On the Optimal Bit Complexity of Circulant Binary Embedding

Saehoon Kim[1,2]
[1]AItrics, Korea
kshkawa@postech.ac.kr

Jungtaek Kim[2], **Seungjin Choi**[2]
[2]Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 37673, Korea
{jtkim, seungjin}@postech.ac.kr

Abstract

Binary embedding refers to methods for embedding points in $\mathbb{R}^d$ into vertices of a Hamming cube of dimension k, such that the normalized Hamming distance well preserves the pre-defined similarity between vectors in the original space. A common approach to binary embedding is to use random projection with unstructured projection, followed by one-bit quantization to produce binary codes, which has been proven that $k = \mathcal{O}\left(\epsilon^{-2} \log n\right)$ is required to approximate the angle up to ϵ-distortion, where n is the number of data. Of particular interest in this paper is *circulant binary embedding* (CBE) with angle preservation, where a random circulant matrix is used for projection. It yields comparable performance while achieving the nearly linear time and space complexities, compared to embedding methods relying on unstructured projection. To support promising empirical results, several non-asymptotic analysis have been introduced to establish conditions on the number of bits to meet ϵ-distortion embedding, where one of state-of-the-art achieves the optimal sample complexity $k = \mathcal{O}\left(\epsilon^{-3} \log n\right)$ while the distortion rate ϵ^{-3} is far from the optimality, compared to $k = \mathcal{O}\left(\epsilon^{-2} \log n\right)$. In this paper, to support promising empirical results of CBE, we extend the previous theoretical framework to address the optimal condition on the number of bits, achieving that CBE with $k = \mathcal{O}\left(\epsilon^{-2} \log n\right)$ approximates the angle up to ϵ-distortion under mild assumptions. We also provide numerical experiments to support our theoretical results.

Table 1: Comparison of the analysis for BE with unstructured projection and circulant projection in terms of bit complexity and conditions necessary to build the analysis, where ϵ is a distortion rate and n is the number of data points.

Methods	Bit Complexity	Conditions
Unstructured BE	$\mathcal{O}\left(\epsilon^{-2} \log n\right)$	-
Our analysis	$\mathcal{O}\left(\epsilon^{-2} \log n\right)$	*small infinity norm*
(Oymak 2016) (Near-optimal)	$\mathcal{O}\left(\epsilon^{-3} \log n\right)$	*small infinity norm*
(Yu et al. 2015) (Near-optimal)	$\mathcal{O}\left(\epsilon^{-2} \log^2 n\right)$	*small infinity norm*

Introduction

Binary embedding (BE) is a nonlinear dimensionality reduction method, where a mapping is determined to relate high-dimensional points in $\mathbb{R}^d$ to binary codes of length $\mathcal{O}(d)$ whose normalized Hamming distance preserves the pre-defined distance in the original space $\mathbb{R}^d$. It becomes a ubiquitous tool for large-scale data analysis, including approximate nearest neighbor search (Charikar 2002), large-scale machine learning (Gong et al. 2013; Yu et al. 2014), and so on. A notable method is angle-preserving binary embedding (Charikar 2002), where an embedding function is constructed by random projection followed by one-bit quantization such that the angular distance between two points is preserved by the normalized Hamming distance. Besides the angle-preserving binary embedding, various methods have been proposed, including the method preserving the similarity specified by shift-invariant kernels (Raginsky and Lazebnik 2009; Kim and Choi 2015) and the maximum inner product search (Shrivastava and Li 2014).

Most of binary embedding methods involving random projection require long codes to achieve satisfactory performance, so time and space complexities become serious concerns. In the case where the code length is $\mathcal{O}(d)$, both time and space complexities require $O(d^2)$ to construct a single binary code, which becomes expensive when the dimension of data d is large. In order to improve the scalability, both in terms of computational cost and space complexity, a few fast binary embedding methods have been recently developed by accelerating matrix-vector multiplication with some structured matrices: a randomized Hadamard matrix (Dasgupta, Kumar, and Sarlós 2011), the Kronecker product of multiple small matrices (Gong et al. 2013; Kim and Choi 2015; Zhang et al. 2015), a circulant matrix (Yu et al. 2014), and two consecutive structured matrices (Yi, Caramanis, and Price 2015; Choromanska et al. 2016). These can be interpreted as nonlinear extensions of random projection with Walsh-Hadamard matrix (Ailon and Chazelle 2009), the Kronecker product of two small matrices (Eftekhari, Babaie-Zadeh, and Moghaddam 2011), and a circulant matrix (Hinrichs and Vybíral 2011).

Of particular interest is the *circulant binary embedding* (CBE) (Yu et al. 2014), which yields comparable performance while reducing time complexity from $\mathcal{O}(d^2)$ to $\mathcal{O}(d \log d)$ and the space complexity from $\mathcal{O}(d^2)$ to $\mathcal{O}(d)$ by employing the randomized circulant matrix. To support such promising results, several theoretical guarantees (Oymak 2016; Yu et al. 2015; Choromanska et al. 2016;

Dirsken and Stollenwerk 2016) have been proposed by introducing non-asymptotic analysis of binary embedding with random structured matrices, which is mainly interpreted as analyzing the number of bits to approximate the angular distance up to ϵ-distortion with high probability. This type of analysis is analogous to the well-studied theory for fast random projection (Ailon and Chazelle 2009; Eftekhari, Babaie-Zadeh, and Moghaddam 2011; Hinrichs and Vybíral 2011), providing conditions on preserving pairwise Euclidean distance in a low-dimensional space up to ϵ-distortion, which is built on Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss 1984).

In case of binary embedding with unstructured projection, (Yi, Caramanis, and Price 2015) provides the optimal bit complexity, showing that $\mathcal{O}(\epsilon^{-2}\log n)$ bits are required to preserve the angular distance up to ϵ-distortion, where n is the number of data points. Similarly, (Oymak 2016; Yu et al. 2015; Choromanska et al. 2016; Dirsken and Stollenwerk 2016) propose the theoretical frameworks to establish conditions on the number of bits, but they achieve only near-optimal bit complexities compared to $\mathcal{O}(\epsilon^{-2}\log n)$. Specifically, (Oymak 2016) obtain the optimal sample complexity, $\mathcal{O}(\epsilon^{-3}\log n)$, but the distortion rate ϵ^{-3} is far from the optimality. In this paper, to support promising empirical results of CBE, we extend the framework (Oymak 2016) to address the optimal condition on the number of bits to meet ϵ-distortion binary embedding. Our work contains the following technical improvements over the previous analyses:

- A non-trivial extension of (Oymak 2016) is developed, where a Gaussian random sequence is replaced by Rademacher entries, i.e. an independent Bernoulli random sequence. It matches the original implementation of CBE (Yu et al. 2014).
- Compared to existing analyses (Choromanska et al. 2016; Dirsken and Stollenwerk 2016; Yu et al. 2015; Oymak 2016), our analysis achieves the optimal complexity of CBE, matching the optimality of unstructured projection in case that ϵ-distortion binary embedding is interested, under reasonable conditions on the number of bits and coherence of datasets, which is summarized in Table 1.

