NEAR-OPTIMAL SAMPLE COMPLEXITY BOUNDS FOR CIRCULANT BINARY EMBEDDING

Samet Oymak†, Christos Thrampoulidis‡, Babak Hassibi‡

† Google Inc.
‡ California Institute of Technology

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 in $\mathbb{R}^n$ 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 the proposed 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 $\boldsymbol{A}\boldsymbol{x} \to \text{sgn}(\boldsymbol{A}\boldsymbol{x})$ where $\text{sgn}(\cdot)$ returns the vector of signs. 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

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

Here, $\|\cdot,\cdot\|_H$ is the Hamming distance between two $\{0,1\}^k$ vectors and $\text{ang}(\cdot)$ is the angular distance which returns the smaller 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 $k \approx \delta^{-2} \log N$. Embedding continuous sets is a more challenging problem and it is studied in a series of papers [1–5] with results mostly restricted to Gaussian ensemble. These are of interest for sparse estimation and subspace learning 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). Fast transforms reduces embedding time to $\mathcal{O}(n \log n)$ from $\mathcal{O}(kn)$, which is significantly more efficient in the regime $k = \mathcal{O}(\text{poly}(n))$. The theoretical results for fast binary embedding techniques are rather limited [2, 6, 7]. Related to this work, 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. [8]. Here, the authors speed up Kernel approximation [9] by making use of FJLT however the number of required Fourier features scale quadratically due to suboptimal concentration bounds. There are also several works on the applications of fast binary projections in large scale image retrieval and hashing algorithms [10–12].

Contributions: A natural question is whether fast 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

Emails: sametoymak@gmail.com, {cthrampo, hassibi}@caltech.edu

Fast JL Transform not only works well for linear embedding but also for highly nonlinear problems and the embedding behavior is essentially same. 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 $\boldsymbol{x} \to \boldsymbol{H}\text{diag}(\boldsymbol{b})\boldsymbol{x}$ where $\boldsymbol{H}$ is the Hadamard transform and $\text{diag}(\boldsymbol{b})$ is a diagonal matrix with independent Rademacher diagonal entries. The overall embedding takes the form $\boldsymbol{v} \to \text{sgn}(\boldsymbol{A}\boldsymbol{H}\text{diag}(\boldsymbol{b})\boldsymbol{v})$ where $\boldsymbol{A}$ is the binary embedding matrix. 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.

2 Main results

To achieve optimal binary embedding guarantees, we rely on circulant embedding matrices where the projection matrix is given by $\boldsymbol{A} = \boldsymbol{R}\boldsymbol{C}_{\boldsymbol{h}}\text{diag}(\boldsymbol{r})$. Here,

- $\boldsymbol{R} \in \mathbb{R}^{k\times n}$ 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.
- $\boldsymbol{C}_{\boldsymbol{h}}$ is a circulant matrix whose first row is equal to $\boldsymbol{h}^*$.
- $\text{diag}(\boldsymbol{r})$ is the diagonal matrix obtained from the vector $\boldsymbol{r}$.

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

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

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

$$\rho = \max\{\sup_{1\le i\le N} \|\boldsymbol{v}_i\|_{\ell_\infty}, \sup_{1\le i\ne j\le N} \frac{\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_{\ell_\infty}}{\|\boldsymbol{v}_i - \boldsymbol{v}_j\|_{\ell_2}}\}.$$

Coherence naturally lies between $1/\sqrt{n}$ and 1. 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. $c_2\delta k\rho \log n < 1$.
3. $\delta \ge c_3 k\rho$.

