# Is Dimensionality a Barrier for Retrieval Models?

**Kiril Bangachev**[*†] kirilb@mit.edu

**Guy Bresler**[*‡] guy@mit.edu

**Jonathan Kogan** jkogan1@mit.edu

**Yury Polyanskiy**[*§] yp@mit.edu

Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA, 02139

## Abstract

*Why does the low dimensionality of representations, typically $d \approx 1000$, not prevent modern embedding-based retrieval models from scaling to billions, or even trillions, of data points?* To answer this question, we study maximal-margin embeddings in the following retrieval model, classically studied in communication complexity [PS86] and more recently in embedding-based retrieval [WBNL26]. Let $A \in \{0,1\}^{N \times n}$ be a matrix indicating whether each of $N$ queries is relevant to each of $n$ documents. We are interested in the largest *margin* $m > 0$, denoted by $\mathsf{m}^{\mathsf{rd}}(d, A)$, for which there exist unit norm embeddings of the queries and documents $\{U_j\}_{j=1}^{N}, \{V_i\}_{i=1}^{n}$ with the following property. $\langle U_j, V_i \rangle \geq m$ whenever $A_{ji} = 1$ and $\langle U_j, V_i \rangle \leq -m$ otherwise. A large margin is a key proxy for representation quality: it controls both robustness to perturbations and compositional generalization across queries. Our main theorem establishes that the best possible margin without a restriction on the dimension, $\mathsf{m}^{\mathsf{rd}}(+\infty, A)$, can be nearly achieved in dimension $d = O(\mathsf{m}^{\mathsf{rd}}(+\infty, A)^{-2} \log n)$ which improves a theorem of [BDES02]. Together with a matching lower bound in Theorem 1.5, we conclude that when $A \in \{0,1\}^{\binom{n}{k} \times n}$ is the matrix containing all possible $k$-sparse rows once, dimension $d = O(k \log(n/k))$ is necessary and sufficient for the maximal possible margin $\mathsf{m}^{\mathsf{rd}}(+\infty, A) = \Theta(k^{-1/2})$ in this setting. This fully resolves the setup of [WBNL26]. We also give several constructions for large margins when $d = o(k \log(n/k))$. Our proofs are based on connections between this problem and the literature on compressed sensing and the restricted isometry property. Finally, we empirically test the InfoNCE and sigmoid losses for producing large margin embeddings and demonstrate a clear advantage of the sigmoid loss. Code is available at github/TopK.

[*] Author names appear in alphabetical order by last name.

[†] Supported by NAE Grand Challenge Vest Fellowship.
[‡] Supported by NSF Award 2428619.
[§] Supported by NSF Grant CCF-21-12665 and by generous gifts from Jane Street and Google Research.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

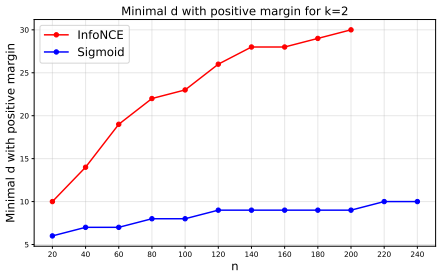



Figure 1: Minimal dimension needed to achieve a non-zero margin after 100000 training steps for the InfoNCE and sigmoid losses. The sigmoid succeeds in much smaller dimensions.




Figure 2: Maximal margin achieved during 100000 training steps for $d \in \{6, 18, 30\}$, using the InfoNCE and sigmoid losses showing clear advantage of the sigmoid loss.

# 1 Introduction

## 1.1 Background and Motivation

Modern information retrieval systems rely on *dense neural embeddings of data*. Depending on the modality, a variety of training paradigms and architectures exist for representing data as vectors: for text, popular choices include transformer-based architectures trained with contrastive or self-supervised objectives over query-passage pairs such as SBERT, DPR, E5, GTE, BGE, EmbeddingGemma [RG19, KOM+20, WYH+22, LZZ+23, CXZ+24, VDZ+25]; for images, popular choices include vision transformers trained with self-distillation or reconstruction objectives such as the DINO family and ViT-MAE [CTM+21, O+23, HCX+22, SVS+25]; finally, in multimodal settings contrastive models such as CLIP and SigLIP are common [RKH+21, ZMKB23, TGW+25].

Once data is represented compactly in vector form (typically $d \leq 4096$ across the wide range of encoders), retrieval is often done via a simple nearest neighbor search around the query embedding [JDS11, JDJ19, MY20, RKH+21, CTM+21, CKNH20, DAT+21, BHS+25] (and many others). The *compactness of the embeddings* enables low-memory caching of representations and fast algorithmic primitives (such as nearest neighbor search). Both speed and low-cost storage are integral to high-performance AI. However, compactness of representations is not enough on its own. For the design of accurate, reliable, and precise AI systems, it is also necessary that embeddings be of *high-quality*.

