Near-Optimal Bounds for Binary Embeddings of Arbitrary Sets

Samet Oymak*† Benjamin Recht†‡

Abstract

We study embedding a subset K of the unit sphere to the Hamming cube $\{-1,+1\}^m$. We characterize the tradeoff between distortion and sample complexity m in terms of the Gaussian width $\omega(K)$ of the set. For subspaces and several *structured-sparse* sets we show that Gaussian maps provide the optimal tradeoff $m \sim \delta^{-2}\omega^2(K)$, in particular for δ distortion one needs $m \approx \delta^{-2}d$ where d is the subspace dimension. For general sets, we provide sharp characterizations which reduces to $m \approx \delta^{-4}\omega^2(K)$ after simplification. We provide improved results for local embedding of points that are in close proximity of each other which is related to locality sensitive hashing. We also discuss faster binary embedding where one takes advantage of an initial sketching procedure based on Fast Johnson-Lindenstauss Transform. Finally, we list several numerical observations and discuss open problems.

1 Introduction

Thanks to applications in high-dimensional statistics and randomized linear algebra, recent years saw a surge of interest in dimensionality reduction techniques. As a starting point, Johnson-Lindenstauss lemma considers embedding a set of points in a high dimensional space to a lower dimension while approximately preserving pairwise ℓ_2-distances. While ℓ_2 distance is the natural metric for critical applications, one can consider embedding with arbitrary norms/functions. In this work, we consider embedding a high-dimensional set of points K to the hamming cube in a lower dimension, which is known as binary embedding. Binary embedding is a natural problem arising from quantization of the measurements and is connected to 1-bit compressed sensing as well as locality sensitive hashing [1, 3, 14, 20, 22, 25, 27]. In particular, given a subset K of the unit sphere in $\mathbb{R}^n$, and a dimensionality reduction map $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we wish to ensure that

$$\left|\frac{1}{m}\|\text{sgn}(\boldsymbol{A}\boldsymbol{x}),\text{sgn}(\boldsymbol{A}\boldsymbol{y})\|_H - \text{ang}(\boldsymbol{x},\boldsymbol{y})\right| \le \delta. \tag{1.1}$$

Here $\text{ang}(\boldsymbol{x},\boldsymbol{y}) \in [0,1]$ is the geodesic distance between the two points which is obtained by normalizing the smaller angle between $\boldsymbol{x},\boldsymbol{y}$ by π. $\|\boldsymbol{a},\boldsymbol{b}\|_H$ returns the Hamming distance between the vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ i.e. the number of entries for which $\boldsymbol{a}_i \ne \boldsymbol{b}_i$. Relation between the geodesic distance and Hamming distance arises naturally when one considers a Gaussian map. For a vector $\boldsymbol{g} \sim \mathcal{N}(0,\mathbf{I}_n)$, $\|\text{sgn}(\boldsymbol{g}^T\boldsymbol{x}),\text{sgn}(\boldsymbol{g}^T\boldsymbol{y})\|_H$ has a Bernoulli distribution with mean $\text{ang}(\boldsymbol{x},\boldsymbol{y})$. This follows from the fact that $\boldsymbol{g}$ corresponds to a uniformly random hyperplane and $\boldsymbol{x},\boldsymbol{y}$ lies on the opposite sides of the hyperplane with probability $\text{ang}(\boldsymbol{x},\boldsymbol{y})$. Consequently, when $\boldsymbol{A}$ has independent standard normal entries, a standard application of Chernoff bound shows that (1.1) holds with probability $1-\exp(-2\delta^2 m)$ for a given $(\boldsymbol{x},\boldsymbol{y})$ pair. This argument is sufficient to ensure binary embedding of finite set of points in a similar manner to Johnson-Lindenstrauss Lemma i.e. $m \ge \mathcal{O}(\frac{\log p}{\delta^2})$ samples are sufficient to ensure (1.1) for a set of p points.

Embedding for arbitrary sets: When dealing with sets that are not necessarily discrete, highly nonlinear nature of the sign function makes the problem more challenging. Recent literature shows that for a set K with infinitely many points, the notion of size can be captured by mean width $\omega(K)$

$$\omega(K) = \mathbb{E}_{\boldsymbol{g}\sim\mathcal{N}(0,\mathbf{I}_n)}[\sup_{\boldsymbol{v}\in K}\boldsymbol{g}^T\boldsymbol{v}]. \tag{1.2}$$

*Simons Institute, UC Berkeley

†Department of Electrical Engineering and Computer Science, UC Berkeley

‡Department of Statistics, UC Berkeley, Berkeley CA

For the case of linear embedding, where the goal is preserving ℓ_2 distances of the points, it is well-known that $m \sim \mathcal{O}(\delta^{-2}\omega^2(K))$ samples ensure δ-distortion for a large class of random maps. The reader is referred to a growing number of literature for results on ℓ_2 embedding of continuous sets [4,16,17,23]. There is much less known about the the randomized binary embeddings of arbitrary sets and the problem was initially studied by Plan and Vershynin [21] when the mapping $\boldsymbol{A}$ is Gaussian. Remarkably, the authors show that embedding (1.1) is possible if $m \gtrsim \delta^{-6}\omega^2(K)$. This result captures the correct relation between the sample complexity m and the set size $\omega(K)$ in a similar manner to ℓ_2-embedding. On the other hand, it is rather weak when it comes to the dependence on the distortion δ. Following from the result on finite set of points, intuitively, this dependence should be δ^{-2} instead of δ^{-6}. For the set of sparse vectors Jacques et al. obtained near-optimal sample complexity $\mathcal{O}(\delta^{-2}s\log(n/\delta))$ [13]. Related to this, Yi et al. [29] considered the same question where they showed that δ^{-4} is achievable for simpler sets such as intersection of a subspace and the unit sphere. More recently in [12] Jacques studies a related problem in which samples $\boldsymbol{Ax}$ are uniformly quantized instead of being discretized to $\{+1,-1\}$. There is also a growing amount of literature related to binary embedding and one-bit compressed sensing [1,3,22,25]. In this work, we significantly improve the distortion dependence bounds for the binary embedding of arbitrary sets. For structured set such as subspaces, sparse signals, and low-rank matrices we obtain the optimal dependence $\mathcal{O}(\delta^{-2}\omega^2(K))$. In particular, we show that a d dimensional subspace can be embedded with $m = \mathcal{O}(\delta^{-2}d)$ samples. For general sets, we find a bound on the embedding size in terms of the covering number and local mean-width which is a quantity always upper bounded by mean width. While the specific bound will be stated later on, in terms of mean-width we show that $m = \mathcal{O}(\delta^{-4}\omega^2(K))$ samples are sufficient for embedding. For important sets such as subspaces, set of sparse vectors and low-rank matrices our results have a simple message: Binary embedding works just as well as linear embedding.

Locality sensitive properties: In relation to locality sensitive hashing [2,6], we might want $\boldsymbol{x},\boldsymbol{y}$ to be close to each other if and only if $\text{sgn}(\boldsymbol{Ax}),\text{sgn}(\boldsymbol{Ay})$ are close to each other. Formally, we consider the following questions regarding a binary embedding.

- How many samples do we need to ensure that for all $\boldsymbol{x},\boldsymbol{y}$ satisfying $\text{ang}(\boldsymbol{x},\boldsymbol{y}) > \delta$, we have that $\frac{1}{m}\|\text{sgn}(\boldsymbol{Ax}),\text{sgn}(\boldsymbol{Ay})\|_H \gtrsim \mathcal{O}(\delta)$?
- How many samples do we need to ensure that for all $\boldsymbol{x},\boldsymbol{y}$ satisfying $\frac{1}{m}\|\text{sgn}(\boldsymbol{Ax}),\text{sgn}(\boldsymbol{Ay})\|_H > \delta$, we have that $\text{ang}(\boldsymbol{x},\boldsymbol{y}) \gtrsim \mathcal{O}(\delta)$?

These questions are less restrictive compared to requiring (1.1) for all $\boldsymbol{x},\boldsymbol{y}$ as a result they should intuitively require less sample complexity. Indeed, we show that the distortion dependence can be improved by a factor of δ e.g. for a subspace one needs only $m = \mathcal{O}(\frac{\omega^2(K)}{\delta}\log\delta^{-1})$ samples and for general sets one needs only $m = \mathcal{O}(\frac{\omega^2(K)}{\delta^3}\log\delta^{-1})$ samples. Our distortion dependence is an improvement over the related results of Jacques [12].

Sketching for binary embedding: While providing theoretical guarantees are beneficial, from an application point of view, efficient binary embedding is desirable. A good way to accomplish this is to make use of matrices with fast multiplication. Such matrices include subsampled Hadamard or Fourier matrices as well as super-sparse ensembles. Efficient sketching matrices found applications in a wide range of problems to speed up machine learning algorithms [7,10,28]. For binary embedding, recent works in this direction include [29,30] which have limited theoretical guarantees for discrete sets. In Section 6 we study this procedure for embedding arbitrary sets and provide guarantees to achieve faster embeddings.

2 Main results on binary embedding

Let us introduce the main notation that will be used for the rest of this work. $\mathcal{S}^{n-1}$ and $\mathcal{B}^{n-1}$ are the unit Euclidian sphere and ball respectively. $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes a Gaussian vector with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. $\mathbf{I}_n$ is the identity matrix of size n. A matrix with independent $\mathcal{N}(0,1)$ entries will be called standard Gaussian matrix. Gaussian width $\omega(\cdot)$ is defined above in (1.2) and will be used to capture the size of our set of interest $K \subset \mathbb{R}^n$. $c, C, \{c_i\}_{i=0}^{\infty}, \{C_i\}_{i=0}^{\infty}$ denote positive constants that may vary from line to line. ℓ_p norm is denoted by $\|\cdot\|_{\ell_p}$ and $\|\cdot\|_0$ returns the sparsity of a vector. Given a set K denote $\text{cone}(K) = \{\alpha \boldsymbol{v} \mid \alpha \geq 0,\ \boldsymbol{v} \in K\}$ by $\hat{K}$ and define the binary embedding as follows.

Definition 2.1 (δ-binary embedding) *Given $\delta \in (0,1)$, $f : \mathbb{R}^n \to \{0,1\}^m$ is a δ-binary embedding of the set $\mathcal{C}$ if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$, we have that*

$$|\frac{1}{m}\|f(\boldsymbol{x}), f(\boldsymbol{y})\|_H - ang(\boldsymbol{x}, \boldsymbol{y})| \leq \delta.$$

With this notation, we are in a position to state our first result which provides distortion bounds on binary embedding in terms of the Gaussian width $\omega(K)$.

Theorem 2.2 *Suppose $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ has independent $\mathcal{N}(0,1)$ entries. Given set $K \subset \mathcal{S}^{n-1}$ and a constant $1 > \delta > 0$, there exists positive absolute constants c_1, c_2 such that the followings hold.*

- $\hat{K}$ **is a subspace:** *Whenever $m \geq c_1 \frac{\omega^2(K)}{\delta^2}$, with probability $1 - \exp(-c_2\delta^2 m)$, $\boldsymbol{A}$ (i.e. $f : \boldsymbol{x} \to sgn(\boldsymbol{A}\boldsymbol{x})$) is a δ-binary embedding for K.*
- K **is arbitrary:** *Whenever $m \geq c_1 \frac{\omega^2(K)}{\delta^4} \log\frac{1}{\delta}$, with probability $1 - \exp(-c_2\delta^2 m)$, $\boldsymbol{A}$ is a δ-binary embedding for K.*

For arbitrary sets, we later on show that the latter bound can be improved to $\frac{\omega^2(K)}{\delta^4}$. This is proven by applying a sketching procedure and is deferred to Section 6.

The next theorem states our results on the local properties of binary embedding namely it characterizes the behavior of the embedding in a small neighborhood of points.