Observe that in the maximally incoherent case ($\rho = \mathcal{O}(n^{-1/2})$), 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*

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

This result applies to arbitrary set of points; however, it depends on the incoherence of the set ρ via Condition 2.1. 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

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

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

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*

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

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 $\rho = C\delta(\log n)^{-1/2}n^{-1/3}$. Consequently, as soon as $k \ge c_1\delta^{-3}\log N$, with probability $1 - \exp(-c_4\delta^3 k)$,

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

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) with high probability. In particular, combine Theorem 2.2 with Lemma B.2 of [13][2]. ■

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

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

[2] Additional lemmas and full proofs can be found in the extended manuscript [13]

Theorem 2.4 *Consider the binary embedding via the operator $\boldsymbol{x} \to sgn(\boldsymbol{A_H x})$. 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

- $|S_{good}| \ge (1 - c_5 n^{-2})N$ *and*
- *for all $\boldsymbol{v} \in S_{good}$: $\|\boldsymbol{H} diag(\boldsymbol{b})\boldsymbol{v}\|_{\ell_\infty} \le \rho$ where $\rho = C\sqrt{\log n / n}$.*

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

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

Proof This result follows from the fact that all but a small fraction of the set of points obtained by the map $\boldsymbol{v}_i \to \boldsymbol{H b v}_i$ has desirable geometric features (small ρ) with high probability. In particular, combine Theorem 2.2 with Lemma B.3 of [13]. Pick $p = n^{-2}$ in Lemma B.3. ∎

3 Proof strategy

As mentioned in the introduction, Gaussian projections have superior embedding performance and their properties are rather well understood. The challenge with proposed fast projection method is the fact that $\boldsymbol{Ax}$ do not have statistically independent entries. This makes it difficult to rely on basic tools available for i.i.d. random variables such as Chernoff bound. On the other hand, observe that by construction, the individual rows of $\boldsymbol{A}$ have i.i.d. Gaussian entries. Furthermore, a standard application of Hanson-Wright Theorem [14] can show that if $\boldsymbol{x}$ is a diffused vector (i.e. small $\|\boldsymbol{x}\|_{\ell_\infty}$), entries of $\boldsymbol{Ax}$ have very low pairwise correlations. These two properties imply that $\boldsymbol{A}$ behaves similar to a Gaussian map up to certain extent. Our proof strategy focuses on rigorously characterizing this similarity and carefully using this characterization to obtain optimal embedding bounds.

To illustrate our strategy, we will work on two unit vectors $\boldsymbol{x}, \boldsymbol{y}$. Denote $\mathrm{diag}(\boldsymbol{r})\boldsymbol{x}$ by $\boldsymbol{x^r}$. With this definition, observe that

$$\boldsymbol{Ax} = \boldsymbol{RC_h} \mathrm{diag}(\boldsymbol{r})\boldsymbol{x} = \boldsymbol{RC}^L_{\boldsymbol{x^r}} \boldsymbol{h}.$$

where $\boldsymbol{C}^L_{\boldsymbol{x^r}}$ is the circulant matrix where each new row is rotated to left rather than right and first row is equal to $\boldsymbol{x^{r*}}$. The right hand side is reduced to a form which involves multiplication of a matrix and an i.i.d. Gaussian vector. Recall that multiplication of a unitary matrix and a standard normal vector is still standard normal. Hence, if $\boldsymbol{C}^L_{\boldsymbol{x^r}}$ is approximately a unitary matrix, this would mean that entries of $\boldsymbol{Ax}$ are approximately i.i.d. Gaussian. To put these in a more manageable form, we introduce the following. Given $\boldsymbol{x}$, let $s_i(\boldsymbol{x})$ be the vector obtained by circularly rotating $\boldsymbol{x}$, namely $s_i(\boldsymbol{x})_j = \boldsymbol{x}_{j-i \ (mod\ n)}$. Let the restriction $\boldsymbol{R}$ pick up the rows $\{h_i\}_{i=1}^k$ from $\{1, 2, \ldots, n\}$.