That *compactness* and *quality* are at odds is no surprise. Low dimension (rank) limits the geometric configurations available to item representations. What should be surprising instead is that models based on low-dimensional representations still achieve excellent performance on retrieval from huge datasets. For instance, DPR embeds data in $d = 768$ dimensions and achieves remarkable performance when retrieving from a pool of $\approx 2 \times 10^7$ passages [KOM+20]. This raises the question:

*Q: How can embedding models be so successful despite their low dimensionality?*

We approach this question theoretically by analyzing the top-$k$ retrieval model recently introduced in [WBNL26]. By studying a key quantity governing the quality of embeddings – *the margin* – we show a surprising low-rank saturation result (formally stated in Theorem 1.4):

*Informal Main Result: The highest possible quality embeddings are already achieved in a dimension growing only logarithmically with the data size.*

Hence, embedding-based retrieval models do not require a high dimension for optimal performance. This main result opens the door to further theoretical results related to precise dimension recommendations for encoders and linking different notions of "quality" of representations. To obtain these results, we connect the model of [WBNL26] to the classical literature on compressed sensing and the restricted isometry property [CT05, FPRU10, FR13, Nov95], communication complexity [AFR85, She09, LS09], and learning theory [FSSS03].

### 1.2 Mathematical Model and Problem Formulation

**Basic Retrieval Model.** In the model of [WBNL26], an incidence matrix $A \in \{0,1\}^{N \times n}$ encodes the relevance relationship between $N$ queries and $n$ objects. A query $j \in \{1, 2, \ldots, N\} = [N]$ is relevant to object $i \in [n]$ if and only if $A_{ji} = 1$. This abstraction captures a variety of practical settings, for example when $[n]$ corresponds to a corpus of passages and $[N]$ to a set of queries [KOM+20], or when $[N]$ corresponds to the images in a dataset and $[n]$ to their classes [RDS+15].

The results of [WBNL26] focus primarily on the case where $N = \binom{n}{k}$ and $A = S_{n,k} \in \{0,1\}^{N \times n}$ contains every possible $k$-sparse row exactly once. In this setting, for every group of $k$ objects, there is a query to which exactly those $k$ objects are relevant. Many of our results apply more broadly to matrices $A$ for which all rows are distinct and $k$-sparse (so that only some groups of $k$ documents can be retrieved, which is more realistic). Typically, we will think about the quantity $k$ as a large constant or slowly growing with $n$, even though our results apply for any $k$ unless otherwise stated. We note that [BBNP25] analyzes the case when $k = 1$, $S_{n,k} = I$.

Corresponding to $A$, models $f_\theta$ and $g_\phi$ produce embeddings, respectively $\{f_\theta(O_i)\}_{i=1}^n = \{V_i\}_{i=1}^n$ of the objects $\{O_i\}_{i=1}^n \subseteq \mathcal{O}$ and $\{g_\phi(Q_j)\}_{j=1}^N = \{U_j\}_{j=1}^N$ of the queries $\{Q_j\}_{j=1}^N \subseteq \mathcal{Q}$. These embeddings are assumed to be unit norm throughout the paper, which is without loss of generality as only cosine similarities between them will be considered.

The goal of the representations is to enable retrieval of objects relevant to a given query. One property that ensures this is if there exists some $\tau \in \mathbb{R}$ such that:

$$\langle U_j, V_i \rangle > \tau \text{ if } A_{ji} = 1 \text{ and } \langle U_j, V_i \rangle < \tau \text{ if } A_{ji} = 0. \tag{1}$$

Indeed, this would imply that the documents relevant to query $j$ are those whose embeddings have cosine similarity with $U_j$ exceeding $\tau$. These documents can be retrieved by the standard procedure of finding the nearest neighbors of $U_j$ and filtering out those with cosine similarity less than $\tau$.

**Retrieval Margin.** While property (1) guarantees correct retrieval, it is too brittle to be useful in practice. A minor modification $O_i'$ of object $O_i$, for example flipping few pixels, may lead to an incorrect retrieval by some query $Q_j$, i.e. $\langle f_\theta(O_i), g_\phi(Q_j) \rangle < \tau < \langle f_\theta(O_i'), g_\phi(Q_j) \rangle$. To capture the fact that strong models should succeed on retrieval even when data is slightly perturbed, we introduce the *margin* of a model. Borrowing the terminology of [BBNP25], we make the following definition.