Definition 2.3 (Local δ-binary embedding) *Given $\delta \in (0,1)$, $f : \mathbb{R}^n \to \{0,1\}^m$ is a local δ-binary embedding of the set $\mathcal{C}$ if there exists constants c_{up}, c_{low} such that*

- *For all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ satisfying $\|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2} \geq \delta$: $|\frac{1}{m}\|f(\boldsymbol{x}), f(\boldsymbol{y})\|_H - ang(\boldsymbol{x}, \boldsymbol{y})| \geq c_{low}\delta$.*
- *For all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$ satisfying $\|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2} \leq \delta/\sqrt{\log \delta^{-1}}$: $|\frac{1}{m}\|f(\boldsymbol{x}), f(\boldsymbol{y})\|_H - ang(\boldsymbol{x}, \boldsymbol{y})| \leq c_{up}\delta$.*

Theorem 2.4 *Suppose $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is a standard Gaussian matrix. Given set $K \subset \mathcal{S}^{n-1}$ and a constant $1 > \delta > 0$, there exist positive absolute constants c_1, c_2 such that the followings hold.*

- $\hat{K}$ **is a subspace:** *Whenever $m \geq c_1 \frac{\omega^2(K)}{\delta} \log\frac{1}{\delta}$, with probability $1 - \exp(-c_2\delta m)$, $\boldsymbol{A}$ is a local δ-binary embedding for K.*
- K **is arbitrary:** *Whenever $m \geq c_1 \frac{\omega^2(K)}{\delta^3} \log\frac{1}{\delta}$, with probability $1 - \exp(-c_2\delta m)$, $\boldsymbol{A}$ is a local δ-binary embedding for K.*

We should emphasize that the second statements of Theorems 2.2 and 2.4 are simplified versions of a better but more involved result. To state this, we need to introduce two relevant definitions.

- **Covering number:** Let N_ε be the ε-covering number of K with respect to the ℓ_2 distance.
- **Local set:** Given $\alpha > 0$, define the local set K_ε to be

$$K_\varepsilon = \{\boldsymbol{v} \in \mathbb{R}^n \;\Big|\; \boldsymbol{v} = \boldsymbol{a} - \boldsymbol{b},\ \boldsymbol{a}, \boldsymbol{b} \in K,\ \|\boldsymbol{a} - \boldsymbol{b}\|_{\ell_2} \leq \varepsilon\}.$$

In this case, our tighter sample complexity estimate is as follows.

Theorem 2.5 *Suppose $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is a standard Gaussian matrix. Given set $K \subset \mathcal{S}^{n-1}$ and a constant $1 > \delta > 0$, there exist positive absolute constants c, c_1, c_2 such that the followings hold. Setting $\varepsilon = c\delta/\sqrt{\log \delta^{-1}}$,*

- *whenever* $m \geq c_1 \max\{\delta^{-2}\log N_\varepsilon, \delta^{-3}\omega^2(K_\varepsilon)\}$*, with probability* $1-\exp(-c_2\delta^2 m)$*,* $\boldsymbol{A}$ *is a* δ*-binary embedding for* K*.*
- *whenever* $m \geq c_1 \max\{\delta^{-1}\log N_\varepsilon, \delta^{-3}\omega^2(K_\varepsilon)\}$*, with probability* $1-\exp(-c_2\delta m)$*,* $\boldsymbol{A}$ *is a local* δ*-binary embedding for* K*.*

Implications for structured sets: This result is particularly beneficial for important low-dimensional sets for which we have a good understanding of the local set K_ε and covering number N_ε. Such sets include

- $\hat{K}$ is a subspace,
- $\hat{K}$ is a union of subspaces for instance set of d-sparse signals (more generally signals that are sparse with respect to a dictionary),
- $\hat{K}$ is set of rank-d matrices in $\mathbb{R}^{n_1\times n_2}$ where $n = n_1 \times n_2$,
- $\hat{K}$ is the set of (d,l) group-sparse signals. This scenario assumes small entry-groups $\{G_i\}_{i=1}^N \subset \{1,2,\dots,n\}$ which satisfy $|G_i| \leq l$. A vector is considered group sparse if its nonzero entries lie in at most d of these groups. Sparse signals is a special case for which each entry is a group and group size $l = 1$.

These sets are of fundamental importance for high-dimensional statistics and machine learning. They also have better distortion dependence. In particular, as a function of the model parameters (e.g. d,n,l) there exists a number $C(K)$ such that (e.g. [5])

$$\log N_\varepsilon \leq C(K)\log\frac{1}{\varepsilon},\ \omega^2(K_\varepsilon) \leq \varepsilon^2 C(K). \tag{2.1}$$

For all of the examples above, we either have that $C(K) \sim \omega^2(K)$ or the simplified closed form upper bounds of these quantities (in terms of d,l,n) are the same. [5,26]. Consequently, Theorem 2.5 ensures that for such low-dimensional sets we have the near-optimal dependence for binary embedding namely $m \sim \frac{\omega^2(K)}{\delta^2}\log\frac{1}{\delta}$. This follows from the improved dependencies $\omega^2(K_\varepsilon) \sim \delta^2\omega^2(K)/\log\delta^{-1}$ and $\log N_\varepsilon \sim \omega^2(K)\log\delta^{-1}$.

While these bounds are near-optimal they can be further improved and can be matched to the bounds for linear embedding namely $m \sim \delta^{-2}\omega^2(K)$. The following theorem accomplishes this goal and allows us to show that binary embedding performs as good as linear embedding for a class of useful sets..

Theorem 2.6 *Suppose* (2.1) *holds for all* $\varepsilon > 0$*. Then, there exists* c_1, c_2 *such that if* $m > c_1\delta^{-2}C(K)$*,* $\boldsymbol{A}$ *is a* δ*-binary embedding for* K *with probability* $1-\exp(-c_2\delta^2 m)$*.*

The rest of the paper is organized as follows. In Section 3 we give a proof of Theorem 2.2. Section 4 specifically focuses on obtaining optimal distortion guarantees for structured sets and provides the technical argument for Theorem 2.6. Section 5 provides a proof for Theorem 2.4 which is closely connected to the proof of Theorem 2.2. Section 6 focuses on sketching and fast binary embedding techniques for improved guarantees. Numerical experiments and our observations are presented in Section 7.

3 Main proofs

Let K be an arbitrary set over the unit sphere, $\bar{K}_\varepsilon$ be an ε covering of K and $N_\varepsilon = |\bar{K}_\varepsilon|$. Due to Sudakov Minoration there exists an absolute constant $c > 0$ for which $\log N_\varepsilon \leq \frac{c\omega^2(K)}{\varepsilon^2}$. This bound is often suboptimal for instance, if $\hat{K}$ is a d dimensional subspace, then we have that

$$\log N_\varepsilon \leq \omega^2(K)\log(\frac{c}{\varepsilon})$$

where $\omega^2(K) \approx d$. Recall that our aim is to ensure that for near optimal values of m, all $\boldsymbol{x}, \boldsymbol{y} \in K$ obeys

$$|\text{ang}(\boldsymbol{x}, \boldsymbol{y}) - \frac{1}{m}\|\text{sgn}(\boldsymbol{A}\boldsymbol{x}), \text{sgn}(\boldsymbol{A}\boldsymbol{y})\|_H| \le \delta. \tag{3.1}$$

This task is simpler when K is a finite set. In particular, when $\boldsymbol{A}$ has standard normal entries we have that $\|\text{sgn}(\boldsymbol{A}\boldsymbol{x}), \text{sgn}(\boldsymbol{A}\boldsymbol{y})\|_H$ is sum of m i.i.d. Bernoulli random variables with mean $\text{ang}(\boldsymbol{x}, \boldsymbol{y})$. This ensures

$$\text{ang}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}[\frac{1}{m}\|\text{sgn}(\boldsymbol{A}\boldsymbol{x}), \text{sgn}(\boldsymbol{A}\boldsymbol{y})\|_H],$$
$$\mathbb{P}(|\text{ang}(\boldsymbol{x}, \boldsymbol{y}) - \frac{1}{m}\|\text{sgn}(\boldsymbol{A}\boldsymbol{x}), \text{sgn}(\boldsymbol{A}\boldsymbol{y})\|_H| > \delta) \le \exp(-2\delta^2 m).$$

Consequently, as long as we are dealing with finite sets one can use a union bound. This argument yields the following lemma for $\bar{K}_\varepsilon$.

Lemma 3.1 *Assume $m \ge \frac{2}{\delta^2}\log N_\varepsilon$. Then, with probability $1 - \exp(-\delta^2 m)$ all points $\boldsymbol{x}, \boldsymbol{y}$ of $\bar{K}_\varepsilon$ obeys (3.1).*

Using this as a starting point, we will focus on the effect of continuous distortions to move from $\bar{K}_\varepsilon$ to K. The following theorem considers the second statement of Definition 2.3 and states our main result on local deviations.

Theorem 3.2 *Suppose $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is a standard Gaussian matrix. Given $1 > \delta > 0$, pick $c > 0$ to be a sufficiently large constant, set $c\varepsilon = \delta(\log\frac{1}{\delta})^{-1/2}$ and assume that*

$$m \ge c\max\{\delta^{-3}\omega^2(K_\varepsilon), \frac{1}{\delta}\log N_\varepsilon\}.$$

Then, with probability $1 - 2\exp(-\delta m/64)$, for all $\boldsymbol{x}, \boldsymbol{y} \in K$ obeying $\|\boldsymbol{y} - \boldsymbol{x}\|_{\ell_2} \le \varepsilon$ we have

$$m^{-1}\|\boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y}\|_H \le \delta.$$

This theorem is stated in terms of Gaussian width of K_ε and the covering number N_ε. Making use of the fact that $\omega(K_\varepsilon) \le 2\omega(K)$ and $\log N_\varepsilon \le c\frac{\omega^2(K)}{\varepsilon^2}$ (and similar simplification for subspaces), we arrive at the following corollary.

Corollary 3.3 *Suppose $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is a standard Gaussian matrix.*

- *When K is a general set, set*

$$m \ge c\delta^{-3}\log\frac{1}{\delta}\omega^2(K),$$

- *When $\hat{K}$ is a d-dimensional subspace, set*

$$m \ge c\delta^{-1}\log\frac{1}{\delta}d,$$

for sufficiently large constant $c > 0$. Then, with probability $1 - 2\exp(-\delta m/64)$, for all $\boldsymbol{x}, \boldsymbol{y} \in K$ obeying $\|\boldsymbol{y} - \boldsymbol{x}\|_{\ell_2} \le c^{-1}\delta(\log\frac{1}{\delta})^{-1/2}$ we have

$$m^{-1}\|\boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y}\|_H \le \delta.$$

3.1 Preliminary results for the proof

We first define relevant quantities for the proof. Given a vector $\boldsymbol{x}$ let $\tilde{\boldsymbol{x}}$ denote the vector obtained by sorting absolute values of $\boldsymbol{x}$ decreasingly. Define

$$\|\boldsymbol{x}\|_{k+} = \sum_{i=1}^{k} \tilde{\boldsymbol{x}}, \;\; \|\boldsymbol{x}\|_{k-} = \sum_{i=n-k+1}^{n} \tilde{\boldsymbol{x}}.$$

In words, $k+$ and $k-$ functions returns the ℓ_1 norms of the top k and bottom k entries respectively. The next lemma illustrates why $k+$ and $k-$ functions are useful for our purposes.

Lemma 3.4 *Given vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, suppose $\|\boldsymbol{x}\|_{k-} > \|\boldsymbol{y}\|_{k+}$. Then, we have that*

$$\|\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{y}\|_H < k.$$

Proof Suppose at ith location $\text{sgn}(\boldsymbol{x}_i) \neq \text{sgn}(\boldsymbol{x}_i + \boldsymbol{y}_i)$. This implies $|\boldsymbol{y}_i| \geq |\boldsymbol{x}_i|$. If $\|\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{y}\|_H \geq k$ it implies that there is a set $S \subset \{1, 2, \dots, n\}$ of size k. Over this subset $\|\boldsymbol{y}_S\|_{\ell_1} \geq \|\boldsymbol{x}_S\|_{\ell_1}$ which yields

$$\|\boldsymbol{y}\|_{k+} \geq \|\boldsymbol{y}_S\|_{\ell_1} \geq \|\boldsymbol{x}_S\|_{\ell_1} \geq \|\boldsymbol{x}\|_{k-}.$$

This contradicts with the initial assumption. ■

The next two subsections obtains bounds on $k+$ and $k-$ functions of a Gaussian vector in order to be able to apply Lemma 3.4 later on.

3.1.1 Obtain an estimate on $k+$

The reader is referred to Lemma A.4 for a proof of the following result.

