin(x)$ is increasing over $[0, \pi/2]$ and decreasing over $[\pi/2, \pi]$, we have that

$$2\int_0^{(\theta'-\theta)/2} \sin(x)dx \leq \int_\theta^{\theta'} \sin(x)dx.$$

If $\theta' - \theta < \pi/2$, using the fact that $\sin(x)/x$ is decreasing over $[0, \pi/2]$

$$2\int_0^{(\theta'-\theta)/2} \sin(x)dx \geq (2\sqrt{2}\pi^{-1})^2(\theta' - \theta)^2 \geq 0.9^2(\theta' - \theta)^2.$$

This implies $0.9^2(\theta' - \theta)^2 \leq \alpha$. Otherwise, $\alpha \geq 2\int_0^{(\theta'-\theta)/2} \sin(x)dx \geq 2\int_0^{\pi/4} \sin(x)dx \geq 0.5 \implies \sqrt{\alpha} \geq 0.7$. On the other hand $\theta' - \theta < \pi$ which implies $\theta' - \theta < (\pi/0.7)\sqrt{\alpha}$. Consequently $\theta' - \theta \leq \max\{\pi/0.7, 1.2\}\sqrt{\alpha}$.

Suppose $\boldsymbol{x}$ is a unit length vector and $\boldsymbol{x}''$ be the projection of $\boldsymbol{x}$ on $\boldsymbol{x}'$. Clearly $\|\boldsymbol{x}' - \boldsymbol{x}\|_{\ell_2} \geq \|\boldsymbol{x}'' - \boldsymbol{x}\|_{\ell_2}$ and $\mathrm{ang}(\boldsymbol{x}'', \boldsymbol{x}) = \mathrm{ang}(\boldsymbol{x}', \boldsymbol{x})$. $\|\boldsymbol{x}'' - \boldsymbol{x}\|_{\ell_2}$ has a simple form namely it is equal to $\sin(\theta)$. Now if $\theta < \pi/4$, $\sin(\theta) > 2\sqrt{2}\pi^{-1}\theta$ so that $\mathrm{ang}(\boldsymbol{x}', \boldsymbol{x}) \leq 2\sqrt{2}\pi^{-1}\|\boldsymbol{x}' - \boldsymbol{x}\|_{\ell_2}$. If $\theta > \pi/4$, $\|\boldsymbol{x}' - \boldsymbol{x}\|_{\ell_2} \geq \sqrt{1/2}$ and $\theta \leq \pi$ which implies $\theta \leq \sqrt{2}\pi\|\boldsymbol{x} - \boldsymbol{x}'\|_{\ell_2}$. ■

B Results on random matrices

Lemma B.1 *Let $\boldsymbol{g} \in \mathbb{R}^n$ be a standard Gaussian vector and E be an event over $\boldsymbol{g}$ that holds with probability p. We have that*

$$\mathbb{E}[\|\boldsymbol{g}\|_{\ell_2}^2 \big| E]\mathbb{P}(E) \leq (9n + 2\log p^{-1})p.$$

Setting $p = n^{-3}$ yields right hand side is at most $15n^{-2}$.

Proof Let $r > 0$ be the number for which $\mathbb{P}(\|\boldsymbol{g}\|_{\ell_2} \geq r) = p$ and L be the associated event. Then $\mathbb{P}(E \cap \bar{L}) = \mathbb{P}(\bar{E} \cap L)$ and

$$\mathbb{E}[\|\boldsymbol{g}\|_{\ell_2}^2 \big| E \cap \bar{L}] \leq \mathbb{E}[\|\boldsymbol{g}\|_{\ell_2}^2 \big| \bar{E} \cap L].$$

This implies that

$$\mathbb{E}[\|\boldsymbol{g}\|_{\ell_2}^2 \big| E]p \leq \mathbb{E}[\|\boldsymbol{g}\|_{\ell_2}^2 \big| L]p = \int_{\|\boldsymbol{g}\|_{\ell_2} > r} \|\boldsymbol{g}\|_{\ell_2}^2 \, d\boldsymbol{g}.$$

Let $a = \|\boldsymbol{g}\|_{\ell_2}$ and $p(t)$ be the density function of a and $Q(t) = \mathbb{P}(a > t)$. Using Lipschitzness of ℓ_2 norm, we have that for $t > \sqrt{n}$, $Q(t) \leq \exp(-(t - \sqrt{n})^2)$.