**Definition 1.** *Given an incidence matrix $A \in \{0,1\}^{N \times n}$, we say that the vectors $\{V_i\}_{i=1}^n, \{U_j\}_{j=1}^N$ are a margin-$m$, relative-bias-$\tau$ embedding of $A$ for $m \geq 0$ if*

$$\langle U_j, V_i \rangle \geq \tau + m \text{ if } A_{ji} = 1 \text{ and } \langle U_j, V_i \rangle \leq \tau - m \text{ if } A_{ji} = 0. \tag{2}$$

The main question that we will study is bounding the maximal possible margin of the embedding. We note that for the rest of the paper, we will only focus on the case of $\tau = 0$ for simplicity of exposition. Following [WBNL26], this is without loss of generality up to increasing the dimension by 1 and multiplying the margin by a small absolute constant. See Appendix A for completeness.

**Problem 1.** *For a given $A \in \{0,1\}^{N \times n}$, find the optimal restricted-dimension margin*

$$\mathsf{m}^{rd}(d, A) := \sup\{m \ : \ \exists \text{ margin-}m, \text{ relative-bias-}0 \text{ embeddings of } A \text{ in dimension } d\}.$$

*Whenever there does not exist a margin-$m$ embedding of $A$ for any $m \geq 0$, we set $\mathsf{m}^{rd}(d, A) = -\infty$.*

The quantity without the dimension restriction, i.e. $\mathsf{m}^{rd}(+\infty, A) := \sup_d \mathsf{m}^{rd}(d, A)$ is known as the *margin complexity* in the communication complexity literature [She09]. [5] The minimal $d$ for which an embedding exists is the *sign-rank*, $\mathsf{sr}(A) := \min\{d \ : \mathsf{m}^{rd}(d, A) > 0\}$ [AFR85].

**Significance of Large-Margin Embeddings.** Large-margin learning has been of interest for many years to the machine learning community [Nov62, CV95, Bar98, FS99, CSSS13, ZRHT03, CS02, SB00, BBV06, AV06, BFT17]. Concrete works on representation learning which highlight the importance of large margins include [BBS08, HCL06, WS09, SKP15, DGXZ19, SGSR17, EKM+18, RCSJ21, BBNP25]. We outline several reasons why it is concretely important for retrieval. The statements and proofs are simple and delayed to Appendix G.

[5] Dimension $n$ is enough to obtain maximal margin, but we stick to the $+\infty$ notation to emphasize the absence of dimension restriction. Also, the sup can be replaced by max due to compactness.

- *Large Margin and Robustness:* We already stated why small-margin embeddings are brittle. We formalize this idea for Lipschitz encoders in Proposition 6.
- *Large Margin and Quantization:* In particular, as large margin embeddings are more robust, they allow for coarser quantization while retaining perfect retrieval. See Proposition 8.
- *Large Margin and Compositional Generalization:* In addition to generalization to correctly retrieving perturbed objects, larger margins also yield higher compositional generalization. Concretely, in the $S_{n,k}$ model, suppose the object embeddings $\{V_i\}_{i=1}^n$ are such that every subset of $k$ of them can be retrieved with margin $m$ via a query embedding. Then every subset of up to $k + 2mk$ objects can also be retrieved. Hence, a larger margin leads to a larger class of queries that can be correctly retrieved. See Proposition 4.
- *Large Margin and Sigmoid Loss:* Following [BBNP25], we show that the global minimizers of sigmoid contrastive loss (with positive pairs determined by the incidence matrix $A$) exactly coincide with margin-$m$, relative-bias-$\tau$ embeddings and the loss has an inductive bias towards large-margin embeddings. See Proposition 7 and [SHN+18] for related work.

**Some Preliminary Observations.** Note that if $B$ is a submatrix of $A$, for any $d$, $\mathsf{m}^{rd}(d, B) \geq \mathsf{m}^{rd}(d, A)$ (since one can just take the embeddings corresponding to the respective rows and columns). Hence, in particular, $\mathsf{m}^{rd}(d, A) \geq \mathsf{m}^{rd}(d, S_{n,k})$ for any $A$ with $k$-sparse rows as well. Similarly, $d \longrightarrow \mathsf{m}^{rd}(d, A)$ is non-decreasing for any fixed $A$.

### 1.3 Prior work

**Sign Rank.** In [PS86], the authors establish that for square matrices $A$, $\log(\mathsf{sr}(A))$ equals (up to additive error 1) the two-way communication complexity of computing $A$ in the model of Yao [Yao79]. In [AMY16], on the other hand, the authors show that a dual version of the sign rank is equivalent (up to a multiplicative factor of 2) to the VC dimension of the function class corresponding to rows of $A$. The following results from these two lines of work are relevant. We provide proofs in Appendix C for completeness (as they are stated in a different language in the original papers).