Background

In this section, we briefly review binary embedding methods: (1) standard binary embedding where random projection is performed with an *unstructured* matrix; (2) circulant binary embedding where the randomized circulant matrix is used (Yu et al. 2014). We also review the existing results of theoretical guarantees in the case of binary embedding with unstructured projection.

Binary Embedding: Unstructured Projection

A common approach to binary embedding constitutes a random projection followed by a one-bit quantization, to relate a vector $\boldsymbol{x} \in \mathcal{S}^{d-1}$ to a binary string of length k (Charikar 2002):

$$\begin{aligned} h(\boldsymbol{x}) &\triangleq \operatorname{sgn}\left(\boldsymbol{G}^\top \boldsymbol{x}\right), \\ &= [h_1(\boldsymbol{x}), \ldots, h_k(\boldsymbol{x})]^\top, \end{aligned} \tag{1}$$

where $\mathcal{S}^{d-1}$ is referred to as $(d-1)$-sphere and $\operatorname{sgn}(\cdot)$ is an element-wise one-bit quantizer. Elements of $\boldsymbol{G} = [\boldsymbol{g}_1, \ldots, \boldsymbol{g}_k] \in \mathbb{R}^{d\times k}$ are drawn independently from Gaussian distribution, $\mathcal{N}(0,1)$, with zero mean and unit variance.

It was shown in (Charikar 2002) that the Hamming distance between $h_l(\boldsymbol{x}_i) = \operatorname{sgn}(\boldsymbol{g}_l^\top \boldsymbol{x}_i)$ and $h_l(\boldsymbol{x}_j) = \operatorname{sgn}(\boldsymbol{g}_l^\top \boldsymbol{x}_j)$ is an unbiased estimator of the angular distance between two vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, i.e.,

$$\mathbb{E}\Big[\mathcal{I}\left[h_l(\boldsymbol{x}_i) \neq h_l(\boldsymbol{x}_j)\right]\Big] = \frac{\theta_{\boldsymbol{x}_i,\boldsymbol{x}_j}}{\pi}, \quad l = 1, \cdots, k \tag{2}$$

where $\mathcal{I}[\,\cdot\,]$ is the indicator function which returns 1 whenever the input argument is true and 0 otherwise. The angle between two vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is denoted by $\theta_{\boldsymbol{x}_i,\boldsymbol{x}_j}$ and the angle is normalized by π such that the value of the angular distance lies between 0 and 1.

Definition 1. *The* normalized Hamming distance *between two binary codes $h(\boldsymbol{x}_i) \in \{1,0\}^k$ and $h(\boldsymbol{x}_j) \in \{1,0\}^k$ of length k is defined as the average of k independent Bernoulli indicator variables $\mathcal{I}_l(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathcal{I}\left[h_l(\boldsymbol{x}_i) \neq h_l(\boldsymbol{x}_j)\right]$ ($l = 1, \ldots, k$), i.e.,*

$$d_H(h(\boldsymbol{x}_i), h(\boldsymbol{x}_j)) \triangleq \frac{1}{k}\sum_{l=1}^{k} \mathcal{I}\left[h_l(\boldsymbol{x}_i) \neq h_l(\boldsymbol{x}_j)\right]. \tag{3}$$

Definition 2. *Given $\epsilon \in (0,1)$ and any finite set of d-dimensional vectors, $\mathcal{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$, a mapping $h : \mathcal{S}^{d-1} \to \{0,1\}^k$ is said to be an* ϵ-distortion binary embedding *if*

$$\left| d_H(h(\boldsymbol{x}_i), h(\boldsymbol{x}_j)) - \frac{\theta_{\boldsymbol{x}_i,\boldsymbol{x}_j}}{\pi} \right| \leq \epsilon, \tag{4}$$

for $\forall \boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{D}$.

Of particular interest is the *bit complexity* of ϵ-distortion binary embedding, establishing a certain condition on the number of bits, k, which guarantees Eq. 4, given a dataset $\mathcal{D}$. The bit complexity in the case of Eq. 1 is summarized in the following theorem (Jacques et al. 2013).

Theorem 1. *Given $\epsilon \in (0,1)$ and any finite data set $\mathcal{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \subset \mathcal{S}^{d-1}$, with probability at least $1 - \exp(-c\epsilon^2 k)$, $k = \mathcal{O}\left(\frac{1}{\epsilon^2}\log n\right)$ implies that we have $h : \mathbb{R}^d \to \{0,1\}^k$ such that for all $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{D}$*

$$\left| d_H(h(\boldsymbol{x}_i), h(\boldsymbol{x}_j)) - \frac{\theta_{\boldsymbol{x}_i,\boldsymbol{x}_j}}{\pi} \right| \leq \epsilon,$$

where $c > 0$ is a constant.

Proof. Since the proof is the standard application of Hoeffdings' inequality, we defer it to supplementary material. □

Circulant Binary Embedding: Structured Projection

We briefly review circulant binary embedding (CBE) (Yu et al. 2014). Given a d-dimensional vector $\boldsymbol{g} = [g_1, \ldots, g_d]^\top$, where $\{g_i\}_{i=1}^d$ are assumed to be independently drawn from

Gaussian distribution $\mathcal{N}(0,1)$, a circulant matrix $\boldsymbol{G}_c \in \mathbb{R}^{d\times d} = \mathrm{circ}(\boldsymbol{g})$ is defined as

$$\begin{pmatrix} g_1 & g_d & \cdots & g_3 & g_2 \\ g_2 & g_1 & \cdots & g_4 & g_3 \\ \vdots & g_2 & g_1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots & g_d \\ g_d & g_{d-1} & \cdots & g_2 & g_1 \end{pmatrix}. \tag{5}$$