Lemma 3.5 *Suppose $\boldsymbol{g} \sim \mathcal{N}(0, \mathbf{I}_n)$ and $0 < \delta < 1$ is a sufficiently small constant. Set $k = \delta n$. There exists constants $c, C > 0$ such that for $n > C\delta^{-1}$ we have that*

$$\mathbb{E}[\|\boldsymbol{g}\|_{k+}] = \mathbb{E}[\sum_{i=1}^{\delta n} \tilde{\boldsymbol{g}}_i] \leq c\delta n \sqrt{\log \frac{1}{\delta}}.$$

The next lemma provides an upper bound for $\sup_{\boldsymbol{x}\in\mathcal{C}} \|\boldsymbol{A}\boldsymbol{x}\|_{k+}$ in expectation.

Lemma 3.6 *Let $\boldsymbol{A}$ be a standard Gaussian matrix and $\mathcal{C} \subset \mathbb{R}^n$. Define $diam(\mathcal{C}) = \sup_{\boldsymbol{v}\in\mathcal{C}} \|\boldsymbol{v}\|_{\ell_2}$. Set $k = \delta m$ for a small constant $0 < \delta < 1$. Then, there exists constants $c, C > 0$ such that for $m > C\delta^{-1}$*

$$\mathbb{E}[\sup_{\boldsymbol{x}\in\mathcal{C}} \|\boldsymbol{A}\boldsymbol{x}\|_{k+}] \leq c\, diam(\mathcal{C}) m\delta \log \frac{1}{\delta} + \sqrt{m\delta}\,\omega(\mathcal{C}).$$

Proof Let $S_k = \{\boldsymbol{v} \in \mathbb{R}^n \;\big|\; \|\boldsymbol{v}\|_0 \leq k, \;\; \|\boldsymbol{v}\|_{\ell_\infty} \leq 1\}$. Observe that

$$\sup_{\boldsymbol{x}\in\mathcal{C}} \|\boldsymbol{A}\boldsymbol{x}\|_{k+} = \sup_{\boldsymbol{x}\in\mathcal{C}, \boldsymbol{v}\in S_k} \boldsymbol{v}^T \boldsymbol{A}\boldsymbol{x}.$$

Applying Slepian's Lemma A.6 (also see [24] or [11]), we find that, for $\boldsymbol{g} \sim \mathcal{N}(0, \mathbf{I}_m), \boldsymbol{h} \sim \mathcal{N}(0, \mathbf{I}_n), g \sim \mathcal{N}(0, 1)$,

$$\begin{aligned}
\mathbb{E}[\sup_{\boldsymbol{x}\in\mathcal{C}} \sup_{\boldsymbol{v}\in S_k} \boldsymbol{v}^T \boldsymbol{A}\boldsymbol{x}] &\leq \mathbb{E}[\sup_{\boldsymbol{x}\in\mathcal{C}, \boldsymbol{v}\in S_k} \|\boldsymbol{x}\|_{\ell_2} \boldsymbol{v}^T \boldsymbol{g} + \boldsymbol{x}^T \boldsymbol{h} \|\boldsymbol{v}\|_{\ell_2}] + \mathbb{E}[\sup_{\boldsymbol{x}\in\mathcal{C}, \boldsymbol{v}\in S_k} \|\boldsymbol{x}\|_{\ell_2} \|\boldsymbol{v}\|_{\ell_2} |g|] \\
&\leq \text{diam}(\mathcal{C})\, \mathbb{E}[\|\boldsymbol{g}\|_{k+}] + \sqrt{k}\omega(\mathcal{C}) + \text{diam}(\mathcal{C})\sqrt{k} \\
&\leq c\text{diam}(\mathcal{C}) m\delta \sqrt{\log \frac{1}{\delta}} + \sqrt{m\delta}\,\omega(\mathcal{C})
\end{aligned}$$

which is the advertised result. For the final line we made use of the estimate obtained in Lemma 3.5. ■

3.1.2 Obtain an estimate on $k-$

As the next step we obtain an estimate of $\|\boldsymbol{g}\|_{k-}$ for $\boldsymbol{g} \sim \mathcal{N}(0, \mathbf{I}_m)$ by finding a simple deviation bound.

Lemma 3.7 *Suppose $\boldsymbol{g} \sim \mathcal{N}(0, \mathbf{I}_m)$. Let γ_α be the number for which $\mathbb{P}(|g| \le \gamma_\alpha) = \alpha$ where $g \sim \mathcal{N}(0,1)$ (i.e. the inverse cumulative density function of $|g|$). Then,*

$$\mathbb{P}(\|\boldsymbol{g}\|_{\delta m-} \ge m\frac{\delta\gamma_{\delta/2}}{4}) \ge 1 - \exp(-\frac{\delta m}{32}).$$

γ_α trivially obeys $\gamma_\alpha \ge \sqrt{\frac{\pi}{2}}\alpha$. This yields

$$\mathbb{P}(\|\boldsymbol{g}\|_{\delta m-} \ge \frac{m\delta^2}{8}) \ge 1 - \exp(-\frac{\delta m}{32}).$$

Proof We show this by ensuring that, with high probability, among the bottom $m\delta$ entries at least $m\delta/4$ of them are greater than $\gamma_{\delta/2}$. This is a standard application of Chernoff bound. Let a_i be a random variable which is 1 if $|\boldsymbol{g}_i| \le \beta$ and 0 else. Then, if $s := \sum_{i=1}^n a_i \le \frac{3\delta}{4}$, it would imply that at most $\frac{3\delta}{4}$ of entries of $|\boldsymbol{g}|$ are less than β. We will argue that, this is indeed the case with the advertised probability for $\beta = \gamma_{\delta/2}$. For $\beta = \gamma_{\delta/2}$, we have that $\mathbb{P}(a_i = 1) = \frac{\delta}{2}$. Hence, applying a standard Chernoff bound, we obtain

$$\mathbb{P}(\sum_{i=1}^m a_i > m\frac{3\delta}{4}) \le \exp(-\frac{\delta m}{32}).$$

With this probability, we have that out of the bottom δm entries of $\boldsymbol{g}$ at least $\frac{\delta}{4}$ of them are greater than or equal to $\gamma_{\delta/2}$. This implies $\|\boldsymbol{g}\|_{\delta m-} \ge m\frac{\delta\gamma_{\delta/2}}{4}$. To conclude we use the standard fact that inverse absolute-gaussian cumulative function obeys $\gamma_\alpha \ge \sqrt{\frac{\pi}{2}}\alpha$. ■

3.2 Proof of Theorem 3.2

We are in a position to prove Theorem 3.2. Without losing generality, we may assume $\delta < \delta'$ where δ' is a sufficiently small constant. The result for $\delta \ge \delta'$ is implied by the case $\delta = \delta'$.

Proof Set $\varepsilon > 0$ to be $c'\varepsilon = \delta/\sqrt{\log\delta^{-1}}$ for a constant $c' \ge 1$ to be determined. Using Lemma 3.4, the proof can be reduced to showing the following claim: Under given conditions, all $\boldsymbol{x} \in \bar{K}_\varepsilon$, $\boldsymbol{y} \in K$ satisfying $\|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2} \le \varepsilon$ obey

$$\|\boldsymbol{A}\boldsymbol{x}\|_{\delta m-} \ge \|\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x})\|_{\delta m+}. \tag{3.2}$$

To show this, we shall apply a union bound. For a particular $\boldsymbol{x} \in \bar{K}_\varepsilon$ applying Lemma 3.7, we know that

$$\mathbb{P}(\|\boldsymbol{A}\boldsymbol{x}\|_{\delta m-} \ge \frac{m\delta^2}{8}) \ge 1 - \exp(-\frac{\delta m}{32}). \tag{3.3}$$

Using a union bound, we find that if $N_\varepsilon < \exp(\frac{\delta m}{64})$, with probability $1 - \exp(-\frac{\delta m}{64})$, all $\boldsymbol{x} \in \bar{K}_\varepsilon$ obeys the relation above. This requires $m \ge \frac{64}{\delta}\log N_\varepsilon$ which is satisfied by assumption.

We next show that given $\boldsymbol{x} \in \bar{K}_\varepsilon$ and $\boldsymbol{y} \in K$ that is in the ε neighborhood of $\boldsymbol{x}$ we have that

$$\|\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x})\|_{\delta m+} \le \frac{m\delta^2}{8}.$$

Observe that $\boldsymbol{x} - \boldsymbol{y} \in K_\varepsilon$, consequently, applying Lemma 3.6 and using $\varepsilon\sqrt{\delta m}$-Lipschitzness of $f(\boldsymbol{A}) = \sup_{\boldsymbol{v}\in K_\varepsilon}\|\boldsymbol{A}\boldsymbol{v}\|_{\delta m+}$, with probability $1 - \exp(-\delta m)$, for an absolute constant $c > 0$ we have that

$$\sup_{\boldsymbol{v}\in K_\varepsilon}\|\boldsymbol{A}\boldsymbol{v}\|_{\delta m+} \le c\varepsilon\delta\sqrt{\log\frac{1}{\delta}}m + \sqrt{\delta m}\omega(K_\varepsilon). \tag{3.4}$$

Following (3.3) and (3.4), we simply need to determine the conditions for which

$$c\varepsilon\delta\sqrt{\log\frac{1}{\delta}}m + \sqrt{m}\delta\omega(K_\varepsilon) \leq \frac{m\delta^2}{8}.$$

This inequality holds if

- $cc' \leq 1/16$,
- $m \geq 256\delta^{-3}\omega^2(K_\varepsilon)$.

To ensure this we can pick m, c' to satisfy the conditions while keeping initial assumptions intact. With these, (3.2) is guaranteed to hold concluding the proof. ∎

3.2.1 Proof of Corollary 3.3

We need to substitute the standard covering and local-width bounds to obtain this result while again setting $c\varepsilon = \delta/\sqrt{\log\delta^{-1}}$. For general K, simply use the estimates $\omega(K_\varepsilon) \leq 2\omega(K)$, $\log N_\varepsilon \leq c\omega(K)^2/\varepsilon^2$. When $\hat{K}$ is a d dimensional subspace we use the estimates $\omega(K_\varepsilon) \leq \varepsilon\sqrt{d}$ and $\log N_\varepsilon \leq cd\log\varepsilon^{-1}$ and use the fact that $\log\varepsilon^{-1} \sim \log\delta^{-1}$.

3.3 Proof of Theorem 2.5: First Statement

Proof of the first statement of Theorem 2.5 follows by merging the discrete embedding and "small deviation" arguments. We restate it below.

Theorem 3.8 *Suppose $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is a standard Gaussian matrix. Set $c\varepsilon = \delta/\sqrt{\log\delta^{-1}}$. $\boldsymbol{A}$ provides a δ-binary embedding of K with probability $1 - \exp(-c_2\delta^2 m)$ if the number of samples satisfy the bound*

$$m \geq c_1 \max\{\delta^{-2}\log N_\varepsilon, \delta^{-3}\omega^2(K_\varepsilon)\}.$$

Proof Given $\delta > 0$, $c\varepsilon = \delta/\sqrt{\log\delta^{-1}}$, applying Lemma 3.1 we have that all $\boldsymbol{x}', \boldsymbol{y}' \in \bar{K}_\varepsilon$ obeys (3.1) whenever $m \geq 2\delta^{-2}\log N_\varepsilon$. Next whenever the conditions of Theorem 3.2 are satisfied all $\boldsymbol{x}, \boldsymbol{y} \in K$, $\boldsymbol{x}', \boldsymbol{y}' \in \bar{K}_\varepsilon$ satisfying $\|\boldsymbol{x}' - \boldsymbol{x}\|_{\ell_2} \leq \varepsilon$, $\|\boldsymbol{y}' - \boldsymbol{y}\|_{\ell_2} \leq \varepsilon$, obeys

$$m^{-1}\|\boldsymbol{A}\boldsymbol{x}', \boldsymbol{A}\boldsymbol{x}\|_H \leq \delta,\ m^{-1}\|\boldsymbol{A}\boldsymbol{y}', \boldsymbol{A}\boldsymbol{y}\|_H \leq \delta.$$

Combining everything and using the fact that $\text{ang}(\boldsymbol{x}, \boldsymbol{x}'), \text{ang}(\boldsymbol{y}, \boldsymbol{y}') \leq \pi^{-1}\varepsilon$ (as angular distance is locally Lipschitz around 0 with respect to the ℓ_2 norm), under given conditions with probability $1 - \exp(-C\delta^2 m)$ we find that

$$\begin{aligned}|m^{-1}\|\boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y}\|_H - \text{ang}(\boldsymbol{x}, \boldsymbol{y})| \leq\ & |\text{ang}(\boldsymbol{x}, \boldsymbol{y}) - \text{ang}(\boldsymbol{x}', \boldsymbol{y})| + |\text{ang}(\boldsymbol{x}', \boldsymbol{y}) - \text{ang}(\boldsymbol{x}', \boldsymbol{y}')| \\ & + |m^{-1}\|\boldsymbol{A}\boldsymbol{x}', \boldsymbol{A}\boldsymbol{y}'\|_H - \text{ang}(\boldsymbol{x}', \boldsymbol{y}')| + m^{-1}\|\boldsymbol{A}\boldsymbol{x}', \boldsymbol{A}\boldsymbol{x}\|_H + m^{-1}\|\boldsymbol{A}\boldsymbol{y}', \boldsymbol{A}\boldsymbol{y}\|_H \\ & \leq 2\pi^{-1}\varepsilon + 2\delta + \delta \leq 4\delta.\end{aligned}$$

For the second line we used the fact that

$$\|\boldsymbol{x}, \boldsymbol{y}\|_H - \|\boldsymbol{x}', \boldsymbol{y}\|_H \leq \|\boldsymbol{x}, \boldsymbol{x}'\|_H.$$

Using the adjustment $\delta' \leftrightarrow 4\delta$ we obtain the desired continuous binary embedding result. The only additional constraint to the ones in Theorem 3.2 is the requirement $m \geq 2\delta^{-2}\log N_\varepsilon$ which is one of the assumptions of Theorem 2.2 thus we can conclude with the result. ∎

Observe that this result gives following bounds for δ-binary embedding.