$$\mathbb{E}[\|\boldsymbol{g}\|_{\ell_2}^2 \big| L]p = \int_{a>r} a^2 p(a)da = -\int_{a>r} a^2 dQ(a) = \int_{a>r} Q(a)da^2 - [Q(a)a^2]_r^\infty \qquad = \int_{a>r} Q(a)da^2 + Q(r)r^2.$$

We also have that

$$\int_{a>r} p(a)da = Q(r) \leq \exp(-(r - \sqrt{n})^2).$$

which implies $p \leq \exp(-(r - \sqrt{n})^2) \implies r \leq \sqrt{\log p^{-1}} + \sqrt{n}$. Construct an alternative distribution where $p'(a) = p(a)$ for $a \leq r$, $p'(a) = 0$ for $r < a \leq \sqrt{\log p^{-1}} + \sqrt{n}$ and $Q'(a) = \exp(-(a - \sqrt{n})^2)$ for $a > \sqrt{\log p^{-1}} + \sqrt{n}$. This choice ensures that $Q'(a) > Q(a)$ for all $a \geq 0$ hence

$$\int_{a>r} a^2 p'(a)da = \int_{a>r} Q'(a)da^2 + Q'(r)r^2 \geq \int_{a>r} Q(a)da^2 + Q(r)r^2.$$

Consequently, we will use Q' to upper bound the Gaussial tail. We have that

$$\int_{a>r} a^2 p'(a)da = \int_{a>\sqrt{\log p^{-1}}+\sqrt{n}} a^2 p'(a)da.$$

Finally, we need to estimate the right hand side. For $r' = \sqrt{n} + c$ and $c \geq 1$, we have that

$$\int_{a>r'} 2aQ'(a)da = \int_{u>c} 2(\sqrt{n}+u)\exp(-u^2/2)du = 7\sqrt{n}\exp(-c^2/2).$$

We also have the estimate $Q(r')r'^2 \leq 2(n+c^2)\exp(-c^2/2)$. Setting $c = \sqrt{\log p^{-1}}$ and $p = n^{-3}$ we find that

$$\mathbb{E}[\|\boldsymbol{g}\|_{\ell_2}^2 \big| L]p \leq (9n + 2\log p^{-1})p \leq 15n^{-2}.$$

■

Lemma B.2 (Infinity norm of random modulation) *Let $\{\boldsymbol{v}_i\}_{i=1}^N$ be a finite set of points. Let $\boldsymbol{b} \in \mathbb{R}^n$ be a vector with independent Rademacher entries and let $\boldsymbol{U} \in \mathbb{R}^n$ be the unitary Hadamard matrix where entries are $\pm\sqrt{1/n}$. Let $\boldsymbol{w}_i = \boldsymbol{U}\,diag(\boldsymbol{b})\boldsymbol{v}_i$. With probability $1 - \exp(-c_0 \log N)$, for all $1 \leq i \leq N$, we have that*

$$\sup_{1\leq i\leq N} \|\boldsymbol{w}_i\|_{\ell_\infty} \leq \frac{\sqrt{\log n} + \sqrt{\log N}}{\sqrt{n}}.$$

Proof Observe that $\boldsymbol{w}_i = \boldsymbol{U}\mathrm{diag}(\boldsymbol{v}_i)\boldsymbol{b}$ hence each entry of $\boldsymbol{w}_i$ is a weighted linear combination of subgaussians where the weights are $\boldsymbol{U}_{jk}\boldsymbol{v}_{ki} = \pm\boldsymbol{v}_{ki}/\sqrt{n}$. In particular $\sum_{k=1}^n |\boldsymbol{U}_{jk}\boldsymbol{v}_{ki}|^2 = 1/n$ hence $\boldsymbol{w}_{ij}$ has $\mathcal{O}(1/\sqrt{n})$ subgaussian norm. Consequently for any $1 \leq j \leq n$

$$\mathbb{P}(|\boldsymbol{w}_{ij}| \geq t) \leq \exp(-cnt^2/2).$$

Pick $t = c'(\frac{\log n + \log N}{n})^{1/2}$ and apply a union bound over all $1 \leq j \leq n$ and all $1 \leq i \leq N$ to conclude. ■