**Theorem 1.1** ( [AFR85, AMY16]). *If every row of $A \in \{0,1\}^{N\times n}$ is $k$-sparse, then $\mathsf{sr}(A) \leq 2k+1$ [AFR85]. Conversely, when $n \geq 3k+1$, $\mathsf{sr}(S_{n,k}) \geq 2k+1$ [AMY16].*

In Appendix C, we analyze the margin of the $\mathsf{sr}(A) \leq 2k+1$ construction to obtain Corollary 2.

**Margin Complexity.** It turns out that the margin complexity $\mathsf{m}^{rd}(+\infty, A)$ equals (up to a multiplicative error of 8) the discrepancy of $A$ [LS09]. Discrepancy is also used in communication complexity lower bounds [Yao83, BFS86, She09]. Important for us is the following spectral bound on margin complexity, which follows by combining the margin-complexity discrepancy equivalence of [LS09] and the spectral bound of [KN96]. We give a more direct and simpler proof in Appendix D.

**Theorem 1.2** (Spectral Bound on $\mathsf{m}^{rd}(+\infty, A)$ [For02, She09, KN96]). *For any $A \in \{0,1\}^{N\times n}$,*

$$\mathsf{m}^{rd}(+\infty, A) \leq \inf \|J\|_{op}\sqrt{Nn} \text{ over } J \in \mathbb{R}^{N\times n}, \text{ s.t.,}$$
$$J_{ji} \geq 0 \;\forall\; A_{ji} = 1, J_{ji} \leq 0 \;\forall\; A_{ji} = 0, \text{ and } \sum_{j,i} |J_{ji}| = 1.$$

Here, $\|\cdot\|_{op}$ is the operator norm. Concretely, $\|M\|_{op} = \sup_{v\,:\,\|v\|_2=1} \|Mv\|_2$. This upper bound on the margin complexity can be computed efficiently as it minimizes a convex objective over a convex space. In the case of $S_{n,k}$ we can choose $J_{ji} = \frac{1}{2Nk}$ whenever $A_{ji} = 1$ and $J_{ji} = -\frac{1}{2N(n-k)}$ otherwise to conclude the $1/(2\sqrt{k})$ upper bound on $\mathsf{m}^{rd}(+\infty, S_{n,k})$. A simple construction shows that this is tight. Both upper bound and construction are in Appendix D.

**Corollary 1.** *For $k = o(n)$, $\mathsf{m}^{rd}(+\infty, S_{n,k}) = \frac{1+o_k(1)}{2\sqrt{k}}$.*

Distributional versions of margin complexity have also been studied [KMS20].

**Embedding-Based Retrieval.** While [WBNL26] and the follow-up work [WYL+26] mostly focus on sign rank (existence of any positive margin), the following result is relevant.

**Theorem 1.3** ([WBNL26]). *For any dimension-$d$ margin-$m$ embedding of $S_{n,k}$, it holds that $d \geq \frac{\log\binom{n}{k}}{\log(1+1/m)}$. Rewriting, $\mathsf{m}^{rd}(d, S_{n,k}) \leq \left(\binom{n}{k}^{1/d} - 1\right)^{-1}$.*

While a valuable first step, this result falls short of fully answering our motivating question and Problem 1. First, it is a negative result, giving only an upper bound on the achievable margin. [WBNL26] does not provide a complementary positive result demonstrating whether the bound is tight (as we will see shortly, the bound is **not** tight). In particular, it does not elucidate why low-dimensional retrieval models are so successful in practice.

**(Random) Spherical Graphs.** Finally, we note that Definition 1 is related to bipartite masks of spherical geometric graphs. Concretely, a spherical graph $G$ on vertex set $[p]$ is defined by unit vectors $X_1, X_2, \ldots, X_p \in \mathbb{S}^{d-1}$ associated to the vertices. Two vertices $a, b$ are adjacent whenever $\langle X_a, X_b \rangle \geq \tau$ for some fixed $\tau$. In our model, $p = n + N$, the vertices form a bipartition $[N] \cup \{N+1, \ldots, N+n\}$, the latent vectors are $\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n$, and considered are only edges between the two parts $[N], \{N+1, \ldots, N+n\}$. This is exactly the bipartite mask setting in [BBH21]. Recent work on spherical geometric graphs focuses on the setting when the vectors are i.i.d. uniformly random on the sphere [BDER16, LMSY23, MSX25]. We want to emphasize the paper [Gam20] which takes a rather different direction. In this work, the author shows how to construct a geometric graph satisfying a strong Ramsey property from a matrix satisfying a strong restricted isometry property (RIP). Interestingly, our results are also based on techniques and ideas which originated in the study of the restricted isometry property.