- For an arbitrary K, $m \geq \omega^2(K)\delta^{-4}\log\delta^{-1}$ samples are sufficient. This yields the corresponding statement of Theorem 2.2.
- When $\hat{K}$ is a d dimensional subspace, $m \geq \omega^2(K)\delta^{-2}\log\delta^{-1}$ samples are sufficient.

More generally, one can plug improved covering bounds for the scenarios $\hat{K}$ is the set of d sparse vectors or set of rank d matrices to show that for these sets $m \geq \omega^2(K)\delta^{-2}\log\delta^{-1}$ is sufficient. A useful property of these sets are the fact that $K - K$ is still low-dimensional for instance if $\hat{K}$ is the set of d sparse vectors then the elements of $K - K$ are at most $2d$ sparse.

3.3.1 Proof of Theorem 2.6 via improved structured embeddings

So far our subspace embedding bound requires a sample complexity of $\mathcal{O}(d\delta^{-2}\log\delta^{-1})$ which is slightly suboptimal compared to the linear Johnson-Lindenstrauss embedding with respect to ℓ_2 norm. To correct this, we need a more advanced discrete embedding result which requires a more involved argument. In particular we shall use Theorem 4.1 of the next section which is essentially a stronger version of the straightforward result Lemma 3.1 when $\hat{K}$ is a structured set obeying (2.1).

Proof Create an $\varepsilon = \delta^{3/2}$ covering $\bar{K}_\varepsilon$ of K. In order for covering elements to satisfy the embedding bound (3.1) Theorem 4.1 requires $m \geq C\delta^{-2}C(K)$. Next we need to ensure that local deviation properties still hold. In particular given $\boldsymbol{x}', \boldsymbol{y}' \in \bar{K}_\varepsilon$ and $\boldsymbol{x}, \boldsymbol{y} \in K$ we still have $\text{ang}(\boldsymbol{x}, \boldsymbol{x}') \leq \pi^{-1}\varepsilon \leq \pi^{-1}\delta$ and for the rest we repeat the proof of Theorem 3.2 which ends up yielding the following conditions (after an application of Lemma 3.4)

$$c\varepsilon\delta\sqrt{\log\frac{1}{\delta}m} + \sqrt{m\delta}\omega(K_\varepsilon) \leq \frac{m\delta^2}{8},\ m \geq \frac{64}{\delta}\log N_\varepsilon$$

Both of these conditions trivially hold when we use the estimates (2.1), namely $\omega(K_\varepsilon) \leq \varepsilon C(K)$ and $\log N_\varepsilon \leq C(K)\log\varepsilon^{-1}$. ∎

4 Optimal embedding of structured sets

As we mentioned previously, naive bounds for subspaces require a sample complexity of $\mathcal{O}(d\delta^{-2}\log\delta^{-1})$ where d is the subspace dimension. On the other hand, for linear embedding it is known that the optimal dependence is $\mathcal{O}(d/\delta^2)$. We will show that it is in fact possible to achieve optimal dependence via a more involved argument based on "generic chaining" strategy. The main result of this section is summarized in the following theorem.

Theorem 4.1 *Suppose K satisfies the bounds (2.1) for all $\varepsilon > 0$. There exists constants $c, c_1, c_2 > 0$ and an $\varepsilon = c\delta^{3/2}$ covering $\bar{K}_\varepsilon$ of K such that if $m \geq c_1\delta^{-2}C(K)$, with probability $1 - 10\exp(-c_2\delta^2 m)$ all $\boldsymbol{x}, \boldsymbol{y} \in \bar{K}_\varepsilon$ obey*

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

Proof Let $\mathcal{C}_i$ be a $\frac{1}{2^i}$ ℓ_2-cover of K. From structured set assumption (2.1), for some constant $C_0 > 0$ the cardinality of the covering satisfies

$$\log|\mathcal{C}_i| \leq iC(K).$$

Consider covers $\mathcal{C}_i$ for $1 \leq i \leq N = \lceil\log_2\frac{1}{\varepsilon}\rceil$. This choice of N ensures that $\mathcal{C}_N$ is an ε cover.

Given points $\boldsymbol{x} = \boldsymbol{x}_N, \boldsymbol{y} = \boldsymbol{y}_N \subset \mathcal{C}_N$, find $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{N-1}, \boldsymbol{y}_1, \dots, \boldsymbol{y}_{N-1}$ in the covers that are closest to $\boldsymbol{x}_N, \boldsymbol{y}_N$ respectively. For notational simplicity, define

$$4d(\boldsymbol{x}, \boldsymbol{y}) = \|\text{sgn}(\boldsymbol{Ax}) - \text{sgn}(\boldsymbol{Ay})\|_{\ell_2}^2,$$
$$4d(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) = \langle\text{sgn}(\boldsymbol{Ax}) - \text{sgn}(\boldsymbol{Ay}), \text{sgn}(\boldsymbol{Ay}) - \text{sgn}(\boldsymbol{Az})\rangle,$$
$$4d(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{w}) = \langle\text{sgn}(\boldsymbol{Ax}) - \text{sgn}(\boldsymbol{Ay}), \text{sgn}(\boldsymbol{Az}) - \text{sgn}(\boldsymbol{Aw})\rangle.$$

Each of $d(\boldsymbol{x},\boldsymbol{y}), d(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}), d(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{w})$ are sum of m i.i.d. random variables that take values $+1,-1,0$. For instance consider $d(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z},\boldsymbol{w})$. In this case, the random variables are of the form

$$\boldsymbol{a} = \left\langle \text{sgn}(\boldsymbol{g}^T\boldsymbol{x}) - \text{sgn}(\boldsymbol{g}^T\boldsymbol{y}), \text{sgn}(\boldsymbol{g}^T\boldsymbol{z}) - \text{sgn}(\boldsymbol{g}^T\boldsymbol{w})\right\rangle, \text{ where } \boldsymbol{g}\sim\mathcal{N}(0,\mathbf{I}_n),$$

which is either $-1,0,1$. Furthermore, it is 0 as soon as either $\boldsymbol{x},\boldsymbol{y}$ or $\boldsymbol{z},\boldsymbol{w}$ induces the same sign which means $\mathbb{P}(\boldsymbol{a}\neq 0)\leq \min\{\text{ang}(\boldsymbol{x},\boldsymbol{y}),\text{ang}(\boldsymbol{z},\boldsymbol{w})\}$.

Given points $\{\boldsymbol{x}_N,\dots,\boldsymbol{x}_1,\boldsymbol{y}_1,\dots,\boldsymbol{y}_N\}$ we have that

$$\begin{aligned} d(\boldsymbol{x}_i,\boldsymbol{y}_i) =& d(\boldsymbol{x}_i,\boldsymbol{x}_{i-1}) + d(\boldsymbol{y}_i,\boldsymbol{y}_{i-1}) + 2d(\boldsymbol{x}_i,\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1}) + 2d(\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1},\boldsymbol{y}_i) + 2d(\boldsymbol{x}_i,\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1},\boldsymbol{y}_i), \\ &+ d(\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1}). \end{aligned} \tag{4.1}$$

The term $d(\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1})$ on the second line will be used for recursion. We will show that each of the remaining terms (first line) concentrate around their expectations. Since the argument is identical, to prevent repetitions, we will focus on $d(\boldsymbol{x}_i,\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1},\boldsymbol{y}_i)$.

Recall that $d(\boldsymbol{x}_i,\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1},\boldsymbol{y}_i)$ is sum of m i.i.d. random variables $\boldsymbol{a}_i$ that takes values $\{-1,0,1\}$ where $\boldsymbol{a}_i$ satisfies

$$\mathbb{P}(|\boldsymbol{a}_i|=1)\leq\min\{\text{ang}(\boldsymbol{x}_i,\boldsymbol{x}_{i-1}),\text{ang}(\boldsymbol{y}_i,\boldsymbol{y}_{i-1})\}\leq\frac{2}{2^i}\coloneqq 2\delta_i$$

as $\text{ang}(\boldsymbol{x},\boldsymbol{y})\leq\|\boldsymbol{x}-\boldsymbol{y}\|_{\ell_2}/2$. Assuming $\varepsilon_i\leq\delta_i$ (will be verified at (4.3)), for a particular quadruple $(\boldsymbol{x}_i,\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1},\boldsymbol{y}_i)$, applying Lemma 4.2 for $i\geq 4$ (which ensures $\mu/2\leq\mathbb{P}(|\boldsymbol{a}_i|=1)\leq 1/6$), we have that

$$\mathbb{P}(|d(\boldsymbol{x}_i,\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1},\boldsymbol{y}_i)-\mathbb{E}[d(\boldsymbol{x}_i,\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1},\boldsymbol{y}_i)]|\geq\varepsilon_i m)\leq 2\exp(-\frac{\varepsilon_i^2 m}{4\delta_i}). \tag{4.2}$$

Pick $\varepsilon_i=\sqrt{i}2^{-i/2}\delta$. Observe that, $\frac{\varepsilon_i}{\delta_i}$ can be bounded as

$$\frac{\varepsilon_i}{\delta_i}=\sqrt{i}\frac{2^{-i/2}}{2^{-i}}\delta=\sqrt{i}2^{i/2}\delta\leq\sqrt{N}2^{N/2}\delta\leq\sqrt{N}2^{N/2}2^{-2(N-1)/3}<1/4 \tag{4.3}$$

for ε (or δ) sufficiently small (which makes N large) where we used the fact that $2^{-(N-1)}\geq\varepsilon=\delta^{3/2}$. Consequently the bound (4.2) is applicable.