Definition 3.1 (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 $\{\boldsymbol{X}_i\}_{i=1}^k$ such that $\boldsymbol{X}_i = s_{h_i}(diag(\boldsymbol{r})\boldsymbol{x})$ for $1 \le i \le k$. Also let $\boldsymbol{X} = [\boldsymbol{X}_1\ \boldsymbol{X}_2\ \ldots\ \boldsymbol{X}_k]$. Define $\boldsymbol{Y}_i, \boldsymbol{Y}$ in the identical manner given vector $\boldsymbol{y}$ for the same choice of $\boldsymbol{r}$.*

With this definition, we have that

$$\boldsymbol{Ax} = \boldsymbol{RC_h} \mathrm{diag}(\boldsymbol{r})\boldsymbol{x} = \boldsymbol{X}^* \boldsymbol{h}.$$

$\boldsymbol{X}$ is a randomly subsampled circulant matrix and the vector $\boldsymbol{x^r}$ that generates $\boldsymbol{X}$ is random as well. For our proof to work, it is of interest to understand the properties of the random matrix $\boldsymbol{X}$. In a similar fashion to Gram-Schmitt orthogonalization, let $\boldsymbol{X}'_0 = \boldsymbol{X}_0$ and for $i > 0$, write $\boldsymbol{X}_i = \boldsymbol{X}'_i + \boldsymbol{P}_{\boldsymbol{X},i}$ where $\boldsymbol{P}_{\boldsymbol{X},i} \subset \mathrm{span}(\{\boldsymbol{X}'_j\}_{j<i})$ and $\{\boldsymbol{X}'_j\}_{j \le i}$ are orthogonal. Define $\boldsymbol{Y}'_i, \boldsymbol{P}_{\boldsymbol{Y},i}$ similarly.

Writing $\boldsymbol{X} = \boldsymbol{X}' + \boldsymbol{P}_{\boldsymbol{X}}$, we have that $\boldsymbol{X}'$ has orthogonal columns and hence $\boldsymbol{X'^* h}$ has independent Gaussian entries which is trivial to analyze. Here, $\boldsymbol{P}_{\boldsymbol{X}}$ is the matrix of perturbation error and smaller $\boldsymbol{P}_{\boldsymbol{X}}$ shall mean entries of $\boldsymbol{X}^* \boldsymbol{h}$ is closer to being independent. The following lemma characterizes the size of the perturbation $\boldsymbol{P}_{\boldsymbol{X}}$ and helps us specify a notion of approximate independence.

Lemma 3.2 *Let S_i be the subspace spanned by $\{\boldsymbol{X}_{r_j}\}_{j=1}^{i-1}$. With probability $1 - 4\exp(-\delta^2 k)$, we have that for all $1 \le i \le k$*

$$\max\{\|\mathcal{P}_{S_i}(\boldsymbol{X}_i)\|_{\ell_2} = \|\boldsymbol{P}_{\boldsymbol{X},i}\|_{\ell_2}\} \le c_1 \delta k \rho.$$

While Lemma 3.2 upper bounds individual columns, to obtain sharper results, we need to understand the matrix $\boldsymbol{X}$ as a whole. Our main technical result provides an understanding of the conditioning of $\boldsymbol{X}$ and is summarized as follows. The proof heavily relies on earlier results of Tropp on randomly subsampled subdictionaries [15]. Please see the extended manuscript for the detailed technical arguments [13].

Theorem 3.3 (Random circulant subdictionaries) *Pick unit vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ satisfying $\boldsymbol{x}^* \boldsymbol{y} = 0$ and $\|\boldsymbol{x}\|_{\ell_\infty}, \|\boldsymbol{y}\|_{\ell_\infty} \le \rho$. Form a matrix $\boldsymbol{M} \in \mathbb{R}^{n \times 2k}$ by constructing $\boldsymbol{X}, \boldsymbol{Y}$ as described above and setting $\boldsymbol{M} = [\boldsymbol{X}\ \boldsymbol{Y}]$. With probability $1 - 2\exp(-\delta^2 k)$ (over the generation of $\boldsymbol{r}$ and $\{h_i\}_{i=1}^k$'s), we have that*

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

As a side, this result implies that both $\boldsymbol{X}^* \boldsymbol{X}$ and $\boldsymbol{Y}^* \boldsymbol{Y}$ are fairly close to the identity matrix with respect to the spectral norm. Theorem 3.3, allows us to obtain better estimates on the impact of perturbation on the embedding error. From a random matrix theory perspective, this result provides insight about the conditioning of randomly subsampled randomized circulant matrices.