Given $\boldsymbol{x} \in \mathbb{R}^d$, circulant binary embedding (Yu et al. 2014) uses the randomized circulant matrix Eq. 5 (Hinrichs and Vybíral 2011) to produce a d-bit binary code:

$$h^C(\boldsymbol{x}) = \mathrm{sgn}\left(\boldsymbol{G}_c^\top \boldsymbol{D}\boldsymbol{x}\right), \tag{6}$$

where $\boldsymbol{G}_c \in \mathbb{R}^{d\times d}$ is given in Eq. 5 and $\boldsymbol{D} \in \mathbb{R}^{d\times d}$ is a diagonal matrix with a Rademacher sequence, i.e., diagonal entries being either +1 or -1 (independently drawn from Bernoulli distribution with probability 1/2). Pre-multiply $\boldsymbol{x}$ by $\boldsymbol{D}$ is equivalent to applying random sign flipping to each entry of $\boldsymbol{x}$. Since the sign flipping can be performed as a preprocessing for each data $\boldsymbol{x}$, $\boldsymbol{D}$ is dropped out for the sake of simplicity in the subsequent analysis. It was shown in (Yu et al. 2014) that CBE improves the time complexity from $\mathcal{O}(d^2)$ to $\mathcal{O}(d\log d)$ and the space complexity from $\mathcal{O}(d^2)$ to $\mathcal{O}(d)$. If one desires to select $k(< d)$ bits, one can randomly select k bits in d entries of $h^C(\boldsymbol{x}) \in \mathbb{R}^d$. Resorting to FFT, $h^C(\boldsymbol{x})$ can be efficiently computed by

$$h^C(\boldsymbol{x}) = \mathrm{sgn}\Big(\mathcal{F}^{-1}(\mathcal{F}(\boldsymbol{g}) \odot \mathcal{F}(\boldsymbol{D}\boldsymbol{x})\Big), \tag{7}$$

where $\mathcal{F}(\cdot)$ represents the discrete Fourier transform, $\mathcal{F}^{-1}(\cdot)$ is the inverse discrete Fourier transform, and $\odot$ is element-wise product. Since (inverse) discrete Fourier transform requires $O(d\log d)$ in time, the time complexity of CBE is $O(d\log d)$.

Note that in the case of CBE, Bernoulli indicator variables $\left\{\mathcal{I}_l^C(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathcal{I}\left[h_l^C(\boldsymbol{x}_i) \neq h_l^C(\boldsymbol{x}_j)\right]\right\}_{l=1}^d$ are not marginally independent, in contrast to the unstructured projection, where $h_l^C(\boldsymbol{x})$ represents the l-th bit of $h^C(\boldsymbol{x})$ defined in Eq. 6. This imposes a technical challenge for the conditions to meet ϵ-distortion embedding. The normalized Hamming distance between two binary codes $h^C(\boldsymbol{x}_i)$ and $h^C(\boldsymbol{x}_j)$ is defined as

$$d_H\left(h^C(\boldsymbol{x}_i), h^C(\boldsymbol{x}_j)\right) = \frac{1}{k}\sum_{l=1}^{k} \mathcal{I}_l^C(\boldsymbol{x}_i, \boldsymbol{x}_j). \tag{8}$$

For the sake of simplicity, we omit input arguments in $\mathcal{I}_l^C(\boldsymbol{x}_i, \boldsymbol{x}_j)$, i.e., use $\mathcal{I}_l^C$ in obvious cases.

Motivation

Our analysis is motivated by empirical success of CBE (Yu et al. 2014) and recently proposed theoretical analysis to establish *large deviation* theory (Yi, Caramanis, and Price 2015; Yu et al. 2015; Choromanska et al. 2016; Dirsken and Stollenwerk 2016; Oymak 2016), leading to the bit complexity. In this section, we discuss the technical challenge and limitation of the previous works to enhance our motivation.

Among the theoretical work to analyze CBE, the difficulty is raised by the statistical dependence between indicator variables. Formally, due to the dependence between the column vectors in $\mathrm{cir}(\boldsymbol{g})$, the Bernoulli indicator variables, $\{\mathcal{I}_l^C\}_{l=1}^k$ are not marginally independent, which imposes a technical challenge to establish certain conditions on the number of bits required to meet ϵ-distortion binary embedding. To alleviate the dependence, most of works implicitly or explicitly make use of the asymptotic behavior of indicator variables, in which the dependence becomes very weak, as the data dimension grows (see Figure 1). This interesting phenomenon leads to develop promising theoretical frameworks (Yi, Caramanis, and Price 2015; Yu et al. 2015; Oymak 2016; Choromanska et al. 2016; Dirsken and Stollenwerk 2016) to analyze the behavior of CBE. Unfortunately, these frameworks have the following limitations:

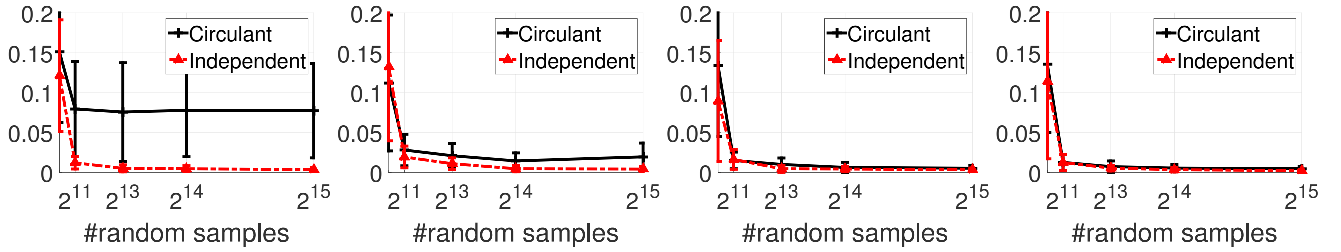
- Near-optimal complexity. (Yu et al. 2015; Choromanska et al. 2016; Oymak 2016) only achieve near-optimal bit complexities, motivating us to establish an analysis towards the optimality. Specifically, given restricted geometrical configurations, (Yu et al. 2015) shows that $k = \mathcal{O}(\epsilon^{-2}\log^2 n)$ implies that CBE meets ϵ-distortion binary embedding. (Oymak 2016) follows the similar framework of (Yu et al. 2015) to achieve $k = \mathcal{O}(\epsilon^{-3}\log n)$ under mild conditions, where the optimal sample complexity is introduced but the distortion rate ϵ^{-3} is far from the optimality.
- One-layer vs. Two-layer binary embedding. (Yi, Caramanis, and Price 2015; Dirsken and Stollenwerk 2016) propose two-layer binary embedding by applying additional random projection with structured matrices. Under mild conditions, this direction achieves the same time and space complexities with the almost optimal bit complexity [1] (Yi, Caramanis, and Price 2015; Dirsken and Stollenwerk 2016). We believe, however, that an additional random projection procedure is absolutely not necessary for the optimal complexity, supported by theoretical and empirical analysis to be presented in the subsequent sections.