This choice of ε_i yields a failure probability of $2\exp(-\frac{\varepsilon_i^2 m}{4\delta_i})=2\exp(-i\delta^2 m/4)$. Union bounding (4.2) over all quadruples we find that the probability of success for all $(\boldsymbol{x}_i,\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1},\boldsymbol{y}_i)$ is at least

$$1-2\exp(-i\delta^2 m/4)\exp(4iC(K))\leq 1-2\exp(-i\delta^2 m/8)$$

under initial assumptions. The deviation of the other terms in (4.1) can be bounded with the identical argument. Define η_i to be

$$\eta_i=d(\boldsymbol{x}_i,\boldsymbol{x}_{i-1})+d(\boldsymbol{y}_i,\boldsymbol{y}_{i-1})+2d(\boldsymbol{x}_i,\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1})+2d(\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1},\boldsymbol{y}_i)+2d(\boldsymbol{x}_i,\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1},\boldsymbol{y}_i).$$

So far we showed that for all $\boldsymbol{x}_i,\boldsymbol{y}_i,\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1}$, with probability $1-10\exp(-i\delta^2 m/8)$

$$|\eta_i-\mathbb{E}[\eta_i]|\leq 5\varepsilon_i m=5\sqrt{i}2^{-i/2}\delta m.$$

Since $d(\boldsymbol{x}_i,\boldsymbol{y}_i)-d(\boldsymbol{x}_{i-1},\boldsymbol{y}_{i-1})=\eta_i$ applying a union bound over $4\leq i\leq N$ we find that

$$d(\boldsymbol{x}_N,\boldsymbol{y}_N)=\sum_{i=4}^N\eta_i+d(\boldsymbol{x}_3,\boldsymbol{y}_3) \tag{4.4}$$

We treat $d(\boldsymbol{x}_3,\boldsymbol{y}_3)$ specifically as Lemma 4.2 may not apply. In this case, the cardinality of the cover is small in particular $\log|\mathcal{C}_3|\leq k(\log C_0+3)+\log L$ hence, we simply use Lemma 3.1 to conclude for all $\boldsymbol{x}_3,\boldsymbol{y}_3$

$$|d(\boldsymbol{x}_3,\boldsymbol{y}_3)-\mathbb{E}[d(\boldsymbol{x}_3,\boldsymbol{y}_3)]|\leq\delta$$

with probability $1-\exp(-\delta^2 m)$. Merging our estimates, and using $\mathbb{E}[d(\boldsymbol{x}_N,\boldsymbol{y}_N)]=\sum_{i=4}^N \mathbb{E}[\eta_i]+\mathbb{E}[d(\boldsymbol{x}_3,\boldsymbol{y}_3)]$, we find

$$|d(\boldsymbol{x}_N,\boldsymbol{y}_N)-\mathbb{E}[d(\boldsymbol{x}_N,\boldsymbol{y}_N)]|\le\delta+\sum_{i=4}^N 5\sqrt{i}2^{-i/2}\delta\le c'\delta$$

for an absolute constant $c'>0$ with probability $1-10\sum_{i=1}^N\exp(-i\delta^2 m/8)\le 1-10\frac{\exp(-\delta^2 m/8)}{1-\exp(-\delta^2 m/8)}$. Clearly, our initial assumption allows us to pick $\delta^2 m\ge 16$ so that $1-\exp(-\delta^2 m/8)\ge 0.5$ which makes the probability of success $1-20\exp(-\delta^2 m/8)$. To obtain the advertised result simply use the change of variable $c'\delta\to\delta$. ■

The remarkable property of this theorem is the fact that we can fix the sample complexity to $\delta^{-2}C(K)$ while allowing the ε-net to get tighter as a function of the distortion δ. We should emphasize that $\varepsilon\sim\delta^{3/2}$ dependence can be further improved to $\varepsilon\sim\delta^{2-\alpha}$ for any $\alpha>0$. We only need to ensure that (4.3) is satisfied. The next result is a helper lemma for the proof of Theorem 4.1.

Lemma 4.2 *Let* $\{\boldsymbol{x}_i\}_{i=1}^m$ *be i.i.d. random variables taking values in* $\{-1,0,1\}$. *Suppose* $\max\{\mathbb{P}(\boldsymbol{x}_1=-1)=p_-,\mathbb{P}(\boldsymbol{x}_1=1)\}=p_+\}\le\mu/2$ *for some* $0\le\mu\le 1/3$. *Then, whenever* $\varepsilon\le\frac{\mu}{2}$

$$\mathbb{P}(|m^{-1}\sum_{i=1}^m\boldsymbol{x}_i-\mathbb{E}[m^{-1}\sum_{i=1}^m\boldsymbol{x}_i]|\le\varepsilon)\ge 1-2\exp(-\frac{\varepsilon^2 m}{4\mu}).$$

Proof Let us first estimate the number of nonzero components in $\{\boldsymbol{x}_i\}$. Using a standard Chernoff bound (e.g. Lemma A.1), for $\varepsilon<\mu/2$, we have that

$$\mathbb{P}(|m^{-1}\sum_{i=1}^m|\boldsymbol{x}_i|-(p_++p_-)|\le\varepsilon)\ge 1-\exp(-\frac{\varepsilon^2 m}{4\mu}).\tag{4.5}$$

Conditioned on $\sum_{i=1}^m|\boldsymbol{x}_i|=c\le(\mu+\varepsilon)m\le 1.5\mu m$, the c nonzero elements in $\boldsymbol{x}_i$ are $+1,-1$ with the normalized probabilities $\frac{p_+}{p_++p_-},\frac{p_-}{p_++p_-}$. Denoting these variables by $\{b_i\}_{i=1}^c$, and applying another Chernoff bound, we have that

$$\mathbb{P}(|\sum_{i=1}^c b_i-\mathbb{E}[\sum_{i=1}^c b_i]|\le\varepsilon_2 c)\ge 1-\exp(-2\varepsilon_2^2 c).$$

Picking $\varepsilon_2=\frac{\varepsilon m}{c}$, we obtain the bound

$$\mathbb{P}(|\sum_{i=1}^c b_i-\mathbb{E}[\sum_{i=1}^c b_i]|\le\varepsilon m)\ge 1-\exp(-2\varepsilon^2 m^2/c)\tag{4.6}$$

which shows that $|\sum_{i=1}^c b_i-\frac{p_+-p_-}{p_++p_-}c|\le\varepsilon m$ with probability $1-\exp(-2\varepsilon^2 m^2/c)$. Since $c\le 2\mu m$, $1-\exp(-2\varepsilon^2 m^2/c)\ge 1-\exp(-\varepsilon^2 m/\mu)$. Finally observe that

$$|\frac{p_+-p_-}{p_++p_-}c-(p_+-p_-)m|\le\frac{|p_+-p_-|}{p_++p_-}\varepsilon m\le\varepsilon m\implies|\sum_{i=1}^c b_i-(p_+-p_-)m|\le 2\varepsilon m.$$

To conclude recall that $\sum_{i=1}^c b_i=\sum_{i=1}^m\boldsymbol{x}_i$ and $(p_+-p_-)m=\mathbb{E}[\sum_{i=1}^m\boldsymbol{x}_i]$, then apply a union bound combining (4.5) and (4.6). ■

5 Local properties of binary embedding

For certain applications such as locality sensitive hashing, we are more interested with the local behavior of embedding, i.e. what happens to the points that are around each other. In this case, instead of preserving the distance, we can ask for $\text{sgn}(\boldsymbol{A}\boldsymbol{x}),\text{sgn}(\boldsymbol{A}\boldsymbol{y})$ to be close if and only if $\boldsymbol{x},\boldsymbol{y}$ are close.

The next theorem is a restatement of the second statement of Theorem 2.5 and summarizes our result on local embedding.

Theorem 5.1 *Given* $0 < \delta < 1$, *set* $\varepsilon = c\delta/\sqrt{\log \delta^{-1}}$. *There exists* $c, c_1, c_2 > 0$ *such that if*

$$m \geq c_1 \max\{\frac{1}{\delta} \log N_\varepsilon, \delta^{-3}\omega^2(K_\varepsilon)\}.$$

with probability $1 - \exp(-c_2\delta m)$, *the following statements hold.*

- *For all* $\boldsymbol{x}, \boldsymbol{y} \in K$ *satisfying* $ang(\boldsymbol{x}, \boldsymbol{y}) \leq \varepsilon$, *we have* $m^{-1}\|sgn(\boldsymbol{Ax}), sgn(\boldsymbol{Ay})\|_H \leq \delta$.
- *For all* $\boldsymbol{x}, \boldsymbol{y} \in K$ *satisfying* $m^{-1}\|sgn(\boldsymbol{Ax}), sgn(\boldsymbol{Ay})\|_H \leq \delta/32$, *we have* $ang(\boldsymbol{x}, \boldsymbol{y}) \leq \delta$.

Proof First statement is already proved by Theorem 3.2. For the second statement, we shall follow a similar argument. Suppose that for some pair $\boldsymbol{x}, \boldsymbol{y} \in K$ obeying $\text{ang}(\boldsymbol{x}, \boldsymbol{y}) > \delta$ we have $\|\text{sgn}(\boldsymbol{Ax}), \text{sgn}(\boldsymbol{Ay})\|_H \leq \delta/32$.

Consider an ε covering $\bar{K}_\varepsilon$ of K where $\varepsilon = c\delta/\sqrt{\log \delta^{-1}}$ for a sufficiently small $c > 0$. Let $\boldsymbol{x}', \boldsymbol{y}'$ be the elements of the cover obeying $\|\boldsymbol{x}' - \boldsymbol{x}\|_{\ell_2}, \|\boldsymbol{y}' - \boldsymbol{y}\|_{\ell_2} \leq \varepsilon$ which also ensures the angular distance to be at most $\pi^{-1}\varepsilon$. Applying Theorem 3.2 under the stated conditions we can ensure that

$$m^{-1}\|\boldsymbol{Ax}, \boldsymbol{Ax}'\|_H, m^{-1}\|\boldsymbol{Ay}, \boldsymbol{Ay}'\|_H \leq \delta/20.$$

Next, since ε can be made arbitrarily smaller than δ we can guarantee that $\text{ang}(\boldsymbol{x}', \boldsymbol{y}') \geq \delta/2$. We shall apply a Chernoff bound over the elements of the cover to ensure that $\|\boldsymbol{Ax}' - \boldsymbol{Ay}'\|_H$ is significant for all $\boldsymbol{x}', \boldsymbol{y}' \in \bar{K}_\varepsilon$. The following version of Chernoff gives the desired bound.

Lemma 5.2 *Suppose* $\{a_i\}_{i=1}^m$ *are i.i.d. Bernoulli random variables satisfying* $\mathbb{E}[a_i] \geq \delta/2$. *Then, with probability* $1 - \exp(-\delta m/16)$ *we have that* $\sum_{i=1}^m a_i \geq \delta m/4$.

Applying this lemma to all pairs of the cover, whenever $m \geq 128\delta^{-1} \log N_\varepsilon$ we find that $\|\boldsymbol{Ax}' - \boldsymbol{Ay}'\|_H \geq \delta m/4$.

We can now use the triangle inequality to achieve

$$|m^{-1}\|\boldsymbol{Ax}, \boldsymbol{Ay}\|_H - \text{ang}(\boldsymbol{x}, \boldsymbol{y})| \geq |m^{-1}\|\boldsymbol{Ax}', \boldsymbol{Ay}'\|_H - \text{ang}(\boldsymbol{x}', \boldsymbol{y}')| \tag{5.1}$$

$$- [|\text{ang}(\boldsymbol{x}, \boldsymbol{y}) - \text{ang}(\boldsymbol{x}', \boldsymbol{y})| + |\text{ang}(\boldsymbol{x}', \boldsymbol{y}) - \text{ang}(\boldsymbol{x}', \boldsymbol{y}')|] \tag{5.2}$$

$$- [|m^{-1}\|\boldsymbol{Ax}, \boldsymbol{Ay}\|_H - m^{-1}\|\boldsymbol{Ax}', \boldsymbol{Ay}\|_H| + |m^{-1}\|\boldsymbol{Ax}', \boldsymbol{Ay}\|_H - m^{-1}\|\boldsymbol{Ax}', \boldsymbol{Ay}'\|_H|] \tag{5.3}$$

(5.1) is at least $\delta/4$, (5.2) is at most $c_0\varepsilon$ and (5.3) is at most $\delta/10$ which ensures that $|m^{-1}\|\boldsymbol{Ax}, \boldsymbol{Ay}\|_H - \text{ang}(\boldsymbol{x}, \boldsymbol{y})| \geq \delta/8$ by picking c small enough. This contradicts with the initial assumption. To obtain this contradiction, we required m to be $m \geq c_1 \max\{\delta^{-1} \log N_\varepsilon, \delta^{-3}\omega^2(K_\varepsilon)\}$ for some constant $c_1 > 0$. ■

6 Sketching for binary embedding

In this section, we discuss preprocessing the binary embedding procedure with a linear embedding. Our goal is to achieve an initial dimensionality reduction that preserves the ℓ_2 distances of K with a linear map and then using binary embedding to achieve reasonable guarantees. In particular we will prove that this scheme works almost as well as the Gaussian binary embedding we discussed so far.