Lemma B.3 (Embedding most vectors) *Let $\{\boldsymbol{v}_i\}_{i=1}^N$ be a finite set of points. Let $\boldsymbol{b} \in \mathbb{R}^n$ be a vector with independent Rademacher entries and let $\boldsymbol{U} \in \mathbb{R}^n$ be the unitary Hadamard matrix where entries are $\pm\sqrt{1/n}$. Let $\boldsymbol{w}_i = \boldsymbol{U}\,diag(\boldsymbol{b})\boldsymbol{v}_i$. With probability $1-p$, for at least $(1 - cp^{-1}n^{-3})N$ points $\boldsymbol{w}_i$ $(1 \leq i \leq N)$, we have that*

$$\sup_{1\leq i\leq N} \|\boldsymbol{w}_i\|_{\ell_\infty} \leq C\frac{\sqrt{\log n}}{\sqrt{n}}.$$

Proof Observe that $\boldsymbol{w}_i = \boldsymbol{U}\mathrm{diag}(\boldsymbol{v}_i)\boldsymbol{b}$ hence each entry of $\boldsymbol{w}_i$ is a weighted linear combination of subgaussians where the weights are $\boldsymbol{U}_{jk}\boldsymbol{v}_{ki} = \pm\boldsymbol{v}_{ki}/\sqrt{n}$. In particular $\sum_{k=1}^n |\boldsymbol{U}_{jk}\boldsymbol{v}_{ki}|^2 = 1/n$ hence $\boldsymbol{w}_{ij}$ has $\mathcal{O}(1/\sqrt{n})$ subgaussian norm. Consequently for any $1 \leq j \leq n$

$$\mathbb{P}(|\boldsymbol{w}_{ij}| \geq t) \leq \exp(-cnt^2).$$

Pick $t = c'(\frac{\log(n)}{n})^{1/2}$ to ensure that $\mathbb{P}(|\boldsymbol{w}_{ij}| \geq t) \leq c''n^{-4}$. Applying a union bound over the entries, this ensures $\mathbb{P}(\|\boldsymbol{w}_i\|_{\ell_\infty} \geq t) \leq c''n^{-3}$. Let N_s be the number of $\boldsymbol{w}_i$ obeying the bound $\|\boldsymbol{w}_i\|_{\ell_\infty} \leq t$. We have that

$$\mathbb{E}[N_s] \geq 1 - c''n^{-3}.$$

Hence $N - N_s$ is a nonnegative random variable obeying $\mathbb{E}[N - N_s] \leq c''n^{-3}$. Applying Markov's inequality $\mathbb{P}(N - N_s > p^{-1}c''n^{-3}) \leq p$. ■

Lemma B.4 *Let $\boldsymbol{A}$ be a random matrix with unit length columns. Suppose $\mathbb{E}\,\|\boldsymbol{A}^*\boldsymbol{A} - \mathbf{I}\| \leq \alpha$. Let $\boldsymbol{B} = \boldsymbol{A}\,diag(\alpha)$ where α is a diagonal matrix whose entries lie between $\sqrt{1\pm\varepsilon}$ and α is allowed to depend on $\boldsymbol{A}$. We have that*

$$\mathbb{E}\,\|\boldsymbol{B}^*\boldsymbol{B} - \mathbf{I}\| \leq 2\alpha + \varepsilon.$$

Proof Let $\phi = \|\boldsymbol{A}^*\boldsymbol{A} - \mathbf{I}\|$. We have that $\|\boldsymbol{B}\|^2 \leq (1+\phi)(1+\varepsilon)$ and $\sigma_{\min}(\boldsymbol{B})^2 \geq (1-\phi)(1-\varepsilon)$. Consequently

$$\mathbb{E}[\|\boldsymbol{B}\|^2 - 1] \leq \mathbb{E}[\phi] + \varepsilon + \mathbb{E}[\phi\varepsilon],\ 1 - \sigma_{\min}(\boldsymbol{B})^2 \leq \mathbb{E}[\phi] + \varepsilon - \mathbb{E}[\phi\varepsilon].$$

■

C Generalizations of Tropp's "Incoherent Subdictionary Theorem"

Definition C.1 *$\Phi \in \mathbb{R}^{n\times 2n}$ is a dictionary with coherence μ where $\mu = \sup_{i\neq j} |\phi_i^* \phi_j|$.*

Definition C.2 (Restriction) *$\boldsymbol{R} \in \mathbb{R}^{m\times m_1}$ is called a restriction operator if $\boldsymbol{AR} \in \mathbb{R}^{n\times m_1}$ is a matrix obtained by selecting m_1 columns of $\boldsymbol{A}$ for any $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ and any $n \geq 1$. If $\boldsymbol{R}$ select m_1 columns uniformly at random, we shall call it random restriction. A random subdictionary of $\boldsymbol{A}$ is obtained by applying the restriction $\boldsymbol{R}$ to get $\boldsymbol{AR}$.*

Define $\|\cdot\|_{1,2}$ norm of a matrix to be the largest ℓ_2 norm of its columns. The next result will be beneficial for the derivation.