Main Results for Optimal Bit Complexity

In this section, we develop a theoretical analysis to address that how many bits are required for CBE to meet ϵ-distortion binary embedding. Our work contains the following technical improvements over previous analysis:

- We extend the framework (Oymak 2016) for the standard CBE, where it assumes that the diagonal entries of $\boldsymbol{D}$ in

[1] As pointed out in (Dirsken and Stollenwerk 2016), the analysis presented in (Yi, Caramanis, and Price 2015) implicitly assumes that the indicator variables are pairwise independent, which is not satisfied in case of a small-dimensional space. This phenomenon has been independently observed as in Figure 1.



(a) dim=5 (b) dim=10 (c) dim=20 (d) dim=30

Figure 1: Empirical evidence for the dependence between $\{\exp(\mathcal{I}_i)\}_{i=1}^m$ with respect to various data dimensions, where $m = 5$ and the angle between two points is $\frac{\pi}{2}$. The vertical line represents $\mathbb{E}\left[\prod_{i=1}^5 \exp(\mathcal{I}_i)\right] - \prod_{i=1}^5 \mathbb{E}\left[\exp(\mathcal{I}_i)\right]$, where the expectation is approximated by the sample average. Black solid line means that the binary codes are generated by CBE and red dotted line means BE with unstructured projection.

Eq. 6 follow i.i.d. isotropic Gaussian distribution. We replace the Gaussian random sequence into the Rademacher entries, i.e., an independent Bernoulli sequence with equal probability. We observe that the Rademacher sequence is more natural for CBE with superior empirical results to the Gaussian random sequence, introduced in experiments.

- Compared to existing analysis (Choromanska et al. 2016; Dirsken and Stollenwerk 2016; Yu et al. 2015; Oymak 2016), our analysis relies on *realistic* conditions on the number of bits and maximum value of vectors in datasets to achieve the optimal complexity of CBE, matching the optimality of unstructured projection in case that ϵ-distortion binary embedding is interested.

To present our analysis, the following assumptions should be introduced.

Condition 1. *Suppose that we have $\mathcal{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \subset \mathcal{S}^{d-1}$. Letting $\rho \triangleq \sup_{1 \leq i \leq n} ||\boldsymbol{x}_i||_\infty$, there exist nonnegative constants c_1, c_2, c_3, such that*

- $k \geq c_1 \epsilon^{-2} \log n$.
- $c_2 \epsilon k \rho \log d < 1$.
- $c_3 \rho k < \epsilon$.
- $c_4 k^3 \rho^2 \epsilon^2 < 1$,

where k is the number of bits, n is the number of data points, and d is the data dimension. In addition to the assumptions, we implicitly assume that $N > d$, which is a certainly desirable scenario.

Condition 1 summarizes the assumptions necessary to build the main analysis. Note that this condition consists of three assumptions where the first condition matches the optimal bit complexity of BE with unstructured projection, the second/third conditions are similarly introduced as in (Oymak 2016), and the final condition appears because of deriving the optimal complexity. Except for the first one, it is not trivial to see the relation of variables, leading to describe details in the next section.

Before introducing technical arguments, our main result is firstly presented:

Theorem 2. *Given $\epsilon \in (0, 1)$ and any finite dataset $\mathcal{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \subset \mathcal{S}^{d-1}$, under Condition 1, with probability at least $1 - \exp(-c_5 \epsilon^2 k)$, $k = \mathcal{O}\left(\epsilon^{-2} \log n\right)$ implies that CBE guarantees ϵ-distortion binary embedding such that for all $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{D}$*

$$\left| d_H(h^C(\boldsymbol{x}_i), h^C(\boldsymbol{x}_j)) - \frac{\theta_{\boldsymbol{x}_i, \boldsymbol{x}_j}}{\pi} \right| \leq \epsilon,$$

where $c_5 > 0$ is a constant.

Proof. Since the proof contains technical arguments, details are discussed in the subsequent sections. □

Theorem 2 achieves the optimal bit complexity composed of the optimal distortion rate and sample complexity, compared to BE. Up to our best knowledge, our analysis is the first attempt to establish the optimal complexity under mild conditions, even though there exist several empirical evidences (Yu et al. 2014; 2015) to support that CBE similarly performs the standard BE.

Discussion on Condition 1

In this section, we discuss on the complex relation of major factors (ρ, ϵ, k, d) in Condition 1. Our interest is to consider low-distortion binary embedding, i.e. $\epsilon = o(1)$ and the case that a upper bound on l_∞ norm decreases with respect to data dimension, meaning that ρ follows a similar form of $\mathcal{O}(d^{-c})$, where $c > 0$ is a positive constant. Then, there exist various choices of k and n by employing the following geometrical configurations.

As pointed out in (Oymak 2016), the maximum incoherence of dataset suggests that $\rho = \mathcal{O}(d^{-1/2})$. It is easy to see that setting by $k = \mathcal{O}(d^{1/3})$ and $n = \mathcal{O}(\epsilon^2 k)$ satisfies the assumptions in Condition 1. (Ailon and Chazelle 2009) preprocess data vectors by applying Hadamard transformation, having $\rho = \mathcal{O}((\log d) d^{-1/2})$ with high probability. Similarly, the assumptions in Condition 1 are easily satisfied by setting $k = \mathcal{O}\left((\log d)^{-2/3} d^{1/3}\right)$ and $n = \mathcal{O}(\epsilon^2 k)$.

Remark that our setting of k is slightly worse than $\mathcal{O}((\log d)^{-1} d^{1/2})$, which is chosen by (Oymak 2016). The difference appears from introducing the optimal distortion rate, and we argue that such difference is not significant.

Proof of Main Theorem 2

In this section, we introduce the proof of Theorem 2 by employing an orthogonal decomposition of circulant pairs which is studied in (Yu et al. 2015; Oymak 2016). Before discussing details of our analysis, several notations are formally introduced.

Definition 3. *Given an example $\boldsymbol{x}_i$ for any $i = 1, \cdots, n$, the m-shifted variable for $\boldsymbol{x}_i$ is defined as $\boldsymbol{x}_{i,m}^c \triangleq [circ(\boldsymbol{D}\boldsymbol{x})]_{:,m}$, where $[circ(\boldsymbol{D}\boldsymbol{x})]_{:,m}$ is the m-th column vector of the circulant matrix for $\boldsymbol{D}\boldsymbol{x}$ defined by Eq. 5.*

Therefore, for any $l = 1, \cdots, k$, $\mathcal{I}_l^C$ can be re-defined as follows.

$$\mathcal{I}_l^C \triangleq \mathcal{I}[\text{sgn}(\boldsymbol{w}^\top \boldsymbol{x}_{i,l}^c) \neq \text{sgn}(\boldsymbol{w}^\top \boldsymbol{x}_{j,l}^c)], \tag{9}$$

where $\boldsymbol{w}$ is drawn from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$.