### 1.4 Results

We highlight a gap in the literature. There is no positive result which guarantees simultaneously *large margin* (quality) and a *small dimensionality* (compactness) of representations.

**Part 1: Main Result.** Our main result directly addresses this gap.

**Theorem 1.4** (Main Result)**.** *Let $A \in \{0,1\}^{N \times n}$. For every $\epsilon > 0$, there exists some $C_\epsilon = O(\epsilon^{-4})$ depending only on $\epsilon$ such that*

$$\mathsf{m}^{rd}(C_\epsilon \times \mathsf{m}^{rd}(+\infty, A)^{-2} \times \log(\min(n, N)),\ A) \geq (1-\epsilon)\mathsf{m}^{rd}(+\infty, A).$$

We interpret this result as follows: one can obtain nearly the optimal margin $\mathsf{m}^{rd}(+\infty, A)$ (i.e., nearly the same as in dimension $d = \infty$) in dimension $O(\mathsf{m}^{rd}(+\infty, A)^{-2} \times \log(\min(n, N)))$. For concreteness, recall that if each row of $A$ is $k$-sparse, then $\mathsf{m}^{rd}(+\infty, A) \geq \frac{1-o_k(1)}{2\sqrt{k}}$. In particular, an optimal margin for any $k$-sparse matrix can be obtained in dimension $O(k \log n)$!

It turns out that one can recover that $\mathsf{m}^{rd}(Ck \log n, S_{n,k}) = \Omega(1/\sqrt{k})$ for a large enough $C$ from results on compressed sensing [CT05]. However, it seems like the approach is limited to margin $\Theta(1/\sqrt{k})$ and dimension $\Theta(k \log n)$ and falls short of our Theorem 1.4. Theorem 1.4 goes further and implies, for example, that for $k$-sparse $A$ with $\mathsf{m}^{rd}(+\infty, A) = \omega_k(k^{-1/2})$, maximal margin can be achieved in an even smaller dimension $o_k(k \log n)$. See Appendix E.

We note that Theorem 1.4 is a substantial strengthening of the result from [BDES02] that states that $\mathsf{sr}(A) = O(\mathsf{m}^{rd}(+\infty, A)^{-2} \log(nN))$. The result of [BDES02] is for sign rank, so it only states that a positive margin embedding exists, without preserving the margin. More importantly, our result has $\log(\min(n, N))$ dependence instead of $\log(nN) = \Theta(\log(\max(n, N)))$. Instead of our $O(k \log(n))$, [BDES02] achieves dimension $O(k^2 \log(n))$ for margin $\Theta(k^{-1/2})$ for $S_{n,k}$.

Both the proof of our Theorem 1.4, presented in Section 2, and the result of [BDES02] use the Johnson-Lindenstrauss lemma (Theorem 2.1) to perform dimensionality reduction. Concretely, the result $\mathsf{sr}(A) = O(\mathsf{m}^{rd}(+\infty, A)^{-2} \log(nN))$ proceeds by taking embeddings $\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n$ attaining the maximal margin $\mathsf{m}^{rd}(+\infty, A)$ and applying JL on the $Nn$ pairs $(U_j, V_i)$ so that inner products are preserved up to an additive error smaller than the margin $\mathsf{m}^{rd}(+\infty, A)$.

The novelty in our proof is that we apply JL so that *only the norms* of (certain convex combinations of the) $\{V_i\}_{i=1}^n$ (when $n \leq N$) are approximately preserved. Then, we use the following non-constructive certificate for the existence of the desired $N$ vectors corresponding to queries instead of explicitly constructing them. The proof of Lemma 1 is simple and delayed to Appendix A.

**Lemma 1.** *Given are vectors $\{V_i\}_{i=1}^n \in \mathbb{R}^D$, a sign vector $\sigma \in \{\pm 1\}^n$, and some $m > 0$. Then,*

($\Longrightarrow$) *Suppose that there exists a unit vector $U$ with the following property. Whenever $\sigma_i = 1$, it holds that $\langle U, V_i\rangle \geq m$ and whenever $\sigma_i = -1$, it holds that $\langle U, V_i\rangle \leq -m$. Then, for every $x \in \mathsf{conv}(\{\sigma_i V_i\}_{i=1}^n)$, it holds that $\|x\|_2 \geq m$.*

($\Longleftarrow$) *Conversely, suppose that $m = \min\left\{\|x\|_2 \ : \ x \in \mathsf{conv}(\{\sigma_i V_i\}_{i=1}^n)\right\}$. Then, there exists a unit vector $U$ with the following property. Whenever $\sigma_i = 1$, it holds that $\langle U, V_i\rangle \geq m$ and whenever $\sigma_i = -1$, it holds that $\langle U, V_i\rangle \leq -m$.*

We show that this property is satisfied by the (convex combinations of) $\{V_i\}_{i=1}^n$ after the JL-projection via Maurey's Approximation Lemma (Theorem 2.2). The idea for using Maurey's approximation lemma originates from the literature on compressed sensing [FPRU10, FR13, Nov95].