Theorem C.3 (Theorem 8 of [19]) *Let $\boldsymbol{A}$ be a matrix with N columns and let $\boldsymbol{R}$ be a restriction to m coordinates chosen uniformly at random. Fix $q \geq 1$. For any $p \geq \max\{2, 2\log(rank\boldsymbol{AR}^*), q/2\}$ we have that*

$$(\mathbb{E}\,\|\boldsymbol{AR}^*\|^q)^{1/q} \leq 3\sqrt{p}\|\boldsymbol{A}\|_{1,2} + \sqrt{m/N}\|\boldsymbol{A}\|.$$

Observe that $(a+b)^q \leq (2\max(a+b))^q \leq (2a)^q + (2b)^q$ hence

$$\mathbb{E}\,\|\boldsymbol{AR}^*\|^q \leq (6\sqrt{p}\|\boldsymbol{A}\|_{1,2})^q + (2\sqrt{m/N}\|\boldsymbol{A}\|)^q.$$

The following is our variation of Tropp's spectral norm bounds on incoherent subdictionaries.

Theorem C.4 *Suppose $\Phi \in \mathbb{R}^{n\times 2n}$ is a random matrix such that all of its realizations are incoherent dictionaries with coherence μ. Pick a random subdictionary $\boldsymbol{X} \in \mathbb{R}^{n\times 2k}$ of Φ. Define the function $f(\boldsymbol{R}) = \mathbb{E}_\Phi \|\boldsymbol{X}^*\boldsymbol{X} - \mathbf{I}\|$. For $u \geq \sqrt{2\log n + 1}$, we have that*

$$\mathbb{P}_{\boldsymbol{R}}(f(\boldsymbol{R}) \geq c'(u\sqrt{m(\log m + 1)}\mu + \frac{m}{n}\|\Phi\|^2)) \leq \exp(-u^2/4). \tag{C.1}$$

Proof The proof exactly follows the work by Tropp, namely Section 6 of [19]. We will only point out the main differences as almost all of the argument overlaps. Let $\boldsymbol{R}$ be the random restriction for which $\boldsymbol{X} = \Phi\boldsymbol{R}$. We first establish the following result.

Theorem C.5 *For $q \geq 2\log n + 1$, we have that*

$$(\mathbb{E}_{\boldsymbol{R}}(\mathbb{E}_\Phi \|\boldsymbol{X}^*\boldsymbol{X} - \mathbf{I}\|)^q)^{1/q} \leq (\mathbb{E}_{\boldsymbol{R},\Phi} \|\boldsymbol{X}^*\boldsymbol{X} - \mathbf{I}\|^q)^{1/q} \leq c(\sqrt{qm(\log m + 1)}\mu + \frac{m}{n}\|\Phi\|^2).$$

Proof For the sake of completeness, we repeat most of the arguments in [19]. First note that $\boldsymbol{X}^*\boldsymbol{X} - \mathbf{I} = \boldsymbol{RHR}^*$ where $\boldsymbol{H} = \Phi^*\Phi - \mathbf{I}$. A standard symmetrization argument (Theorem 9 of [19]) ensures that there exists a submatrix $\hat{\boldsymbol{H}} \in \mathbb{R}^{n/2\times n/2}$ (where columns and rows correspond to disjoint subsets) and restrictions $\boldsymbol{R}_1, \boldsymbol{R}_2$ such that

$$(\mathbb{E}_{\boldsymbol{R}} \|\boldsymbol{X}^*\boldsymbol{X} - \mathbf{I}\|^q)^{1/q} \leq 2(\max_{m_1+m_2=m} \mathbb{E}_{\boldsymbol{R}_1,\boldsymbol{R}_2} \|\boldsymbol{R}_1\hat{\boldsymbol{H}}\boldsymbol{R}_2\|^q)^{1/q}. \tag{C.2}$$