The following definition re-states the procedure of orthogonal decomposition for circulant pairs $\left\{\boldsymbol{x}_{i,l}^c, \boldsymbol{x}_{j,l}^c\right\}_{l=1}^m$ as studied in (Oymak 2016; Yu et al. 2015).

Definition 4. *Given $\boldsymbol{x}_i, \boldsymbol{x}_j$ from $\mathcal{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \subset \mathcal{S}^{d-1}$, there exists an orthogonal decomposition for $\boldsymbol{x}_{i,m}^c, \boldsymbol{x}_{j,m}^c$ defined as:*

$$\begin{aligned} \boldsymbol{x}_{i,m}^c &= \widehat{\boldsymbol{x}}_{i,m}^c + \widehat{\boldsymbol{p}}_{i,m} & (10)\\ \boldsymbol{x}_{j,m}^c &= \widehat{\boldsymbol{x}}_{j,m}^c + \widehat{\boldsymbol{p}}_{j,m}, & (11)\end{aligned}$$

where $m = 1, \cdots, d$. If $m = 1$, both $\widehat{\boldsymbol{p}}_i$ and $\widehat{\boldsymbol{p}}_j$ are zero vectors. Otherwise (i.e. if $m > 1$), $\widehat{\boldsymbol{p}}_i, \widehat{\boldsymbol{p}}_j$ are the projections of $\boldsymbol{x}_{i,m}^c, \boldsymbol{x}_{j,m}^c$ onto the subspace spanned by $\left\{\widehat{\boldsymbol{x}}_{i,l}^c, \widehat{\boldsymbol{x}}_{j,l}^c\right\}_{l=1}^{m-1}$.

A crucial observation of Definition 4 is that $||\widehat{\boldsymbol{p}}_{i,m}||_2$ for any i and m is always bounded with high probability, which is shown in the subsequent corollary.

Corollary 1. *Suppose that $\boldsymbol{x}_i, \boldsymbol{x}_j$ from $\mathcal{D} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\} \subset \mathcal{S}^{d-1}$ satisfies $\max\{||\boldsymbol{x}_i||_\infty, ||\boldsymbol{x}_j||_\infty\} \leq \rho$. Letting $\epsilon \in (0, 1)$, with probability at least $1 - 4\exp(-\epsilon^2 k)$, we achieve that*

$$\max\{||\widehat{\boldsymbol{p}}_{i,m}||_2, ||\widehat{\boldsymbol{p}}_{j,m}||_2\} \leq c_a \epsilon k \rho,$$

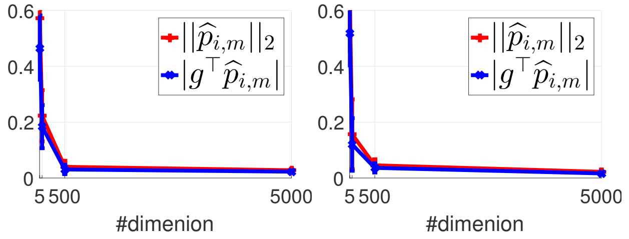
where $c_a > 0$ is a constant. $\{\widehat{\boldsymbol{p}}_{i,m}, \widehat{\boldsymbol{p}}_{j,m}\}_{m=1}^k$ are defined in Definition 4.

Proof. Since the proof is eseentially similar to Lemma 5.1 in (Oymak 2016) by employing Lemma 2 in Appendix, we describe details in the supplementary material. □

Given Condition 1, $c_2 \epsilon k \rho \log d < 1$ induces the following interesting property:

$$\max\{||\widehat{\boldsymbol{p}}_{i,m}||_2, ||\widehat{\boldsymbol{p}}_{j,m}||_2\} \leq c \log^{-1} d, \tag{12}$$

where $c > 0$ is a constant. As d grows, both $||\widehat{\boldsymbol{p}}_{i,m}||_2$ and $||\widehat{\boldsymbol{p}}_{j,m}||_2$ become the zero vectors. The following corollary is a simple consequence of Corollary 1, which means that the absolute value of random projection of $||\widehat{\boldsymbol{p}}_{i,m}||_2$ is bounded with high probability.



(a) $\theta_{\boldsymbol{x}_i, \boldsymbol{x}_j} = \frac{\pi}{2}$ (b) $\theta_{\boldsymbol{x}_i, \boldsymbol{x}_j} = \frac{\pi}{3}$

Figure 2: Plots of $||\widehat{\boldsymbol{p}}_{i,m}||_2$ and the sample average of $|\boldsymbol{g}^\top \widehat{\boldsymbol{p}}_{i,m}|$ for $\theta_{\boldsymbol{x}_i, \boldsymbol{x}_j} = \pi/2$ and $\theta_{\boldsymbol{x}_i, \boldsymbol{x}_j} = \pi/3$ with respect to the data dimension. The sample average is computed by $\frac{1}{n}\sum_{k=1}^n |\boldsymbol{g}_k^\top \widehat{\boldsymbol{p}}_{i,m}|$, where $\boldsymbol{g}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ for $i = 1, \cdots, n$ and $n = 100$. Without loss of generality, we set $m = 2$.

Corollary 2. *Under the same setting of Corollary 1 with Condition 1 including $c_4 k^3 \rho^2 \epsilon^2 \leq 1$, for any $i \in \{1, \cdots, n\}$ and $m \in \{1, \cdots, k\}$, the following holds*

$$\mathbb{P}\left[|\boldsymbol{g}^\top \widehat{\boldsymbol{p}}_{i,m}| \leq \epsilon\right] \geq 1 - \exp(-c_4 \epsilon^2 k),$$

where $\boldsymbol{g}$ follows $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $c_4 > 0$.

Proof. Supposing that $||\widehat{\boldsymbol{p}}_{i,m}||_2 \leq \epsilon$, the standard Gaussian distribution property suggests that

$$\mathbb{P}\left[|\boldsymbol{g}^\top \widehat{\boldsymbol{p}}_{i,m}| \geq t\epsilon\right] \leq \exp(-0.5t^2),$$

where $\boldsymbol{g}$ follows $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Corollary 1 says that $\widehat{\boldsymbol{p}}_{i,m}$ is always bounded by $c_a \epsilon k \rho$ with probability $1 - 4\exp(-\epsilon^2 k)$, which leads to

$$\mathbb{P}\left[|\boldsymbol{g}^\top \widehat{\boldsymbol{p}}_{i,m}| \leq \epsilon\right] \geq 1 - \exp\left(-0.5 c_a^{-2} \rho^{-2} k^{-2}\right).$$

The condition $c_4 k^3 \rho^2 \epsilon^2 \leq 1$ makes the lower bound interesting, i.e.,

$$\begin{aligned} 1 - 4\exp(-\epsilon^2 k) &\geq 1 - \exp(-c_4 \epsilon^2 k), \\ 1 - \exp\left(-0.5 c_a^{-2} \rho^{-2} k^{-2}\right) &\geq 1 - \exp(-c_4 \epsilon^2 k). \end{aligned}$$

Then, given a large n, there exists a non-negative c_4 such that

$$c_4 \leq \min\left\{\frac{1}{2} c_a^{-2} k^{-3} \rho^{-2} \epsilon^{-2}, 1 - \frac{\log 4}{c_1 \log n}\right\},$$

which concludes the proof. □

By setting $k = \mathcal{O}(d^{1/3})$ and $\rho = \mathcal{O}((\log d) d^{-1/2})$, Corollary 2 can be interpreted in terms of the data dimension. It means that as d goes infinity, the absolute value of random projection of $||\widehat{\boldsymbol{p}}_{i,m}||_2$ is bounded with high probability, which is empirically supported in Figure 2.