Repeated applications of Theorem 3.3 yields the following corollary which does not require the orthogonality of $\boldsymbol{x}$ and $\boldsymbol{y}$.

Corollary 3.4 *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 constructing $\boldsymbol{X}, \boldsymbol{Y}$ as described above and setting $\boldsymbol{M} = [\boldsymbol{X}\ \boldsymbol{Y}]$. With probability $1 - 6\exp(-\delta^2 k)$ (over $\boldsymbol{r}$ and selection of $\{\boldsymbol{X}_i\}_{i=1}^k$'s), we have that*

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

where $\mathbf{I}_\theta \in \mathbb{R}^{2k\times 2k}$ is given by the matrix

$$\begin{bmatrix} \mathbf{I}_k & \cos(\theta)\mathbf{I}_k \\ \cos(\theta)\mathbf{I}_k & \mathbf{I}_k \end{bmatrix}.$$

While the fundamental idea is to decouple the entries of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{X}^*\boldsymbol{h}$ into a nicer i.i.d. component $\boldsymbol{X}'^*\boldsymbol{h}$ and a perturbation $\boldsymbol{P}^*_{\boldsymbol{X}}\boldsymbol{h}$, it is still not clear how to relate these arguments to binary embedding. The relation becomes more clear when we consider the implications of sign mismatch between $\boldsymbol{X}'^*\boldsymbol{h}$ and $\boldsymbol{X}^*\boldsymbol{h}$. We already have a good understanding of $\boldsymbol{X}'^*\boldsymbol{h}$ due to its Gaussian nature. This allows us to sharply estimate the Hamming distance $\|\text{sgn}(\boldsymbol{X}'^*\boldsymbol{h}), \text{sgn}(\boldsymbol{Y}'^*\boldsymbol{h})\|_H$. It also means that, if the signs of $\boldsymbol{X}'^*\boldsymbol{h}$ and $\boldsymbol{X}^*\boldsymbol{h}$ mostly match, the problem is essentially solved. To address this, we consider the robuster version of sign mismatch with a parameter $\varepsilon > 0$ by defining the following events

$$\begin{aligned} E_{i,1} =& (\text{sgn}(\boldsymbol{h}^*\boldsymbol{X}'_i) \neq \text{sgn}(\boldsymbol{h}^*\boldsymbol{Y}'_i) \text{ and } |\boldsymbol{h}^*\boldsymbol{X}'_i|, |\boldsymbol{h}^*\boldsymbol{Y}'_i| > \varepsilon) \\ & \text{and } \text{sgn}(\boldsymbol{h}^*\boldsymbol{X}_i) = \text{sgn}(\boldsymbol{h}^*\boldsymbol{Y}_i), \\ E_{i,2} =& (\text{sgn}(\boldsymbol{h}^*\boldsymbol{X}'_i) = \text{sgn}(\boldsymbol{h}^*\boldsymbol{Y}'_i) \text{ and } |\boldsymbol{h}^*\boldsymbol{X}'_i|, |\boldsymbol{h}^*\boldsymbol{Y}'_i| > \varepsilon) \\ & \text{and } \text{sgn}(\boldsymbol{h}^*\boldsymbol{X}_i) \neq \text{sgn}(\boldsymbol{h}^*\boldsymbol{Y}_i). \end{aligned}$$

Here, $E_{i,1}$ and $E_{i,2}$ are the robust versions of the events where the signs associated with $\boldsymbol{X}_i, \boldsymbol{Y}_i$ are not consistent with the ones associated with $\boldsymbol{X}'_i, \boldsymbol{Y}'_i$. Robustness is enforced by requiring the products $\boldsymbol{h}^*\boldsymbol{X}'_i, \boldsymbol{h}^*\boldsymbol{Y}'_i$ to be ϵ away from zero.