Exponentiating both sides, this implies

$$\mathbb{E}_{\boldsymbol{R}} \|\boldsymbol{X}^*\boldsymbol{X} - \mathbf{I}\|^q \leq 2^q \max_{m_1+m_2=m} \mathbb{E}_{\boldsymbol{R}_1,\boldsymbol{R}_2} \|\boldsymbol{R}_1\hat{\boldsymbol{H}}\boldsymbol{R}_2\|^q.$$

Hence, we shall upper bound the right-hand side. This will be done in two steps by first taking expectation over $\boldsymbol{R}_2$ and then $\boldsymbol{R}_1$.

$$\mathbb{E}_{\boldsymbol{R}_1,\boldsymbol{R}_2} \|\boldsymbol{R}_1\hat{\boldsymbol{H}}\boldsymbol{R}_2\|^q = \mathbb{E}_{\boldsymbol{R}_1}\,\mathbb{E}_{\boldsymbol{R}_2} \|\boldsymbol{R}_1\hat{\boldsymbol{H}}\boldsymbol{R}_2\|^q.$$

Applying Theorem C.3 with $p = \max\{2, 2\log(m/2) + 1, q/2\}$ we have that

$$\mathbb{E}_{\boldsymbol{R}_2} \|\boldsymbol{R}_1\hat{\boldsymbol{H}}\boldsymbol{R}_2\|^q \leq (6\sqrt{p}\|\boldsymbol{R}_1\hat{\boldsymbol{H}}\|_{1,2})^q + (\sqrt{8m_2/n}\|\boldsymbol{R}_1\hat{\boldsymbol{H}}\|)^q.$$

Similar to Tropp, the coherence assumption ensures that $\|\boldsymbol{R}_1\hat{\boldsymbol{H}}\|_{1,2} \leq \mu\sqrt{m}$ to obtain

$$\mathbb{E}_{\boldsymbol{R}_2} \|\boldsymbol{R}_1\hat{\boldsymbol{H}}\boldsymbol{R}_2\|^q \leq (6\sqrt{p}\mu\sqrt{m})^q + (\sqrt{8m_2/n}\|\boldsymbol{R}_1\hat{\boldsymbol{H}}\|)^q.$$

The remaining task is to upper bound $\mathbb{E}\,\|\boldsymbol{R}_1\hat{\boldsymbol{H}}\|^q$. Reapplying Theorem C.3, we have that

$$\mathbb{E}\,\|\boldsymbol{R}_1\hat{\boldsymbol{H}}\|^q = \mathbb{E}\,\|\hat{\boldsymbol{H}}^*\boldsymbol{R}_1^*\|^q \le (6\sqrt{p}\mu\sqrt{n})^q + (\sqrt{8m_1/n}\|\hat{\boldsymbol{H}}\|)^q.$$

The combination of the last two inequalities, yields that, for any Φ obeying the coherence and spectral norm bounds, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{R}_1,\boldsymbol{R}_2}\,\|\boldsymbol{R}_1\hat{\boldsymbol{H}}\boldsymbol{R}_2\|^q &\le (6\sqrt{p}\mu\sqrt{m})^q + (\sqrt{8m_2/n}6\sqrt{p}\mu\sqrt{n})^q + (\sqrt{8m_2/n}\sqrt{8m_1/n}\|\hat{\boldsymbol{H}}\|)^q \\ &\le (c_1\sqrt{pm}\mu)^q + (c_2\frac{m}{n}\|\hat{\boldsymbol{H}}\|)^q.\end{aligned}$$

Since this holds for all realizations of Φ, we can take an additional expectation over Φ to conclude

$$\mathbb{E}_{\boldsymbol{R},\Phi}\,\|\boldsymbol{X}^*\boldsymbol{X}-\mathbf{I}\|^q \le (c_1\sqrt{pm}\mu)^q + (c_2\frac{m}{n}\|\hat{\boldsymbol{H}}\|)^q \implies (\mathbb{E}_{\boldsymbol{R},\Phi}\,\|\boldsymbol{X}^*\boldsymbol{X}-\mathbf{I}\|^q)^{1/q} \le c(\sqrt{pm}\mu + \frac{m}{n}\|\hat{\boldsymbol{H}}\|).$$