Lemma 1. *Under the same setting of Corollary 1, for $m = 1, \cdots, k$, with probability at least $1 - 4\exp(-\epsilon^2 k)$, the following holds*

$$\left|ang\left(\widehat{\boldsymbol{x}}_{i,m}^c, \widehat{\boldsymbol{x}}_{j,m}^c\right) - ang\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right| \leq c\epsilon k \rho,$$

where $ang(\boldsymbol{x}, \boldsymbol{y})$ is the angle between two vectors $\boldsymbol{x}, \boldsymbol{y}$ and $c > 0$ is a constant.

Proof. The proof has a similar spirit of the logic in (Oymak 2016). Lemma A.5 in (Oymak 2016) states that

$$\text{ang}\left(\boldsymbol{x}_i, \widehat{\boldsymbol{x}}^c_{i,m}\right) = \text{ang}\left(\boldsymbol{x}_i, \boldsymbol{x}_i - \widehat{\boldsymbol{p}}_{i,m}\right) \leq 5||\widehat{\boldsymbol{p}}_{i,m}||_2.$$

According to Corollary 1, with probability at least $1 - 4\exp(-\epsilon^2 k)$, the following is achieved:

$$\max\left\{\text{ang}\left(\boldsymbol{x}_i, \widehat{\boldsymbol{x}}^c_{i,m}\right), \text{ang}\left(\boldsymbol{x}_j, \widehat{\boldsymbol{x}}^c_{j,m}\right)\right\} \leq 5c_a\epsilon k\rho,$$

where $c_a > 0$ is a constant. Then, the following relation is revealed by applying the triangle inequality of angular distance:

$$\begin{aligned} & \left|\text{ang}\left(\widehat{\boldsymbol{x}}^c_{i,m}, \widehat{\boldsymbol{x}}^c_{j,m}\right) - \text{ang}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)\right| \\ \leq & \left|\text{ang}\left(\widehat{\boldsymbol{x}}^c_{i,m}, \boldsymbol{x}_i\right) + \text{ang}\left(\widehat{\boldsymbol{x}}^c_{j,m}, \boldsymbol{x}_j\right)\right| \leq 10c_a\epsilon k\rho. \end{aligned}$$

By setting $c = 10c_a$, the proof is concluded. □

In the rest of the section, the proof of Theorem 2 is described. We extend the arguments presented in (Oymak 2016) to achieve the optimal bit complexity. Our interest is to build a tight lower bound on the following event:

$$E_{(i,j)} \triangleq \left|\frac{1}{k}\sum_{l=1}^{k} \mathcal{I}^C_l(\boldsymbol{x}_i, \boldsymbol{x}_j) - \text{ang}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right| \leq \epsilon, \quad (13)$$

where $\mathcal{I}^C_l$ is defined as in Eq. 9 and $\epsilon \in (0, 1)$. As observed in Figure 1, $\{\mathcal{I}^C_l\}_{l=1}^k$ are not mutually independent, which means that the event should be decomposed into the following two sub-events:

$$\begin{aligned} E_{(i,j),r} &\triangleq \left|\frac{1}{k}\sum_{l=1}^{k} \mathcal{I}^{r,l}_{(i,j)} - \text{ang}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right| \leq \epsilon & (14) \\ E_{(i,j),p} &\triangleq \cap_k \left[\max\left\{|\boldsymbol{g}^\top \widehat{\boldsymbol{p}}_{i,k}|, |\boldsymbol{g}^\top \widehat{\boldsymbol{p}}_{j,k}|\right\} \leq \epsilon\right], & (15) \end{aligned}$$

where $\mathcal{I}^{r,l}_{(i,j)}$ is the union of two events defined by

$$\begin{aligned} E^1_{margin} &\triangleq \mathcal{I}\left[\boldsymbol{g}^\top \widehat{\boldsymbol{x}}^c_{i,m} > \epsilon \text{ and } \boldsymbol{g}^\top \widehat{\boldsymbol{x}}^c_{j,m} < -\epsilon\right] \\ E^2_{margin} &\triangleq \mathcal{I}\left[\boldsymbol{g}^\top \widehat{\boldsymbol{x}}^c_{i,m} < -\epsilon \text{ and } \boldsymbol{g}^\top \widehat{\boldsymbol{x}}^c_{j,m} > \epsilon\right]. \end{aligned}$$

It is trivial to observe that $E_{(i,j)}$ is satisfied whenever $E_{(i,j),r} \cap E_{(i,j),p}$ holds, which leads to develop a lower bound on the event $E_{(i,j)}$:

$$\begin{aligned} \mathbb{P}\left[E_{(i,j)}\right] &= \mathbb{P}\left[E_{(i,j),r} \cap E_{(i,j),p}\right] & (16) \\ &= 1 - \mathbb{P}\left[E^c_{(i,j),r} \cup E^c_{(i,j),p}\right] & (17) \\ &\geq 1 - \left(\mathbb{P}\left[E^c_{(i,j),r}\right] + \mathbb{P}\left[E^c_{(i,j),p}\right]\right). & (18) \end{aligned}$$

As discussed in (Yu et al. 2015; Oymak 2016), the random variables $\{\boldsymbol{x}^c_{i,m}, \boldsymbol{x}^c_{j,m}\}_{m=1}^k$ are mutually independent, an upper bound on $\mathbb{P}\left[E^c_{(i,j),r}\right]$ can be achieved by applying the standard Hoeffding's inequality, resulting in

$$\mathbb{P}\left[E^c_{(i,j),r}\right] \leq 2\exp(-2\epsilon^2 k), \quad (19)$$

where we place derivation on this inequality in the supplementary material. Now, looking at the second event, the following holds.

$$\begin{aligned} \mathbb{P}\left[E^c_{(i,j),p}\right] &= \mathbb{P}\left[\cup_{m=1}^k \left[\min\left\{|\boldsymbol{g}^\top \widehat{\boldsymbol{p}}_{i,m}|, |\boldsymbol{g}^\top \widehat{\boldsymbol{p}}_{j,m}|\right\} \geq \epsilon\right]\right] \\ &\leq \sum_{m=1}^{k} \mathbb{P}\left[\min\left\{|\boldsymbol{g}^\top \widehat{\boldsymbol{p}}_{i,m}|, |\boldsymbol{g}^\top \widehat{\boldsymbol{p}}_{j,m}|\right\} \geq \epsilon\right] \\ &\leq k\exp(-c_4\epsilon^2 k) = \exp(-c_d\epsilon^2 k), \end{aligned}$$

where the final inequality is derived by Corollary 2 and $0 < c_d \leq c_4 - \frac{\ln k}{c_1 \log n}$ is a constant, which is easy to be satisfied in case of a large n.