Now observe that, in order for $E_{i,j}$'s to occur, we need $\max\{|\boldsymbol{h}^*\boldsymbol{P}_{\boldsymbol{X},i}|, |\boldsymbol{h}^*\boldsymbol{P}_{\boldsymbol{Y},i}|\} > \varepsilon$ i.e. the perturbation error has to be somewhat significant. Guaranteeing small perturbation error via Lemma 3.2 and Theorem 3.4 helps establish that $\{E_{i,j}\}_{i=1}^k$ occur rarely and $\|\text{sgn}(\boldsymbol{X}^*\boldsymbol{h}), \text{sgn}(\boldsymbol{X}'^*\boldsymbol{h})\|_H$ is rather small. With these, we eventually conclude that, for all $\boldsymbol{x}, \boldsymbol{y} \in \{\boldsymbol{v}_i\}_{i=1}^N$,

$$\begin{aligned} \|\text{sgn}(\boldsymbol{A}\boldsymbol{x}), \text{sgn}(\boldsymbol{A}\boldsymbol{y})\|_H &= \|\text{sgn}(\boldsymbol{X}^*\boldsymbol{h}), \text{sgn}(\boldsymbol{Y}^*\boldsymbol{h})\|_H \\ &\approx \|\text{sgn}(\boldsymbol{X}'^*\boldsymbol{h}), \text{sgn}(\boldsymbol{Y}'^*\boldsymbol{h})\|_H \\ &\approx k \cdot \text{ang}(\boldsymbol{x}, \boldsymbol{y}). \end{aligned}$$

4 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 [16–18]. This is the first such result for fast binary embedding and significantly improves over the related literature (e.g. [7, 8]). We believe the tools developed in this work can find broad applications in other nonlinear embedding tasks. For instance, our argument may be used to improve the concentration estimates of Fastfood features [8] 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 $\log N \lesssim \mathcal{O}(\sqrt{n})$. Overall, this work opens up several research directions.

- **Fast embedding in linear regime:** Does optimal fast binary embedding work with embedding dimension $\mathcal{O}(n)$? 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 [7, 10] 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 [19–21]. It would also be of interest to investigate quadratic samples arising in phase retrieval [22, 23].

- **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 [7].

Acknowlegments

S.O. would like to thank Felix Yu for his comments on the manuscript and stimulating discussions.

5 References

[1] Yaniv Plan and Roman Vershynin, "Dimension reduction by random hyperplane tessellations," *Discrete & Computational Geometry*, vol. 51, no. 2, pp. 438–461, 2014.

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

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

[4] Yaniv Plan and Roman Vershynin, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *Information Theory, IEEE Transactions on*, vol. 59, no. 1, pp. 482–494, 2013.

[5] 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*, vol. 59, no. 4, pp. 2082–2102, 2013.

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

[7] 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.

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

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

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

[11] 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*, pp. 425–439. Springer, 2014.

[12] 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*. IEEE, 2011, pp. 817–824.

[13] Samet Oymak, "Near-optimal sample complexity bounds for circulant binary embedding," *arXiv:1603.03178*, 2016.

[14] Mark Rudelson and Roman Vershynin, "Hanson-wright inequality and sub-gaussian concentration," *Electron. Commun. Probab*, vol. 18, no. 0, 2013.

[15] Joel A Tropp, "On the conditioning of random subdictionaries," *Applied and Computational Harmonic Analysis*, vol. 25, no. 1, pp. 1–24, 2008.

[16] 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*. ACM, 2006, pp. 557–563.

[17] Felix Krahmer and Rachel Ward, "New and improved johnson-lindenstrauss embeddings via the restricted isometry property," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 3, pp. 1269–1281, 2011.

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

[19] 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.

[20] Laurent Jacques, "A quantized johnson–lindenstrauss lemma: The finding of buffon's needle," *Information Theory, IEEE Transactions on*, vol. 61, no. 9, pp. 5012–5027, 2015.

[21] Laurent Jacques and Valerio Cambareri, "Time for dithering: fast and quantized random embeddings via the restricted isometry property," *arXiv preprint arXiv:1607.00816*, 2016.

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

[23] Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi, "Phase retrieval via wirtinger flow: Theory and algorithms," *Information Theory, IEEE Transactions on*, vol. 61, no. 4, pp. 1985–2007, 2015.