**Part 2: Impossibility Results.** We next prove that the dimension bound in Theorem 1.4 is tight for the $S_{n,k}$ model. In the maximal margin regime $m = \Theta(1/\sqrt{k})$, Theorem 1.3 of [WBNL26] gives the lower bound $d = \Omega(k\log(n/k)/\log(k))$ while Theorem 1.4 gives an upper bound of $d = O(k\log(n))$. There is still a $\log(k)$ gap between these upper and lower bounds. We close this gap with the following theorem, which we prove in Section 3:

**Theorem 1.5.** *Suppose that $n \geq 3k$ and $\mathsf{m}^{rd}(d, S_{n,k}) > 0$. For some universal constant $C > 0$,*

$$d \geq C\frac{k\log(n/k)}{\log(1 + 2(\mathsf{m}^{rd}(d, S_{n,k})\sqrt{k})^{-1})}.$$

We conclude that the minimal dimension required to obtain margin $m = \Theta(1/\sqrt{k})$ is exactly $\Theta(k\log(n/k))$, thus fully resolving the $S_{n,k}$ case [WBNL26]. Theorem 1.5 can be adapted to general $k$-sparse $A$, but the resulting bound is more complex. See Appendix F.

We briefly compare the proof of our Theorem 1.5 with the proof of the weaker $d = \Omega(k\log(n/k)/\log(1/m))$ lower bound of [WBNL26]. In [WBNL26], the authors simply use the fact that if $\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n$ are a margin $m$ embedding of $S_{n,k}$, then the vectors $\{U_j\}_{j=1}^N$ form a radius-$m$ packing of the sphere, to which they apply standard volume arguments.

Our proof also relies on a volume argument. However, instead of directly applying it on the query embeddings $\{U_j\}_{j=1}^N$ or document embeddings $\{V_i\}_{i=1}^n$, we pursue an idea akin to the proof of Theorem 1.4. We non-constructively show the existence of a collection of vectors which also has size $\exp(\Omega(k\log(n/k)))$ but is a packing with a much higher radius of $\Omega(m\sqrt{k})$. This collection is formed by linear combinations of the vectors $\{V_i\}_{i=1}^n$ and again crucially relies on Lemma 1.

**Part 3: Constructions beyond the maximal margin.** Theorems 1.4 and 1.5 primarily address the question of embedding $k$-sparse matrices with maximal margin $\Theta(1/\sqrt{k})$. However, other non-trivial margins could be possible in dimensions smaller than $\Theta(k\log n)$. For instance, a simple analysis of the $2k+1$ sign-rank construction of [AFR85], provided in Appendix C, yields the following result.

**Corollary 2.** $\mathsf{m}^{rd}(2k+1, A) \geq (2n)^{-2k}(2k)^{-1/4}$ *if all rows of $A$ are $k$-sparse.*

We prove a separate result which establishes a smooth trade-off between dimension and margin. It is based on self-Khatri-Rao lifts [KR68]. Namely, we form the $V_i$ vectors roughly by drawing i.i.d. vectors $a_1, a_2, \ldots, a_n \in \mathbb{S}^{\sqrt{d}-1}$ and setting $V_i = a_i \otimes a_i \in \mathbb{S}^{d-1}$. The improvement is also related to the aforementioned connection to the restricted isometry property. Khatri-Rao random lift matrices are known to satisfy stronger RIP conditions than random matrices [KM18].

**Theorem 1.6.** *Suppose that $A \in \{0,1\}^{N\times n}$ is such that every row has $k$ non-zero entries for some $k \geq 2$. Suppose furthermore that $d \geq \frac{k^2+5k+7}{2}$ and let $r := \lfloor\frac{\sqrt{8d-7}-1}{2}\rfloor = \sqrt{2d} + O(1)$. Then,*

$$\mathsf{m}^{rd}(d, A) \geq \frac{\Delta}{2(\sqrt{k}+1) - \Delta}$$

*where $\Delta = \Theta\left(\frac{r-k}{r} \times \left(\frac{1}{(n-k)N}\right)^{\frac{2}{r-k}}\right)$.*

The proof is in Section 4. We illustrate the bound with the following cases when $N = \binom{n}{k}$, $A = S_{n,k}$.

