On the Optimal Bit Complexity of Circulant Binary Embedding (supplementary material)

Saehoon Kim[1,2], Jungtaek Kim[1], and Seungjin Choi[1]
[1]CSE, POSTECH, Pohang, Korea [2]AItrics, Seoul, Korea
{kshkawa,jtkim,seungjin}@postech.ac.kr

1 Outline

In this supplementary material, we present the proofs of Theorem 1, Corollary 1, and other lemmas in the submitted manuscript. Moreover, we introduce some experimental results for GIST1M and Flickr45K, which are not placed in the manuscript due to the space limit.

2 Proof of Theorem 1

Theorem 1. *(Theorem 1 in the manuscript). 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 $1-\exp(-c\epsilon^2 k)$, $k = \mathcal{O}\left(\epsilon^{-2}\log n\right)$ implies that we have $h : \mathcal{S}^{d-1} \to \{0,1\}^k$ 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. It can be easily proved by Hoeffding's inequality [1] and union bound. Hoeffding's inequality provides the upper-bound on the probability that the sum of independent random variables deviates from its expected value. When entries of $\boldsymbol{G}$ are drawn independently from Gaussian with zero mean and unit variance, Bernoulli indicator variables $\{\mathcal{I}_l(\boldsymbol{x}_i, \boldsymbol{x}_j)\}_{l=1}^k$ are marginally independent, given $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Thus, the probability that the normalized Hamming distance (the sum of $\mathcal{I}_l$) deviates from the angle distance (its mean), is given by

$$\mathbb{P}\left[\left| d_H\left(h(\boldsymbol{x}_i), h(\boldsymbol{x}_j)\right) - \frac{\theta_{\boldsymbol{x}_i,\boldsymbol{x}_j}}{\pi} \right| \geq \epsilon\right] \leq 2\exp\left\{-2k\epsilon^2\right\}. \tag{1}$$

Given a dataset $\mathcal{D}$, a ϵ-distortion binary embedding involves the following probability:

$$\mathbb{P}\left[\bigcap_{\boldsymbol{x}_i,\boldsymbol{x}_j \in \mathcal{D}} \left(\left| d_H\left(h(\boldsymbol{x}_i), h(\boldsymbol{x}_j)\right) - \frac{\theta_{\boldsymbol{x}_i,\boldsymbol{x}_j}}{\pi} \right| \leq \epsilon \right)\right]. \tag{2}$$

Strict positiveness of this probability guarantees ϵ-distortion binary embedding. To this end, we use Boole's inequality (a.k.a. the union bound) which states that for any finite or countable set of events, the probability that at least one of the events happens is no greater than the sum of the probabilities of the individual events, i.e., $\mathbb{P}\left[\cup_i A_i\right] \leq \sum_i \mathbb{P}[A_i]$, for a set of events $\{A_i\}$. Applying Boole's inequality

yields

$$\mathbb{P}\left[\bigcap_{\boldsymbol{x}_i,\boldsymbol{x}_j\in\mathcal{D}}\left(\left|d_H\left(h(\boldsymbol{x}_i),h(\boldsymbol{x}_j)\right)-\frac{\theta_{\boldsymbol{x}_i,\boldsymbol{x}_j}}{\pi}\right|\leq\epsilon\right)\right] \tag{3}$$

$$=1-\mathbb{P}\left[\bigcup_{\boldsymbol{x}_i,\boldsymbol{x}_j\in\mathcal{D}}\left(\left|d_H\left(h(\boldsymbol{x}_i),h(\boldsymbol{x}_j)\right)-\frac{\theta_{\boldsymbol{x}_i,\boldsymbol{x}_j}}{\pi}\right|\geq\epsilon\right)\right] \tag{4}$$

$$\geq 1-\sum_{\boldsymbol{x}_i,\boldsymbol{x}_j\in\mathcal{D}}\mathbb{P}\left[\left|d_H\left(h(\boldsymbol{x}_i),h(\boldsymbol{x}_j)\right)-\frac{\theta_{\boldsymbol{x}_i,\boldsymbol{x}_j}}{\pi}\right|\geq\epsilon\right] \tag{5}$$

$$\geq 1-2n^2\exp\left\{-2k\epsilon^2\right\}. \tag{6}$$

The lower bound on the probability of event becomes interesting (i.e. greater than zero) when the condition $k\geq\mathcal{O}(\epsilon^{-2}\log n)$ is given. □

3 Proof of Corollary 1

In this section, some details for the proof of Corollary 1 in the manuscript are introduced. Lemma 1 1 is the key for the proof, which needs Hanson-Wright inequality for sub-Gaussian random variables described as follows.