For the sake of this section, $\boldsymbol{B} \in \mathbb{R}^{m \times m_{lin}}$ denotes the binary embedding matrix and $\boldsymbol{F} \in \mathbb{R}^{m_{lin} \times n}$ denotes the preprocessing matrix that provides an initial sketch of the data. The overall sketched binary embedding is given by $\boldsymbol{x} \to \text{sgn}(\boldsymbol{Ax})$ where $\boldsymbol{A} = \boldsymbol{BF} \in \mathbb{R}^{m \times n}$. We now provide a brief background on linear embedding.

6.1 Background on Linear Embedding

For a mapping to be a linear embedding, we require it to preserve distances and lengths of the sets.

Definition 6.1 (δ-linear embedding) *Given* $\delta \in (0, 1)$, $\boldsymbol{F}$ *is a* δ*-embedding of the set* $\mathcal{C}$ *if for all* $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{C}$, *we have that*

$$|\,\|\boldsymbol{Fx} - \boldsymbol{Fy}\|_{\ell_2} - \|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2}| \leq \delta, \; |\,\|\boldsymbol{Ax}\|_{\ell_2} - \|\boldsymbol{x}\|_{\ell_2}| \leq \delta. \tag{6.1}$$

Observe that if $\mathcal{C}$ is a subset of the unit sphere, this definition preserves the length of the vectors multiplicatively i.e. obeys $|\|\boldsymbol{A}\boldsymbol{x}\| - \|\boldsymbol{x}\|| \leq \delta\|\boldsymbol{x}\|$. We should point out that more traditional embedding results ask for

$$| \, \|\boldsymbol{F}\boldsymbol{x} - \boldsymbol{F}\boldsymbol{y}\|_{\ell_2}^2 - \|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2}^2 \, | \leq \delta.$$

This condition can be weaker than what we state as it allows for $| \, \|\boldsymbol{F}\boldsymbol{x} - \boldsymbol{F}\boldsymbol{y}\|_{\ell_2} - \|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2} | \sim \sqrt{\delta}$ for small values of $\|\boldsymbol{x} - \boldsymbol{y}\|_{\ell_2}$. However, Gaussian matrices allow for δ-linear embedding with optimal dependencies. The reader is referred to Lemma 6.8 of [18] for a proof.

Theorem 6.2 *Suppose $\boldsymbol{F} \in \mathbb{R}^{m_{lin} \times n}$ is a standard Gaussian matrix normalized by $\sqrt{m_{lin}}$. For any set $\mathcal{C} \in \mathcal{B}^{n-1}$ whenever $\sqrt{m_{lin}} \geq \omega(\mathcal{C}) + \eta + 1$, with probability $1 - \exp(-\eta^2/8)$ we have that*

$$\sup_{\boldsymbol{x} \in \mathcal{C}} | \, \|\boldsymbol{F}\boldsymbol{x}\|_{\ell_2} - \|\boldsymbol{x}\|_{\ell_2} | \leq m_{lin}^{-1/2}(\omega(\mathcal{C}) + \eta).$$

To achieve a δ-linear embedding we can apply this to the sets $K - K$ and K by setting $m_{lin} = \mathcal{O}(\delta^{-2}(\omega(\mathcal{C}) + \eta)^2)$.

Similar results can be obtained for other matrix ensembles including Fast Johnson-Lindenstrauss Transform [17]. A Fast JL transform is a random matrix $\boldsymbol{F} = \boldsymbol{S}\boldsymbol{D}\boldsymbol{R}$ where $\boldsymbol{S} \in \mathbb{R}^{m_{lin} \times n}$ randomly subsamples m rows of a matrix, $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ is the normalized Hadamard transform, and $\boldsymbol{R}$ is a diagonal matrix with independent Rademacher entries. The recent results of [17] shows that linear embedding of arbitrary sets via FJLT is possible. The following corollary follows from their result by using the fact that $| \, \|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2}^2 - \|\boldsymbol{x}\|_{\ell_2}^2 | \leq \delta^2 \implies | \, \|\boldsymbol{A}\boldsymbol{x}\|_{\ell_2} - \|\boldsymbol{x}\|_{\ell_2} | \leq \delta$.

Corollary 6.3 *Suppose $K \subset \mathcal{S}^{n-1}$. Suppose $m_{lin} \geq c(1+\eta)^2\delta^{-4}(\log n)^4\omega^2(K)$ and $\boldsymbol{F} \in \mathbb{R}^{m_{lin} \times n}$ is an FJLT. Then, with probability $1 - \exp(-\eta)$, $\boldsymbol{F}$ is a δ-linear embedding for K.*

Unlike Theorem 6.2, this corollary yields $d \sim \delta^{-4}$ instead of δ^{-2}. It would be of interest to improve such distortion bounds.

6.2 Results on sketched binary embedding

With these technical tools, we are in a position to state our results on binary embedding.

Theorem 6.4 *Suppose $\boldsymbol{F} \in \mathbb{R}^{m_{lin} \times m}$ is a δ-linear embedding for the set $K \cup -K$. Then*

- *$\boldsymbol{B}\boldsymbol{F}$ is a $c\delta$-binary embedding of K if $\boldsymbol{B}$ is a δ-binary embedding for $\mathcal{S}^{m_{lin}-1}$.*
- *Suppose $\hat{K}$ is union of L d-dimensional subspaces, then $\boldsymbol{B}\boldsymbol{F}$ is a $c\delta$-binary embedding with probability α if $\boldsymbol{B}$ is a δ-binary embedding for union of L d-dimensional subspaces with the same probability.*

Proof Given $\boldsymbol{x}, \boldsymbol{y} \in K$ define $\boldsymbol{x}_F = \frac{\boldsymbol{F}\boldsymbol{x}}{\|\boldsymbol{F}\boldsymbol{x}\|}$, $\boldsymbol{y}_F = \frac{\boldsymbol{F}\boldsymbol{y}}{\|\boldsymbol{F}\boldsymbol{y}\|}$ where $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{B}\boldsymbol{F}\boldsymbol{x}$. Since $\boldsymbol{F}$ is a δ-linear embedding for K, for all $\boldsymbol{x}, \boldsymbol{y}$ the following statements hold.

- $\max\{|\|\boldsymbol{F}\boldsymbol{x}\| - \|\boldsymbol{x}\||, |\|\boldsymbol{F}\boldsymbol{y}\| - \|\boldsymbol{y}\||, |\|\boldsymbol{F}\boldsymbol{x} - \boldsymbol{F}\boldsymbol{y}\| - \|\boldsymbol{x} - \boldsymbol{y}\||\} \leq \delta$.
- $|\|\boldsymbol{y}_F - \boldsymbol{x}_F\| - \|\boldsymbol{F}\boldsymbol{x} - \boldsymbol{F}\boldsymbol{y}\|| \leq \|\boldsymbol{F}\boldsymbol{x} - \boldsymbol{x}_F\| + \|\boldsymbol{F}\boldsymbol{y} - \boldsymbol{y}_F\| \leq 2\delta$.

These imply $|\|\boldsymbol{y}_F - \boldsymbol{x}_F\| - \|\boldsymbol{x} - \boldsymbol{y}\|| \leq 3\delta$ which in turn implies $|\text{ang}(\boldsymbol{x}_F, \boldsymbol{y}_F) - \text{ang}(\boldsymbol{x}, \boldsymbol{y})| \leq c\delta$ using standard arguments in particular Lipschitzness of the geodesic distance between 0 and $\frac{\pi}{2}$. We may assume the angle to be between 0 and $\pi/2$ as the set $K \cup -K$ is symmetric and if $\text{ang}(\boldsymbol{x}, \boldsymbol{y}) > \pi/2$ we may consider $\text{ang}(\boldsymbol{x}, -\boldsymbol{y}) < \pi/2$.

If $\boldsymbol{B}$ is a δ binary embedding it implies that $|\text{ang}(\boldsymbol{x}_F, \boldsymbol{y}_F) - m^{-1}\|\text{sgn}(\boldsymbol{A}\boldsymbol{x}), \text{sgn}(\boldsymbol{A}\boldsymbol{y})\|_H| \leq \delta$. In turn, this yields $|\text{ang}(\boldsymbol{x}, \boldsymbol{y}) - m^{-1}\|\text{sgn}(\boldsymbol{A}\boldsymbol{x}), \text{sgn}(\boldsymbol{A}\boldsymbol{y})\|_H| \leq (c+1)\delta$ and overall map is $(c+1)\delta$-binary embedding.

For the second statement observe that if $\hat{K}$ is a union of subspaces so is $\text{cone}(\boldsymbol{F}K)$ and we are given that $\boldsymbol{B}$ is a δ binary embedding for union of subspaces with α probability. This implies that $|\text{ang}(\boldsymbol{x}_F, \boldsymbol{y}_F) -$

$\|\text{sgn}(\boldsymbol{B}\boldsymbol{x}_F), \text{sgn}(\boldsymbol{B}\boldsymbol{y}_F)\|_H| \le \delta$ with probability α and we again conclude with $|\text{ang}(\boldsymbol{x},\boldsymbol{y}) - m^{-1}\|\text{sgn}(\boldsymbol{A}\boldsymbol{x}), \text{sgn}(\boldsymbol{A}\boldsymbol{y})\|_H| \le (c+1)\delta$. ∎

Our next result obtains a sample complexity bound for sketched binary embedding.

Theorem 6.5 *There exists constants $c, C > 0$ such that if $\boldsymbol{B} \sim \mathbb{R}^{m \times m_{lin}}$ is a standard Gaussian matrix satisfying $m \ge c\delta^{-2}m_{lin}$ and*

- *if $\boldsymbol{F}$ is a standard Gaussian matrix normalized by $\sqrt{m_{lin}}$ where $m_{lin} > c\delta^{-2}\omega^2(K)$, with probability $1 - \exp(-C\delta^2 m) - \exp(-C\delta^2 m_{lin})$ $\boldsymbol{x} \to sgn(\boldsymbol{A}\boldsymbol{x})$ for $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{F}$ is a δ-binary embedding of K.*

- *if $\boldsymbol{F}$ is a Fast Johnson-Lindenstrauss Transform where $d > c(1+\eta)^2\delta^{-4}(\log n)^4\omega^2(K)$, with probability $1 - \exp(-\eta) - \exp(-C\delta^2 m)$ $\boldsymbol{x} \to sgn(\boldsymbol{A}\boldsymbol{x})$ for $\boldsymbol{A} = \boldsymbol{B}\boldsymbol{F}$ is a δ-binary embedding of K.*

Observe that for Gaussian embedding distortion dependence is $\delta^{-4}\omega^2(K)$ which is better than what can be obtained via Theorem 2.2. This is due to the fact that here we made use of the improved embedding result for subspaces. On the other hand distortion dependence for FJLT is δ^{-6} which we believe to be an artifact of Corollary 6.3. We remark that better dependencies can be obtained for subspaces.

Proof The proof makes use of the fact that $\boldsymbol{B}$ is a δ binary embedding for $\mathbb{R}^{m_{lin}}$ with the desired probability. Consequently, following Theorem 6.4, we simply need to ensure $\boldsymbol{F}$ is a δ linear embedding. When $\boldsymbol{F}$ is Gaussian this follows from Theorem 6.2. When $\boldsymbol{F}$ is FJLT, it follows from the bound Corollary 6.3 namely $m_{lin} \ge c(1+\eta)^2\delta^{-4}(\log n)^4\omega^2(K)$. ∎

6.3 Computational aspects

Consider the problem of binary embedding a d-dimensional subspace $\hat{K}$ with δ distortion. Using a standard Gaussian matrix as our embedding strategy requires $\mathcal{O}(\delta^{-2}nd)$ operation per embedding. This follows from the $\mathcal{O}(mn)$ operation complexity of matrix multiplication where $m = \mathcal{O}(\delta^{-2}d)$.