1. If $d \longrightarrow +\infty$, as $r = \sqrt{2d}(1 + o_d(1))$, it follows that $\Delta \longrightarrow 1$. We recover the fact $\mathsf{m}^{rd}(d, S_{n,k}) \geq \frac{1+o_k(1)}{2\sqrt{k}}$. Note, however, that this construction requires $d = \Omega(k^2 \log^2 n)$ for margin $\Theta(1/\sqrt{k})$ even though we know that dimension $\Theta(k \log n)$ is sufficient.
2. If $d = \frac{k^2}{2}(1+\alpha)^2$ for some $\alpha > 0$, we get $\mathsf{m}^{rd}(\frac{k^2}{2}(1+\alpha)^2, S_{n,k}) \geq \frac{1}{2\sqrt{k}} n^{-\frac{2(k+1)}{\alpha k}}$. Thus, even for growing $k$, we obtain any inverse polynomial margin in dimension $\Theta(k^2)$.

**Part 4: Experiments.** We test what margins are obtained by low-dimensional representations trained via contrastive learners in the free embedding model of [WBNL26]. Concretely, we set $k = 2$ and, for varying $n$ and $d$, train the embeddings $\{U_j\}_{j=1}^N$ and $\{V_i\}_{i=1}^n \in \mathbb{S}^{d-1}$ using either the InfoNCE loss [OLV18], as in [WBNL26], or the sigmoid loss used in SigLIP [ZMKB23], with positive (negative) samples determined by $(S_{n,k})_{ji} = 1$ ($(S_{n,k})_{ji} = 0$). See Appendix H for full experimental details.

The work [WBNL26] uses the InfoNCE experiments to suggest that $d = \Theta(n^{1/3})$ is needed for any positive margin. Our experiment on InfoNCE in Figure 1 is consistent with [WBNL26, Figure 2].

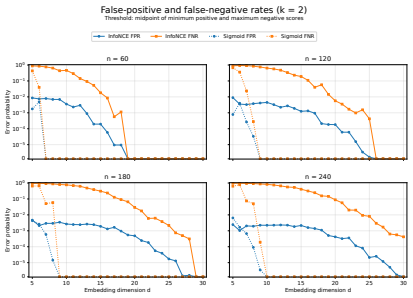



Figure 3: False positive and false negative rates for InfoNCE and sigmoid loss embeddings after 100000 optimization steps. Pairs are classified with respect to the threshold $(\min_{A_{ji}=1}\langle U_j, V_i\rangle + \max_{A_{ji}=0}\langle U_j, V_i\rangle)/2$.

What is surprising in light of [WBNL26] is that the sigmoid loss strongly outperforms InfoNCE. The sigmoid loss needs a much lower dimension, close to the theoretically optimal 5 and nearly independent of $n$, to obtain a positive margin (Figure 1). Likewise, the margins obtained by the sigmoid loss are consistently higher (Figure 2). The behavior for larger $k$ is similar, see Appendix H.

In dimensions where the sigmoid loss, respectively InfoNCE loss, fail to produce perfect embeddings according to Definition 1, the sigmoid loss still has consistently lower false positive and false negative rates (see Figure 3).

A plausible explanation is that the global minimizers of the sigmoid loss (with trainable inverse temperature) coincide exactly with the embeddings of $A$ in the sense of Definition 1. However, in the case of InfoNCE, embeddings need to satisfy some additional combinatorial constraints. See Proposition 7 and Appendix H.

Of course, our experiments only consider margin in the free embedding model. It is possible that for properties of the embeddings relevant to tasks beyond retrieval or for certain real datasets, the InfoNCE loss has advantages over the sigmoid loss.

## 2 Low Dimensions Suffice: Proof of Theorem 1.4

We need the following two classical results to prove Theorem 1.4.

**Theorem 2.1** (Johnson-Lindenstrauss Lemma [JL+84])**.** *Suppose that $\mathcal{X} \subseteq \mathbb{R}^D$ is a set of $n$ points and $\epsilon > 0$. Then, there exists a matrix $\Pi \in \mathbb{R}^{\frac{32 \log n}{\epsilon^2} \times D}$ such that for every $x \in \mathcal{X}$, it holds that $(1-\epsilon)\|x\|_2^2 \leq \|\Pi x\|_2^2 \leq \|x\|_2^2$.*