Theorem 2. *(Hanson-Wright inequality for sub-Gaussian random variables [5]). Let $\boldsymbol{X}=(X_1,\cdots,X_d)\in\mathbb{R}^d$ be a random vector with independent sub-Gaussian random variables X_i which satisfy $\mathbb{E}[X_i]=0$ and $||X_i||_{\psi_2}\leq K$. Let $\boldsymbol{A}$ be an $d\times d$ matrix. Then, for every $t\geq 0$,*

$$\mathbb{P}\left(\left|\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X}-\mathbb{E}\left[\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{X}\right]\right|>t\right)\leq 2\exp\left[-c\min\left(\frac{t^2}{K^4||\boldsymbol{A}||_F^2},\frac{t}{K^2||\boldsymbol{A}||_2}\right)\right]. \tag{7}$$

Lemma 1. *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 i.i.d. Rademacher entries and $\boldsymbol{D}=diag(\boldsymbol{d})$, for all $1\leq i\neq j\leq d$, the following is achieved.*

$$\mathbb{P}\left(|circ(\boldsymbol{D}\boldsymbol{x})_{:,i}^\top circ(\boldsymbol{D}\boldsymbol{y})_{:,j}|>t\right) \leq 2\exp\left(-ct^2\rho^{-2}\right) \tag{8}$$

$$\mathbb{P}\left(|circ(\boldsymbol{D}\boldsymbol{x})_{:,i}^\top circ(\boldsymbol{D}\boldsymbol{x})_{:,j}|>t\right) \leq 2\exp\left(-ct^2\rho^{-2}\right), \tag{9}$$

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

Proof. The proof is basically same with the proof of Lemma 4.5 in [4]. It introduces that $circ(\boldsymbol{D}\boldsymbol{x})_{:,i}^\top circ(\boldsymbol{D}\boldsymbol{y})_{:,j}$ can be represented by $\boldsymbol{d}^\top\boldsymbol{A}\boldsymbol{d}$ where $(m+i,m+j)$th entry of $\boldsymbol{A}$ is $x_{m+i}y_{m+j}$ for $1\leq m\leq d$ and remaining entries are 0. The matrix $\boldsymbol{A}$ has the interesting properties: $||\boldsymbol{A}||_F=\rho, ||\boldsymbol{A}||_2=\rho^2$, and $\mathrm{tr}(\boldsymbol{A})=0$. Consequently, it is easy to show that

$$\mathbb{E}\left[\boldsymbol{d}^\top\boldsymbol{A}\boldsymbol{d}\right] = \mathbb{E}\left[\mathrm{tr}\left(\boldsymbol{d}\boldsymbol{d}^\top\boldsymbol{A}\right)\right] \tag{10}$$

$$= \mathrm{tr}\left(\mathbb{E}\left[\boldsymbol{d}\boldsymbol{d}^\top\right]\boldsymbol{A}\right) \tag{11}$$

$$= \mathrm{tr}\left(\boldsymbol{I}\boldsymbol{A}\right)=\mathrm{tr}\left(\boldsymbol{A}\right)=0. \tag{12}$$

Then, Theorem 2 directly says that

$$\mathbb{P}\left(|\mathrm{circ}(\boldsymbol{D}\boldsymbol{x})_{:,i}^\top\mathrm{circ}(\boldsymbol{D}\boldsymbol{y})_{:,j}|>t\right) \leq 2\exp\left[-c\min\left(\frac{t^2}{\rho^2},\frac{t}{\rho^2}\right)\right] \tag{13}$$

$$= 2\exp\left[-ct^2\rho^{-2}\right], \tag{14}$$

where the last equality is derived by $t\in(0,1)$ and $K=1$ for Bernoulli random variables. The same argument for Eq. 9 is developed to conclude the proof. □

Corollary 1. *(Corollary 1 in the manuscript). 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 in the manuscript.

Proof. The proof is concluded by switching Lemma 4.5 in the proof of Lemma 5.1 in [4] with Lemma 1, where some of the steps are twisted. □

4 Useful Properties

In this section, we present some details to understand the proof of Corollary 2 and Theorem 2 in the manuscript. Lemma 2 is introduced to describe the proof of Corollary 2 in the manuscript. Lemma 3 is described for the proof of Theorem 2 (Eq. (19)) in the manuscript.