For $q \ge 1$, this also implies that

$$(\mathbb{E}_{\boldsymbol{R}}(\mathbb{E}_\Phi\,\|\boldsymbol{X}^*\boldsymbol{X}-\mathbf{I}\|)^q)^{1/q} \le (\mathbb{E}_{\boldsymbol{R},\Phi}\,\|\boldsymbol{X}^*\boldsymbol{X}-\mathbf{I}\|^q)^{1/q} \le c(\sqrt{pm}\mu + \frac{m}{n}\|\hat{\boldsymbol{H}}\|).$$

Picking $p = q(\log m/2+1)$, and using the estimate $\|\hat{\boldsymbol{H}}\| \le \|\boldsymbol{H}\| \le \|\Phi^*\Phi\| + 1 \le 2\|\Phi\|^2$ we obtain that

$$(\mathbb{E}_{\boldsymbol{R}}(\mathbb{E}_\Phi\,\|\boldsymbol{X}^*\boldsymbol{X}-\mathbf{I}\|)^q)^{1/q} \le (\mathbb{E}_{\boldsymbol{R},\Phi}\,\|\boldsymbol{X}^*\boldsymbol{X}-\mathbf{I}\|^q)^{1/q} \le c(\sqrt{qm(\log m+1)}\mu + \frac{m}{n}\|\Phi\|^2).$$

■

Now letting $f(\boldsymbol{R}) = \mathbb{E}_\Phi\,\|\boldsymbol{X}^*\boldsymbol{X}-\mathbf{I}\|$ and applying Proposition 10 of [19], for $u \ge 1$, we obtain that

$$\mathbb{P}(f(\boldsymbol{R}) \ge c'(u\sqrt{m(\log m+1)}\mu + \frac{m}{n}\|\Phi\|^2)) \le \exp(-u^2/4)$$

which is the desired concentration bound. ■

C.1 Asymmetric version of Tropp's incoherent subdictionary result

We first prove the following variation of Theorem 25 of [19]. This result assumes the matrix to have even dimensions but the odd case can be shown with minimal modification of the proof strategy. The proof exactly follows the argument of Tropp however we will provide it here for the sake of completeness. We remark that Tropp's result was based on more classical results due to Bourgain and Tzafriri [3, 12].

Theorem C.6 *Let $\boldsymbol{A}$ be a $2n \times 2n$ matrix with a 0 diagonal. Let $\boldsymbol{R}$ be a restriction to m random coordinates. Fix $q \ge 1$. There exists a partition of the coordinates $\{1,2,\ldots,2n\}$ into two blocks T_1 and T_2 with N elements each so that*

$$\mathbb{E}\,\|\boldsymbol{R}\boldsymbol{A}\boldsymbol{R}^*\|^q \le 2^q \max_{m_1+m_2=m} \mathbb{E}\,\|\boldsymbol{R}_1\boldsymbol{A}_{T_1\times T_2}\boldsymbol{R}_2\|^q.$$

Proof We prove the result for $q=1$. Identical argument applies to $q>1$ case. Let a_{ij} denote the ijth coordinate of the matrix $\boldsymbol{A}$ and e_i denote the ith vector of the standard basis. Define the matrices $\boldsymbol{B}_{jk} = a_{jk}\mathrm{e}_j\mathrm{e}_k^*$. $\delta \in \{0,1\}^{2n}$, be a vector which has exactly m components equal to 1. Then, we have

$$\boldsymbol{R}\boldsymbol{A}\boldsymbol{R}^* = \sum_{j\ne k}\delta_j\delta_k\boldsymbol{B}_{jk}$$

We wish to bound the expectation

$$M = \mathbb{E}_\delta\,\|\sum_{j\ne k}\delta_j\delta_k\boldsymbol{B}_{jk}\|$$

Let $\eta \in \mathbb{R}^{2n}$ be a random vector with exactly n coordinates equal to 1. For $j \ne k$, we have that

$$\mathbb{E}_\eta[\eta_j(1-\eta_k)+\eta_k(1-\eta_j)] = \frac{n}{2n-1}. \tag{C.3}$$