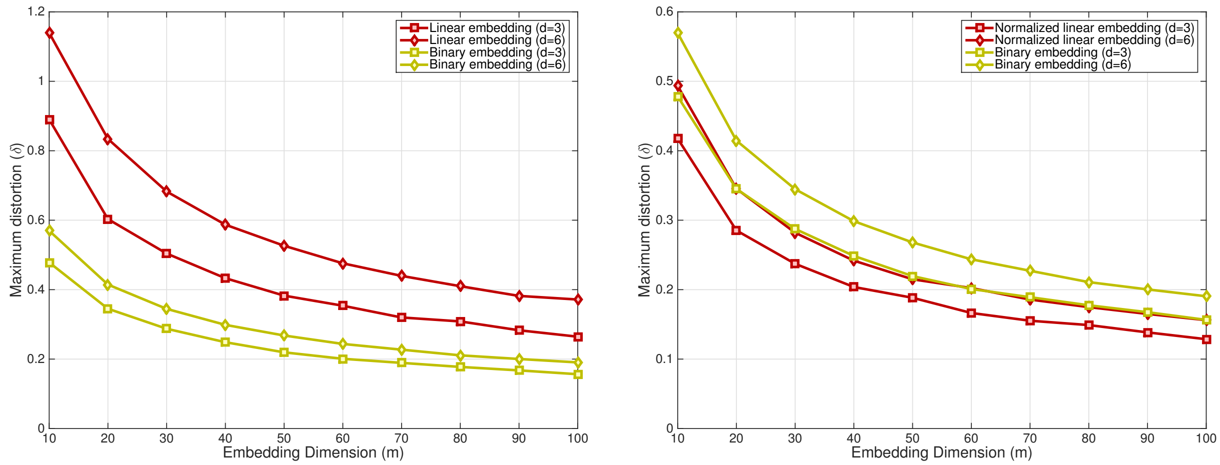
For FJLT sketched binary embedding, setting $m_{lin} = \mathcal{O}(\delta^{-4}(\log n)^4 d)$ and $m = \mathcal{O}(\delta^{-2}d)$ we find that each vector can be embedded in $\mathcal{O}(n\log n + m_{lin}m) \approx \mathcal{O}(n + \delta^{-6}d^2)$ up to logarithmic factors. Consequently, in the regime $d = \mathcal{O}(\sqrt{n})$, sketched embedding strategy is significantly more efficient and embedding can be done in near linear time. The similar pattern arises when we wish to embed an arbitrary set K. The main difference will be the distortion dependence in the computation bounds. We omit this discussion to prevent repetition.

These tradeoffs are in similar flavor to the recent work by Yi et al. [29] that apply to the embedding of finite sets. Here we show that similar tradeoffs can be extended to the case where K is arbitrary.

We shall remark that a faster binary embedding procedure is possible in practice via FJLT. In particular pick $\boldsymbol{A} = \boldsymbol{S}\boldsymbol{D}\boldsymbol{G}\boldsymbol{D}^*\boldsymbol{R}$ where $\boldsymbol{S}$ is the subsampling operator, $\boldsymbol{D}$ is the Hadamard matrix and $\boldsymbol{G}$ and $\boldsymbol{R}$ are diagonal matrices with independent standard normal and Rademacher nonzero entries respectively [15,30]. In this case, it is trivial to verify that for a given pair of $\boldsymbol{x},\boldsymbol{y}$ we have the identity $\text{ang}(\boldsymbol{x},\boldsymbol{y}) = m^{-1}\,\mathbb{E}\,\|\boldsymbol{A}\boldsymbol{x},\boldsymbol{A}\boldsymbol{y}\|_H$. While expectation trivially holds analysis of this map is more challenging and there is no significant theoretical result to the best of our knowledge. This map is beneficial since using FJLT followed by dense Gaussian embedding results in superlinear computation in n as soon as $d \ge \mathcal{O}(\sqrt{n})$. On the other hand, this version of Fast Binary Embedding has near-linear embedding time due to diagonal multiplications. Consequently, it would be very interesting to have a guarantee for this procedure when subspace $\hat{K}$ is in the nontrivial regime $d = \Omega(\sqrt{n})$. Investigation of this map for both discrete and continuous embedding remains as an open future direction.

7 Numerical experiments

Next, we shall list our numerical observations on binary embedding. A computational difficulty of binary embedding for continuous sets is the fact that it is nontrivial to obtain distortion bounds. In particular given



(a) Binary vs linear embedding

(b) Binary vs normalized linear embedding.

Figure 1: (a) Comparison of binary and linear embedding for a Gaussian map as a function of subspace dimension and sample complexity. (b) We keep the same setup however use a normalized distortion function. With this change, linear and binary embedding becomes more comparable.

set K and map $\boldsymbol{A}$ we would like to quantify the maximum distortion given by

$$\delta_{bin} = \sup_{\boldsymbol{x},\boldsymbol{y}\in K} |m^{-1}\|\boldsymbol{A}\boldsymbol{x},\boldsymbol{A}\boldsymbol{y}\|_H - \text{ang}(\boldsymbol{x},\boldsymbol{y})|.$$

To the best of our knowledge there is no guaranteed way of finding this supremum. For instance, for linear embedding we are interested in the bound

$$\delta_{lin} = \sup_{\boldsymbol{x},\boldsymbol{y}\in K} |m^{-1}\,\|\boldsymbol{A}\boldsymbol{x},\boldsymbol{A}\boldsymbol{y}\|_{\ell_2} - \|\boldsymbol{x},\boldsymbol{y}\|_{\ell_2}|$$

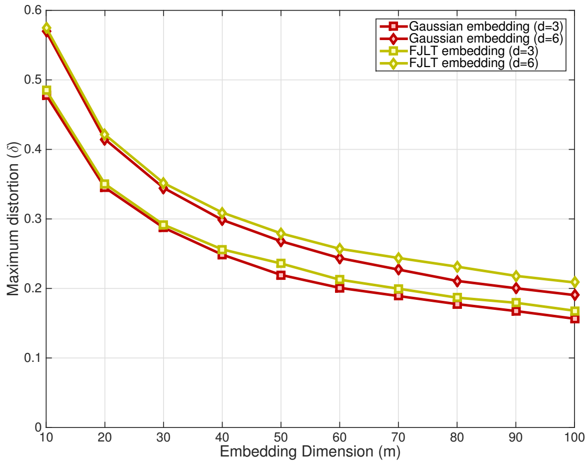
which can be obtained by calculating minimum and maximum singular values of $\boldsymbol{A}$ restricted to K. When K is a subspace, this can be done efficiently by studying the matrix obtained by projecting rows of $\boldsymbol{A}$ onto K.

To characterize the impact of set size on distortion, we sample 200 points from a d-dimensional subspace for $d = 3$ and $d = 6$ where $n = 128$. Clearly sampling finite number of points is not sufficient to capture the whole space however it is a good enough proxy for illustrating the impact of subspace dimension. We additionally vary the number of samples m between 0 and 100. Figure 1a contrasts linear and binary embedding schemes where $\sqrt{m}\boldsymbol{A}$ is a standard Gaussian matrix. We confirm that sampling the points from larger subspace indeed results in a larger distortion for both cases. Interestingly we observe that distortion for binary embedding is smaller than linear. This is essentially due to the fact that cost functions are not comparable rather than their actual performance. Clearly linear embedding stores more information about the signal so we expect it to be more beneficial.

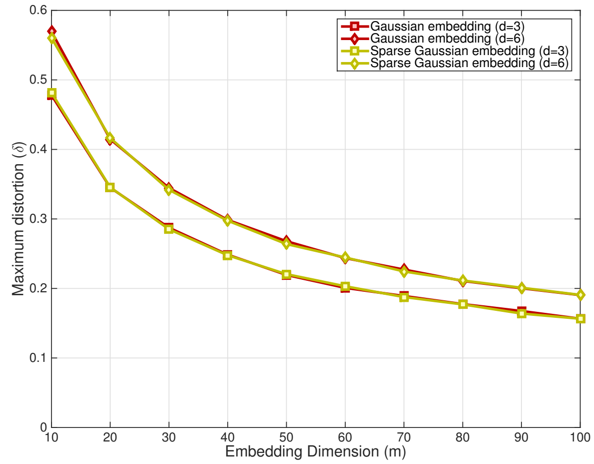
For a better comparison, we normalize the linear distortion function with respect to binary distortion so that if $\|\boldsymbol{A}\boldsymbol{x},\boldsymbol{A}\boldsymbol{y}\|_{\ell_2} = \|\boldsymbol{A}\boldsymbol{x},\boldsymbol{A}\boldsymbol{y}\|_H = 0$ we have δ_{bin} is same as normalized distortion δ_{n-lin}. This corresponds to the function

$$\delta_{n-lin} = \frac{\text{ang}(\boldsymbol{x},\boldsymbol{y})}{\|\boldsymbol{x},\boldsymbol{y}\|_{\ell_2}} \sup_{\boldsymbol{x},\boldsymbol{y}\in K} |m^{-1}\,\|\boldsymbol{A}\boldsymbol{x},\boldsymbol{A}\boldsymbol{y}\|_{\ell_2} - \|\boldsymbol{x},\boldsymbol{y}\|_{\ell_2}|.$$

Figure 1b shows the comparison of δ_{bin} and δ_{n-lin}. We observe that linear embedding results in lower distortion but their behavior is highly similar. Figure 1 shows that linear and binary embedding performs



(a) Gaussian vs FJLT for binary embedding.



(b) Gaussian vs sparse Gaussian for binary embedding.

on par and linear embedding does not provide a significant advantage. This is consistent with the main message of this work.

In Figure 2a we compare Gaussian embedding with fast binary embedding given by $\boldsymbol{A} = \boldsymbol{SDGD^*R}$ as described in Section 6.3. We observe that Gaussian yields slightly better bounds however both techniques perform on par in all regimes. This further motivates theoretical understanding of fast binary embedding which significantly lags behind linear embedding. Sparse matrices are another strong alternative for fast multiplication and efficient embedding [8]. Figure 2b contrasts Gaussian embedding with sparse Gaussian where the entries are 0 with probability 2/3. Remarkably, the distortion dependence perfectly matches. Following from this example, it would be interesting to study the class of matrices that has the same empirical behavior as a Gaussian. For linear embedding this question has been studied extensively [9, 19] and it remains as an open problem whether results of similar flavor would hold for binary embedding.

Acknowledgments

The authors would like to thank Mahdi Soltanolkotabi for helpful conversations. BR is generously supported by ONR awards N00014-11-1-0723 and N00014-13-1-0129, NSF awards CCF-1148243 and CCF-1217058, AFOSR award FA9550-13-1-0138, and a Sloan Research Fellowship. SO was generously supported by the Simons Institute for the Theory of Computing and NSF award CCF-1217058.

References

[1] Albert Ai, Alex Lapanowski, Yaniv Plan, and Roman Vershynin. One-bit compressed sensing with non-gaussian measurements. *Linear Algebra and its Applications*, 441:222–239, 2014.

[2] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 459–468. IEEE, 2006.

[3] 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.

[4] Jean Bourgain and Jelani Nelson. Toward a unified theory of sparse dimensionality reduction in euclidean space. *arXiv preprint arXiv:1311.2542*, 2013.

[5] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359, 2011.

[6] Moses S Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388. ACM, 2002.

[7] Michael Cohen, Sam Elder, Cameron Musco, Christopher Musco, and Madalina Persu. Dimensionality reduction for k-means clustering and low rank approximation. *arXiv preprint arXiv:1410.6801*, 2014.

[8] Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós. A sparse johnson: Lindenstrauss transform. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 341–350. ACM, 2010.

[9] David Donoho and Jared Tanner. Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293, 2009.

[10] Alex Gittens and Michael W Mahoney. Revisiting the nystrom method for improved large-scale machine learning. *arXiv preprint arXiv:1303.1849*, 2013.

[11] Yehoram Gordon. *On Milman's inequality and random subspaces which escape through a mesh in* $\mathbb{R}^n$. Springer, 1988.

[12] Laurent Jacques. Small width, low distortions: quasi-isometric embeddings with quantized sub-gaussian random projections. *arXiv preprint arXiv:1504.06170*, 2015.

[13] 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.

[14] Eyal Kushilevitz, Rafail Ostrovsky, and Yuval Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. *SIAM Journal on Computing*, 30(2):457–474, 2000.

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

[16] Shahar Mendelson, Alain Pajor, and Nicole Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, 2008.

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

[18] Samet Oymak, Benjamin Recht, and Mahdi Soltanolkotabi. Sharp time–data tradeoffs for linear inverse problems. *arXiv preprint arXiv:1507.04793*, 2015.

[19] Samet Oymak and Joel A Tropp. Universality laws for randomized dimension reduction, with applications. *arXiv preprint arXiv:1511.09433*, 2015.