Lemma 2. *Given $||\boldsymbol{x}||_2 \leq \delta$, the following holds*

$$\mathbb{P}\left[\left|\boldsymbol{g}^\top \boldsymbol{x}\right| \geq t\delta\right] \leq 2\exp(-0.5t^2), \tag{15}$$

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

Proof. It is easy to observe that $\boldsymbol{g}^\top \boldsymbol{x}$ follows $\mathcal{N}(0, ||\boldsymbol{x}||_2^2)$. Then,

$$\begin{aligned} \mathbb{P}\left[\frac{\boldsymbol{g}^\top \boldsymbol{x}}{||\boldsymbol{x}||_2} \geq t\frac{\delta}{||\boldsymbol{x}||_2}\right] &\leq \mathbb{P}\left[\frac{\boldsymbol{g}^\top \boldsymbol{x}}{||\boldsymbol{x}||_2} \geq t\right] &(16)\\ &\leq \exp(-0.5t^2), &(17)\end{aligned}$$

where $\frac{\boldsymbol{g}^\top \boldsymbol{x}}{||\boldsymbol{x}||_2}$ follows $\mathcal{N}(0,1)$ and the standard Chernoff inequality is applied. Since the same argument is applicable for $\mathbb{P}\left[\boldsymbol{g}^\top \boldsymbol{x} \geq t\delta\right]$, it concludes the proof. □

Lemma 3. *As in the manuscript, suppose that $E_{(i,j),r}$ is defined as*

$$E_{(i,j),r} \triangleq \left|\frac{1}{k}\sum_{i=1}^{k} \mathcal{I}_{(i,j)}^{r,l} - ang(\boldsymbol{x}_i, \boldsymbol{x}_j)\right| \leq \epsilon, \tag{18}$$

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

$$\begin{aligned} E_{margin}^1 &\triangleq \mathcal{I}\left[\boldsymbol{g}^\top \widehat{\boldsymbol{x}}_{i,m} > \epsilon \text{ and } \boldsymbol{g}^\top \widehat{\boldsymbol{x}}_{j,m} < -\epsilon\right] &(19)\\ E_{margin}^2 &\triangleq \mathcal{I}\left[\boldsymbol{g}^\top \widehat{\boldsymbol{x}}_{i,m} < -\epsilon \text{ and } \boldsymbol{g}^\top \widehat{\boldsymbol{x}}_{j,m} > \epsilon\right]. &(20)\end{aligned}$$

Given Condition 1 in the manuscript, the following holds

$$\mathbb{P}\left[E_{(i,j),r}^c\right] \leq 2\exp(-2\epsilon^2 k), \tag{21}$$

where $\epsilon \in (0,1)$ and $k > 0$ is the binary code length.

Proof. According Corollary 2 in the manuscript and the third condition (i.e. $c_3 \rho k < \epsilon$) in Condition 1, with at least probability $1 - 4\exp(-\epsilon^2 k)$, we have that $\max\{||\widehat{\boldsymbol{p}}_{i,m}||_2, \widehat{\boldsymbol{p}}_{j,m}||_2\} < \epsilon$. Then, it is easy to see that $\boldsymbol{g}^\top \widehat{\boldsymbol{x}}_{i,m}$ and $\boldsymbol{g}^\top \widehat{\boldsymbol{x}}_{j,m}$ follow a uni-variate Gaussian distribution with zero mean and the variance within $(1-\epsilon^2, 1]$. It is trivial to show the following:

$$\mathbb{P}\left[\mathcal{I}_{(i,j)}^{r,l}\right] \leq \text{ang}(\widehat{\boldsymbol{x}}_{i,m}, \widehat{\boldsymbol{x}}_{j,m}) \leq \text{ang}(\boldsymbol{x}_i, \boldsymbol{x}_j) + c_a \epsilon k \rho \leq \text{ang}(\boldsymbol{x}_i, \boldsymbol{x}_j) + \epsilon^2, \tag{22}$$

where the second inequality is derived by Lemma 1 in the manuscript. As discussed in [6], $\{\mathcal{I}^{r,l}_{(i,j)}\}_{l=1}^{k}$ are marginally independent. Hoeffding's inequality leads that

$$\mathbb{P}\left(\frac{1}{k}\sum_{i=1}^{k}\mathcal{I}^{r,l}_{(i,j)} - \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\mathcal{I}^{r,l}_{(i,j)}\right] \geq \epsilon\right) \leq \exp\left(-2\epsilon^2 k\right) \tag{23}$$