Now, a lower bound on the event $E_{(i,j)}$ is achieved by the followings:

$$\begin{aligned} \mathbb{P}\left[E_{(i,j)}\right] &\geq 1 - \left(\mathbb{P}\left[E^c_{(i,j),r}\right] + \mathbb{P}\left[E^c_{(i,j),p}\right]\right) \\ &\geq 1 - \left(2\exp(-2\epsilon^2 k) + \exp(-c_d\epsilon^2 k)\right) \\ &\geq 1 - \left(3\exp(-c_f\epsilon^2 k)\right), \end{aligned}$$

where c_f is set to $\min(2, c_d)$. Obviously, c_f is greater than zero, which makes the bound interesting. The standard procedure with the union bound derives the lower bound on the event $E_{(i,j)}$ for all pairs $1 \leq i, j \leq n$:

$$\begin{aligned} \mathbb{P}\left[\cap_{(i,j)} E_{(i,j)}\right] &\geq 1 - 3n^2\exp(-c_f\epsilon^2 k) \\ &\geq 1 - \exp(-c_5\epsilon^2 k), \end{aligned}$$

where the final inequality holds due to the first condition in Condition 1, i.e., $k \geq c_1\epsilon^{-2}\log n$. It concludes the proof of Theorem 2.

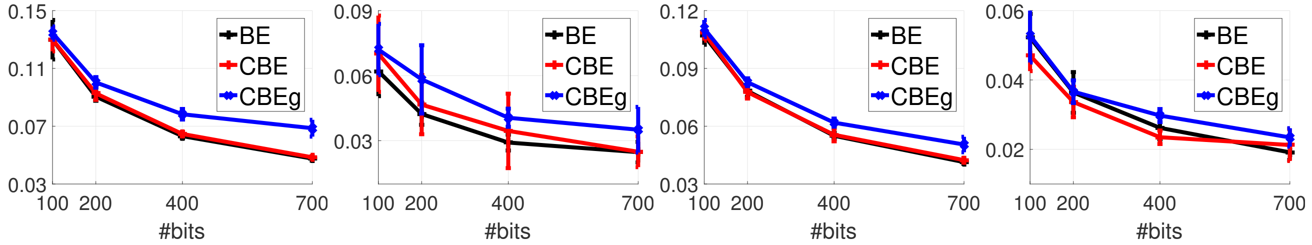
Experiments

In this section, we conducted various numerical experiments to support the theoretical analysis of CBE developed in section with the following datasets:

- MNIST (LeCun et al. 1998) consists of 70,000 handwritten digit images where images are represented by 784-dimensional vectors. We used raw images as high-dimensional vectors.
- CIFAR-10 (Krizhevsky and Hinton 2009) consists of 60,0000 low-resolution images from 10 classes. We trained a residual network (He et al. 2016) with 32 layers and 1,024 filters to achieve around 95% classification accuracy. We extracted 1,024-dimensional features for images, which is obtained by the average pooling layer in residual network.
- GIST1M (Jégou, Douze, and Schmid 2011) consists of 920-dimensional 1 million GIST descriptors with additional 1,000 queries.

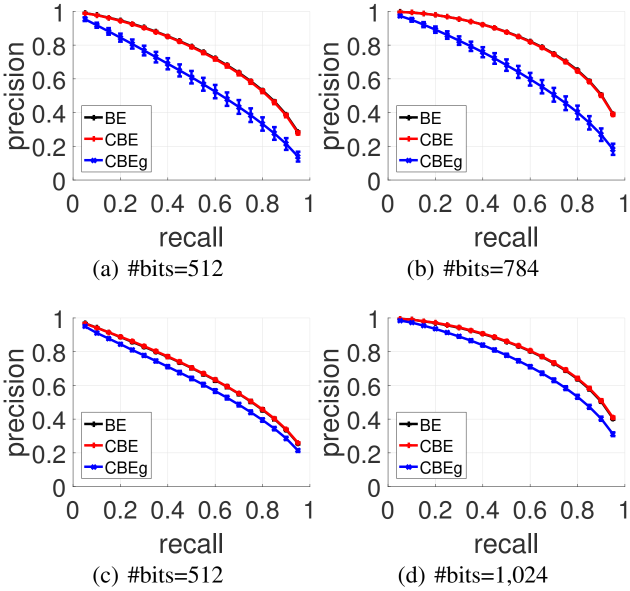
For preprocessing, all vectors in datasets are l_2 normalized. Due to the space limit, we defer the results of GIST1M and another dataset to the supplementary material.

We compared CBE with BE and CBEg in terms of angle preservation and Hamming ranking evaluation, where CBEg means that CBE with Gaussian random sequence instead of Radmecher sequence, proposed in (Oymak 2016). All experiments are repeated five times to avoid any bias.



(a) MNIST ($||\cdot||_F$) (b) MNIST ($||\cdot||_2$) (c) CIFAR-10 ($||\cdot||_F$) (d) CIFAR-10 ($||\cdot||_2$)

Figure 3: Plots for the relative error on approximating the angle between vectors measured by Frobenious and spectral norms, showing that CBE is almost identical to the standard binary embedding.



(a) #bits=512 (b) #bits=784

(c) #bits=512 (d) #bits=1,024

Figure 4: Precision-recall curves from Hamming ranking evaluation to compare CBE with BE and CBEg on MNIST(the top row) and CIFAR-10(the second row).