[20] 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.

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

[22] Yaniv Plan, Roman Vershynin, and Elena Yudovina. High-dimensional estimation with geometric constraints. *arXiv preprint arXiv:1404.3749*, 2014.

[23] Mark Rudelson and Roman Vershynin. On sparse reconstruction from fourier and gaussian measurements. *Communications on Pure and Applied Mathematics*, 61(8):1025–1045, 2008.

[24] Chris Thrampoulidis, Samet Oymak, and Babak Hassibi. A tight version of the gaussian min-max theorem in the presence of convexity. *in preparation*, 2014.

[25] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. The lasso with non-linear measurements is equivalent to one with linear measurements. *arXiv preprint arXiv:1506.02181*, 2015.

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

[27] Jingdong Wang, Heng Tao Shen, Jingkuan Song, and Jianqiu Ji. Hashing for similarity search: A survey. *arXiv preprint arXiv:1408.2927*, 2014.

[28] David P Woodruff. Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*, 2014.

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

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

A Supplementary results

Lemma A.1 *Suppose $\{\boldsymbol{a}_i\}_{i=1}^n$ are Bernoulli-q random variables where $0 \le q \le p$. Suppose $p < 1/3$. For any $\varepsilon \le \frac{p}{2}$, we have that*

$$\mathbb{P}(|\frac{\sum_i \boldsymbol{a}_i}{n} - q| > \varepsilon) \le \exp(-\frac{\varepsilon^2 n}{4p}).$$

Proof For any $\varepsilon > 0$, standard Chenoff bound yields that

$$\mathbb{P}(|\frac{\sum_i \boldsymbol{a}_i}{n} - q| > \varepsilon) \le \exp(-nD(q+\varepsilon\|q)) + \exp(-nD(q-\varepsilon\|q))$$

where $D(\cdot)$ is the KL-divergence. Applying Lemma A.2, we have that

$$\exp(-D(q+\varepsilon\|q)) \le \exp(-D(p+\varepsilon\|p)), \;\; \exp(-D(q-\varepsilon\|q)) \le \exp(-D(p-\varepsilon\|p)).$$

Finally, when $\frac{|\varepsilon|}{p} \le \frac{1}{2}$, we make use of the fact that $D(p+\varepsilon\|p) \ge \frac{\varepsilon^2 n}{4p}$. ■

Lemma A.2 *Given nonnegative numbers p_1, p_2, q_1, q_2, suppose $0.5 > p_1 > p_2$, $d = p_1 - q_1 > 0$, and $p_1 - p_2 = q_1 - q_2$. Then*

$$D(q_1\|q_2) \ge D(p_1\|p_2), \;\; D(q_2\|q_1) \ge D(p_2\|p_1).$$

Proof For $p_2 < p_1 \le 1/2$, using the definition of KL divergence, we have that

$$D(p_1\|p_2) = \int_{p_2}^{p_1} \frac{p_1 - x}{x(1-x)} dx = \int_{q_2}^{q_1} \frac{p_1 - (x+d)}{(x+d)(1-(x+d))} dx = \int_{q_2}^{q_1} \frac{q_1 - x}{(x+d)(1-(x+d))} dx.$$

Now, observe that for $x + d < 1/2$

$$(x+d)(1-(x+d)) \geq x(1-x) \implies \frac{q_1 - x}{(x+d)(1-(x+d))} \leq \frac{q_1 - x}{x(1-x)},$$

which yields $D(p_1\|p_2) \leq D(q_1\|q_2)$. The same argument can be repeated for $D(p_2\|p_1)$. We have that

$$D(p_2\|p_1) = \int_{1-p_1}^{1-p_2} \frac{1-p_2-x}{x(1-x)}dx = \int_{1-q_1}^{1-q_2} \frac{1-p_2-(x-d)}{(x-d)(1-(x-d))}dx = \int_{1-q_1}^{1-q_2} \frac{1-q_2-x}{(x-d)(1-(x-d))}dx.$$

This time, we make use of the inequality $(x-d)(1-(x-d)) \geq x(1-x)$ for $x - d \geq 1/2$ to conclude. ∎

Lemma A.3 *Let $Q(a) = \mathbb{P}(|g| \geq a)$ for $g \sim \mathcal{N}(0,1)$. There exists constants $c_1, c_2 > 0$ such that for $\delta > c_1$ we have that*

$$\int_{Q^{-1}(\delta)}^{\infty} t\sqrt{2/\pi}\exp(-t^2/2)dt \leq c_2\delta\sqrt{\log\delta^{-1}}.$$

Proof Set $\gamma = Q^{-1}(\delta)$. Observe that

$$\sqrt{2/\pi}\int_{\gamma}^{\infty} t\exp(-\frac{t^2}{2})dt = \sqrt{2/\pi}\exp(-\gamma^2/2). \tag{A.1}$$

We need to lower bound the number $\gamma > 0$. Choose δ sufficiently small to ensure $\gamma > 1$. Using standard lower and upper bounds on the Q function,

$$\frac{1}{\sqrt{2\pi}}\frac{1}{\gamma}\exp(-\frac{\gamma^2}{2}) \geq Q(\gamma) = \delta \geq \frac{1}{2\sqrt{2\pi}}\frac{1}{\gamma}\exp(-\frac{\gamma^2}{2}).$$

The left hand side implies that $\gamma \leq \sqrt{2\log 1/\delta}$ as otherwise the left-hand side would be strictly less than δ. Now, using the right hand side

$$\delta\sqrt{2\log 1/\delta} \geq \delta\gamma \geq \frac{1}{2\sqrt{2\pi}}\exp(-\frac{\gamma^2}{2}).$$

This provides the desired upper bound $\exp(-\frac{\gamma^2}{2}) \leq C\delta\sqrt{\log\frac{1}{\delta}}$ for $C = 4\sqrt{\pi}$. Combining this with (A.1), we can conclude. ∎

The following lemma is related to the order statistics of a standard Gaussian vector.

Lemma A.4 *Consider the setup of Lemma 3.5. There exists constants c_1, c_2, c_3 such that for any $\delta > c_1$, we have that*

$$\mathbb{E}[\sum_{i=1}^{\delta n}\tilde{\boldsymbol{g}}_i] \leq c_2\delta n\sqrt{\log\delta^{-1}}$$

whenever $n > c_3\delta^{-1}$.

Proof Let t be the number of entries of $\boldsymbol{g}$ obeying $|\boldsymbol{g}_i| \geq \gamma_{2\delta} = Q^{-1}(2\delta)$. First we show that t is around $2\delta n$ with high probability.

Lemma A.5 *We have that $\mathbb{P}(|t - 2\delta n| \geq \delta n) \leq 2\exp(-\delta n/8)$.*

Proof This again follows from a Chernoff bound. In particular $t = \sum_{i=1}^n a_i$ where $\mathbb{P}(a_i = 1) = 2\delta$ and $\{a_i\}_{i=1}^n$ are i.i.d. Consequently $\mathbb{P}(|\sum_{i=1}^n a_i - 2\delta n| > t) \leq 2\exp(-\delta n/8)$. ∎

Conditioned on t, the largest t entries are i.i.d. and distributed as a standard normal $g \sim \mathcal{N}(0,1)$ conditioned on $g \geq Q^{-1}(2\delta)$. Applying Lemma A.3, this implies that

$$\mathbb{E}[\sum_{i=1}^{\delta n}\tilde{g}_i] \leq \mathbb{E}[\sum_{i=1}^{t}\tilde{g}_i] \leq tc\sqrt{\delta^{-1}} \leq 3cn\delta\sqrt{\log\delta^{-1}}.. \tag{A.2}$$

The remaining event occurs with probability $2\exp(-\delta n/8)$. On this event, for any $t \geq 0$, we have that

$$\mathbb{P}(\tilde{\boldsymbol{g}}_1 \geq \sqrt{2\log n} + t) \leq \exp(\delta n/8)\exp(-t^2/2) \tag{A.3}$$

which implies $\mathbb{E}[\tilde{\boldsymbol{g}}_1] \leq c'(\sqrt{\log n} + \sqrt{\delta n})$ and $\mathbb{E}[\sum_{i=1}^{\delta n} \tilde{g}_i] \leq c'\delta n(\sqrt{\log n} + \sqrt{\delta n})$. Combining the two estimates (A.2) and (A.3) yields

$$\mathbb{E}[\sum_{i=1}^{\delta n} \tilde{\boldsymbol{g}}_i] \leq 3cn\delta\sqrt{\log \delta^{-1}} + c'\delta n(\sqrt{\log n} + \sqrt{\delta n})\exp(-\delta n/8).$$

Setting $n = \beta\delta^{-1}$, we obtain

$$\mathbb{E}[\frac{1}{\delta n}\sum_{i=1}^{\delta n} \tilde{g}_i] \leq 3c\sqrt{\log \delta^{-1}} + c'(\sqrt{\log \delta^{-1} + \log \beta} + \sqrt{\beta})\exp(-\beta/8).$$

We can ensure that the second term is $\mathcal{O}(\sqrt{\log \delta^{-1}})$ by picking β to be a large constant to conclude with the result. ∎

Lemma A.6 (Slepian variation) *Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ be a standard Gaussian matrix. Let $\boldsymbol{g} \in \mathbb{R}^n, \boldsymbol{h} \in \mathbb{R}^m, g$ be standard Gaussian vectors. Given compact sets $\mathcal{C}_1 \subset \mathbb{R}^n, \mathcal{C}_2 \subset \mathbb{R}^m$ we have that*

$$\mathbb{E}[\sup_{\boldsymbol{v}\in\mathcal{C}_1,\boldsymbol{u}\in\mathcal{C}_2} \boldsymbol{u}^*\boldsymbol{A}\boldsymbol{v} + \|\boldsymbol{u}\|_{\ell_2}\|\boldsymbol{v}\|_{\ell_2} g] \leq \mathbb{E}[\sup_{\boldsymbol{v}\in\mathcal{C}_1,\boldsymbol{u}\in\mathcal{C}_2} \boldsymbol{v}^*\boldsymbol{g}\|\boldsymbol{u}\|_{\ell_2} + \boldsymbol{u}^*\boldsymbol{h}\|\boldsymbol{v}\|_{\ell_2}].$$

Proof Consider the Gaussian processes $f(\boldsymbol{v},\boldsymbol{u}) = \boldsymbol{u}^*\boldsymbol{A}\boldsymbol{v} + \|\boldsymbol{u}\|_{\ell_2}\|\boldsymbol{v}\|_{\ell_2} g$ and $g(\boldsymbol{v},\boldsymbol{u}) = \boldsymbol{v}^*\boldsymbol{g}\|\boldsymbol{u}\|_{\ell_2} + \boldsymbol{u}^*\boldsymbol{h}\|\boldsymbol{v}\|_{\ell_2}$. We have that

$$\begin{aligned}
&\mathbb{E}[f(\boldsymbol{v},\boldsymbol{u})^2] = \mathbb{E}[g(\boldsymbol{v},\boldsymbol{u})^2] = 2\|\boldsymbol{u}\|_{\ell_2}^2\|\boldsymbol{v}\|_{\ell_2}^2, \\
&\mathbb{E}[f(\boldsymbol{v},\boldsymbol{u})f(\boldsymbol{v}',\boldsymbol{u}')] - \mathbb{E}[g(\boldsymbol{v},\boldsymbol{u})g(\boldsymbol{v}',\boldsymbol{u}')] \geq (\|\boldsymbol{u}\|_{\ell_2}\|\boldsymbol{u}'\|_{\ell_2} - \langle \boldsymbol{u},\boldsymbol{u}'\rangle)(\|\boldsymbol{v}\|_{\ell_2}\|\boldsymbol{v}'\|_{\ell_2} - \langle \boldsymbol{v},\boldsymbol{v}'\rangle) \geq 0.
\end{aligned}$$

Consequently Slepian's Lemma yield $\mathbb{E}[\sup_{\boldsymbol{v}\in\mathcal{C}_1,\boldsymbol{u}\in\mathcal{C}_2} f] \leq \mathbb{E}[\sup_{\boldsymbol{v}\in\mathcal{C}_1,\boldsymbol{u}\in\mathcal{C}_2} g]$ for finite sets $\mathcal{C}_1, \mathcal{C}_2$. A standard covering argument finishes the proof. ∎