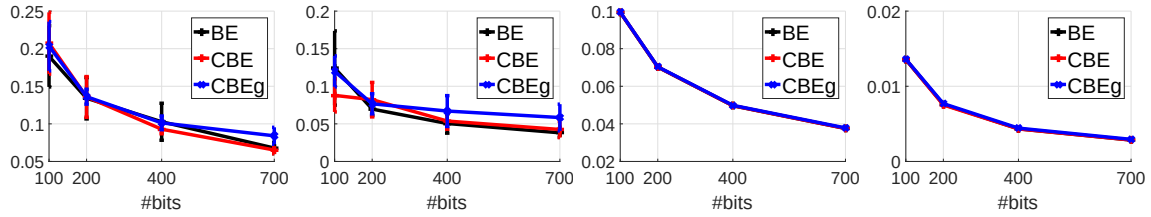
The lower bound on the left-hand side is derived as follows.

$$\begin{aligned}\mathbb{P}\left(\frac{1}{k}\sum_{i=1}^{k}\mathcal{I}^{r,l}_{(i,j)} - \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k}\mathcal{I}^{r,l}_{(i,j)}\right] \geq \epsilon\right) &= \mathbb{P}\left(\frac{1}{k}\sum_{i=1}^{k}\mathcal{I}^{r,l}_{(i,j)} \geq \epsilon + \frac{1}{k}\sum_{i=1}^{k}\mathbb{P}\left[\mathcal{I}^{r,l}_{(i,j)}\right]\right) &(24)\\ &\geq \mathbb{P}\left(\frac{1}{k}\sum_{i=1}^{k}\mathcal{I}^{r,l}_{(i,j)} \geq \epsilon(1+\epsilon) + \text{ang}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right), &(25)\end{aligned}$$

where the last inequality is introduced by Eq. 22 and $\epsilon(1+\epsilon) \approx \epsilon$ because $\epsilon \ll 1$. It concludes that

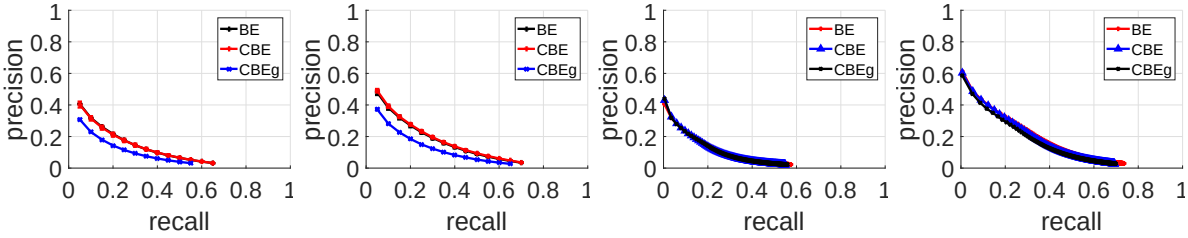
$$\mathbb{P}\left(\frac{1}{k}\sum_{i=1}^{k}\mathcal{I}^{r,l}_{(i,j)} - \text{ang}(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq \epsilon\right) \leq \exp\left(-2\epsilon^2 k\right). \tag{26}$$

A symmetric argument is easily developed to conclude the proof. □



(a) GIST1M ($||\cdot||_F$) (b) GIST1M ($||\cdot||_2$) (c) Flickr45K ($||\cdot||_F$) (d) Flickr45K ($||\cdot||_2$)

Figure 1: Plots for the relative error on approximating the angle between vectors measured by Frobenius and spectral norms for GIST1M and Flickr45K datasets, showing that CBE is almost identical to the standard binary embedding.



(a) #bits=784 (b) #bits=960 (c) #bits=1,024 (d) #bits=2,048

Figure 2: Precision-recall curves from Hamming ranking evaluation to compare CBE with BE and CBEg on GIST1M(the first two subfigures) and Flickr45K(the last two subfigures).

5 Experiments

In this section, we introduce experimental results for angle preservation and hamming ranking evaluation on GIST1M and Flickr45K datasets. Flickr45K consists of 45,000 images randomly chosen from 1 million

Flickr images [2]. VLAD [3] makes use of 500 cluster centers to represent images which are $200 \times 128 = 25,600$-dimensional vectors. To evaluate the angle preservation, we measured the relative error on approximating the angle with the normalized Hamming distance as in the manuscript. To perform Hamming ranking evaluation, we randomly selected 1,000 queries from test datasets and computed 100 nearest neighbors for ground-truths, where angular distance is used.

Clearly, Figure 1 and 2 suggest that CBE preserves the angular distance as well as BE, which has been observed in the manuscript. Unlike MNIST, CIFAR-10, and GIST1M, the performance of CBEg is indistinguishable from the ones of CBE and BE for Flickr45K, which means that the distortion induced by Gaussian random sequences is not serious in case of a large data dimension.

References

[1] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[2] M. J. Huiskes and M. S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.

[3] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, 2010.

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

[5] M. Rudelson and R. Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electron. Commun. Probab*, 18, 2013.

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