Angle Preservation

To evaluate the angle preservation, we measured the relative error on approximating the angle with the normalized Hamming distance:

$$\frac{||\boldsymbol{H}-\boldsymbol{A}||_F}{||\boldsymbol{A}||_F}, \quad \frac{||\boldsymbol{H}-\boldsymbol{A}||_2}{||\boldsymbol{A}||_2},$$

where $||\cdot||_F$ is Frobenius norm, $||\cdot||_2$ is spectral norm, $\boldsymbol{H}_{ij} = d_H(h(\boldsymbol{x}_i), h(\boldsymbol{x}_j))$ and $\boldsymbol{A}_{ij} = \frac{\theta_{\boldsymbol{x}_i,\boldsymbol{x}_j}}{\pi}$. Similar metrics have been introduced to measure the relative error on Gram matrix (Yang et al. 2014). To evaluate $\boldsymbol{H}$ and $\boldsymbol{A}$, we randomly select 3,000 data points for all datasets. Figure 3 concludes that BE and CBE are indistinguishable in terms of angle preservation. Moreover, we observed that a Gaussian random sequence proposed in (Oymak 2016) is inferior to Rademacher sequence, showing that our theoretical analysis is well suited for CBE.

Hamming Ranking Evaluation

We followed Hamming ranking evaluation to compare the methods in terms of approximate nearest neighbor accuracy, where the close neighbors for queries are retrieved by calculating the normalized Hamming distance between queries and data points. For all datasets, we randomly select 1,000 queries from test datasets and computed 100 nearest neighbors for ground-truths, where angular distance is used. For evaluation measure, we computed the standard precision-recall curves with respect to the different number of bits.

Figure 4 compares the performance of BE and CBE in terms of precision-recall curves, concluding that the performance of BE and CBE is almost identical. These results are also very similar to Figure 3 and are well supported by the analysis in the main section.

Conclusions

Binary embedding (BE) projects the data points in $\mathbb{R}^d$ into $\{0,1\}^k$ such that the normalized Hamming distance should preserve the pre-defined similarity metric. Despite the simplicity of binary embedding, it requires the large time and space complexities, $\mathcal{O}(d^2)$, to precisely estimate the similarity, where the data dimension is denoted by d. A promising approach to reduce the time and space complexities is circulant binary embedding (CBE), which is empirically validated that CBE shows comparable performance. In this paper, we established a condition on the number of bits required for CBE to preserve the angular distance up to ϵ-distortion, $k = \mathcal{O}\left(\epsilon^{-2}\log n\right)$, which is the optimal bit complexity compared to BE with unstructured projection.

Appendix

Lemma 2. *Letting two unit vectors be $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^{d-1}$, suppose that $\max\{||\boldsymbol{x}||_\infty, ||\boldsymbol{y}||_\infty\} \leq \rho$ and $\boldsymbol{x}^\top \boldsymbol{y} = 0$. Given $\boldsymbol{D}$ be a diagonal matrix whose entries are i.i.d. Rademacher entries, for all $1 \leq i \neq j \leq d$, the following is achieved:*

$$\begin{aligned} \mathbb{P}\left(|circ(\boldsymbol{D}\boldsymbol{x})_{:,i}^\top circ(\boldsymbol{D}\boldsymbol{y})_{:,j}| > t\right) &\leq 2\exp\left(-ct^2\rho^{-2}\right) \\ \mathbb{P}\left(|circ(\boldsymbol{D}\boldsymbol{x})_{:,i}^\top circ(\boldsymbol{D}\boldsymbol{x})_{:,j}| > t\right) &\leq 2\exp\left(-ct^2\rho^{-2}\right), \end{aligned}$$

where $t \in (0,1)$, $c > 0$, and $circ(\cdot)$ is defined in Eq. 5.

Proof. We defer the proof to the supplementary material. □

Acknowledgements

A portion of this work was supported by NAVER. S. Kim was partially supported by NAVER Ph.D. fellowship award (2016).

References

Ailon, N., and Chazelle, B. 2009. The fast Johnson-Lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on Computing* 39(1):302–322.

Charikar, M. S. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*.

Choromanska, A.; Choromanski, K.; Bojarski, M.; Jebara, T.; Kumar, S.; and LeCun, Y. 2016. Binary embeddings with structured hashed projections. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Dasgupta, A.; Kumar, R.; and Sarlós, T. 2011. Fast locality-sensitive hashing. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.

Dirsken, S., and Stollenwerk, A. 2016. Fast binary embeddins with gaussian circulant matrices: Improved bounds. arXiv preprint arXiv:1608.06498v1.

Eftekhari, A.; Babaie-Zadeh, M.; and Moghaddam, H. A. 2011. Two-dimensional random projection. *Signal Processing* 91:1589–1603.

Gong, Y.; Kumar, S.; Rowley, H. A.; and Lazebnik, S. 2013. Learning binary codes for high-dimensional data using bilinear projections. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hinrichs, A., and Vybíral, J. 2011. Johnson-Lindenstrauss lemma for circulant matrices. *Random Structures & Algorithms* 39(3):391–398.

Jacques, L.; Laska, J. N.; Boufounos, P. T.; and Baraniuk, R. G. 2013. Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. *IEEE Transactions on Information Theory* 59(4):2082–2101.

Jégou, H.; Douze, M.; and Schmid, C. 2011. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.

Johnson, W. B., and Lindenstrauss, J. 1984. Extensions of Lipschitz mappings into a Hilbert space. *Contemporay Mathematics* 26:189–206.

Kim, S., and Choi, S. 2015. Bilinear random projections for locality-sensitive binary embedding. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Krizhevsky, A., and Hinton, G. E. 2009. Learning multiple layers of features from tiny images. Technical report, Computer Science Department, University of Toronto.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Oymak, S. 2016. Near-optimal sample complexity bounds for circulant binary embedding. arXiv preprint arXiv:1603.03178v2.

Raginsky, M., and Lazebnik, S. 2009. Locality-sensitive binary codes from shift-invariant kernels. In *Advances in Neural Information Processing Systems (NIPS)*, volume 22. MIT Press.

Shrivastava, A., and Li, P. 2014. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In *Advances in Neural Information Processing Systems (NIPS)*, volume 27.

Yang, J.; Sindhwani, V.; Avron, H.; and Mahoney, M. W. 2014. Quasi-monte carlo feature maps for shift-invariant kernels. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Yi, X.; Caramanis, C.; and Price, E. 2015. Binary embedding: Fundamental limits and fast algorithm. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Yu, F. X.; Kumar, S.; Gong, Y.; and Chang, S.-F. 2014. Circulant binary embedding. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Yu, F. X.; Bhaskara, A.; Kumar, S.; Gong, Y.; and Chang, S.-F. 2015. On binary embedding using circulant matrices. arXiv preprint arXiv:1511.06480v2.

Zhang, X.; Yu, F. X.; Guo, R.; Kumar, S.; Wang, S.; and Chang, S.-F. 2015. Fast orthogonal projection based on kronecker product. In *Proceedings of the International Conference on Computer Vision (ICCV)*.