Now, based on these, applying Jensen's inequality, we have the following list of inequalities

$$\begin{aligned} M &= \frac{2n-1}{n}\,\mathbb{E}_\delta \|\sum_{j\neq k}\mathbb{E}_\eta[\eta_j(1-\eta_k)+\eta_k(1-\eta_j)]\delta_j\delta_k\boldsymbol{B}_{jk}\| \\ &< 2\,\mathbb{E}_\delta\,\mathbb{E}_\eta\|\sum_{j\neq k}[\eta_j(1-\eta_k)+\eta_k(1-\eta_j)]\delta_j\delta_k\boldsymbol{B}_{jk}\| \\ &< 2[\mathbb{E}_\delta\,\mathbb{E}_\eta\|\sum_{j\neq k}\eta_j(1-\eta_k)\delta_j\delta_k\boldsymbol{B}_{jk}\| + \mathbb{E}_\delta\,\mathbb{E}_\eta\|\sum_{j\neq k}\eta_k(1-\eta_j)\delta_j\delta_k\boldsymbol{B}_{jk}\|] \\ &< 4\,\mathbb{E}_\delta\,\mathbb{E}_\eta\|\sum_{j\neq k}\eta_j(1-\eta_k)\delta_j\delta_k\boldsymbol{B}_{jk}\| \end{aligned}$$

It follows that there exists a $0-1$ vector η^* containing exactly n 1s such that

$$M < 4\,\mathbb{E}_\delta\|\sum_{j\neq k}\eta_j^*(1-\eta_k^*)\delta_j\delta_k\boldsymbol{B}_{jk}\|$$

Note that this vector η^* partitions the set $\{1,2,\dots,2n\}$ into two parts T_1, T_2 each containing N elements. T_1 corresponds to the coordinates $\eta_i = 1$ and T_2 corresponds to the coordinates $\eta_i = 0$. Calling these parts T_1, T_2, we can rewrite the inequality as

$$M < 4\,\mathbb{E}_\delta\|\sum_{j\in T_1, k\in T_2}\delta_j\delta_k\boldsymbol{B}_{jk}\|$$

Next, let number of active coordinates of δ over T_1 be m_1. Observe that conditioned on the choice of T_1 and m_1, the m_1 and $m-m_1$ active coordinates of δ are distributed uniformly at random over T_1 and T_2. This is due to the fact that δ is independent of η^*. With this, the inequality takes the advertised form

$$M < 4\,\mathbb{E}_{m_1,m_2}\|\boldsymbol{R}_1^*\boldsymbol{A}_{T_1\times T_2}\boldsymbol{R}_2\| \le \max_{m_1+m_2=m} 4\,\mathbb{E}\|\boldsymbol{R}_1^*\boldsymbol{A}_{T_1\times T_2}\boldsymbol{R}_2\|$$

■

Using Theorem C.6 and repeating the proof of Theorem C.4 line by line we can conclude with the following result.

Theorem C.7 *Suppose $\Phi \in \mathbb{R}^{n\times 2n}$ is a random matrix such that all of its realizations are incoherent dictionaries with coherence μ. Pick a random subdictionary $\boldsymbol{X}_1 \in \mathbb{R}^{n\times k}$ from first n columns of Φ. Pick the same k coordinates from the second n columns of Φ and form $\boldsymbol{X}_2$. Define the function $f(\boldsymbol{R}) = \mathbb{E}_\Phi\|\boldsymbol{X}_1^*\boldsymbol{X}_2\|$. For $u \ge \sqrt{2\log n+1}$, we have that*

$$\mathbb{P}_{\boldsymbol{R}}(f(\boldsymbol{R}) \ge c'(u\sqrt{m(\log m+1)}\mu + \frac{m}{n}\|\Phi\|^2)) \le \exp(-u^2/4). \tag{C.4}$$

Proof Denote the first and second n columns of Φ by Φ_1 and Φ_2 respectively and set $\Theta = \Phi_1^*\Phi_2$. Let $\boldsymbol{R} \in \mathbb{R}^{n\times k}$ be a random restriction. Observe that

$$\mathbb{E}_\Phi\|\boldsymbol{X}_1^*\boldsymbol{X}_2\| = \mathbb{E}_{\Phi,\boldsymbol{R}}\|\boldsymbol{R}^*\Phi_1^*\Phi_2\boldsymbol{R}\| = \mathbb{E}_{\Phi,\boldsymbol{R}}\|\boldsymbol{R}^*\Theta\boldsymbol{R}\|$$

Next, split the diagonal and off-diagonal entries of Θ and apply the argument in Theorem C.4 where we replace the inequality (C.2) with the estimate of Theorem C.6. ■