**Theorem 2.2** (Maurey's Lemma, [Pis81])**.** *Suppose that $\mathcal{Y} = \mathsf{conv}(\{Y_i\}_{i=1}^n)$ for some points $\{Y_i\}_{i=1}^n \in \mathbb{R}^D$ of norm at most 1. Let $T \in \mathbb{N}$ and let $x$ be any point inside $\mathcal{Y}$. Then, there exist (not necessarily different) indices $i_1, i_2, \ldots, i_T$ such that $\|x - \frac{1}{T}\sum_{j=1}^T Y_{i_j}\|_2 \leq \frac{1}{\sqrt{T}}$.*

*Proof of Theorem 1.4.* Without loss of generality, suppose that $n \leq N$. Otherwise, we can argue about $A^T$ instead of $A$. Suppose that there exist unit vectors $\{U_j\}_{j=1}^N$ and $\{V_i\}_{i=1}^n$ of some dimension $D$ such that $(-1)^{A_{ji}+1}\langle U_j, V_i\rangle \geq m$ for every $j \in [N], i \in [n]$. By Lemma 1, for every $j \in [N]$ and $x \in \mathcal{X}_j := \mathsf{conv}(\{(-1)^{A_{ji}+1}V_i\})$, it holds that $\|x\|_2 \geq m$. In particular, for some $T$ that we will choose later, $\|x\|_2 \geq m$ holds for every $x$ in the following set:

$$\mathcal{Z}_T := \left\{ \frac{1}{T}\sum_{\ell=1}^{T} (-1)^{A_{j,i_\ell}+1} V_{i_\ell} \;:\; j \in [N], i_1, i_2, \ldots, i_T \in [n] \right\}.$$

Now, observe that $|\mathcal{Z}_T| \le (2n)^T$. Indeed, this holds, since there are $n^T$ choices for $i_1, i_2, \dots, i_T \in [n]$ and at most $2^T$ choices for the signs $(-1)^{A_{j,i_\ell}+1}$.

We will apply the dimensionality reduction Theorem 2.1 to $\mathcal{Z}_\mathcal{T}$. Let $T = \frac{128}{m^2\epsilon^2}$ so that $T \log n = \Theta(\epsilon^{-2}m^{-2}\log n)$, as desired. By Theorem 2.1, there exists some $\Pi \in \mathbb{R}^{O(\epsilon^{-4}m^{-2}\log n)\times D}$ such that for every $v \in \mathcal{Z}_T$, it holds that $(1-\epsilon/2)^2\|v\|_2^2 \le \|\Pi v\|_2^2 \le \|v\|_2^2$.

Define $\widetilde{V}_i \coloneqq \Pi V_i$. For every $i$, at least one of $V_i$ and $-V_i$ belongs to $\mathcal{Z}_T$ (take $i = i_1 = \cdots = i_T$). Hence, $1-\epsilon/2 \le \|\widetilde{V}_i\|_2 \le 1$. To obtain the final $V$-vectors, we will simply normalize.

Before that, we demonstrate the existence of vectors $\widetilde{U}_j$ for $\{\widetilde{V}_i\}_{i=1}^n$ via Lemma 1. Concretely, take some $j \in [N]$ and consider $\widetilde{X_j} = \mathsf{conv}(\{(-1)^{A_{ji}+1}\widetilde{V}_i\})$. Let $\widetilde{x} \in \widetilde{X_j}$. It is enough to show that $\|\widetilde{x}\|_2 \ge (1-\epsilon)m$. To this end, we use the JL Lemma on $\mathcal{Z}_T$ and Theorem 2.2. Namely, by 2.2, there exist some indices $i_1, i_2, \dots, i_T$ such that

$$\Big\|\widetilde{x} - \frac{1}{T}\sum_{\ell=1}^{T}(-1)^{A_{j,i_\ell}+1}\widetilde{V_{i_\ell}}\Big\|_2 \le T^{-1/2} \le m\epsilon/2.$$

On the other hand, by the guarantees of Theorem 2.1,

$$\Big\|\frac{1}{T}\sum_{\ell=1}^{T}(-1)^{A_{j,i_\ell}+1}\widetilde{V_{i_\ell}}\Big\|_2 = \Big\|\Pi\Big(\frac{1}{T}\sum_{\ell=1}^{T}(-1)^{A_{j,i_\ell}+1}V_{i_\ell}\Big)\Big\|_2 \ge (1-\epsilon/2)\Big\|\frac{1}{T}\sum_{\ell=1}^{T}(-1)^{A_{j,i_\ell}+1}V_{i_\ell}\Big\|_2.$$

Finally, as $\frac{1}{T}\sum_{\ell=1}^{T}(-1)^{A_{j,i_\ell}+1}V_{i_\ell} \in \mathcal{X}_j$, we know that $\|\frac{1}{T}\sum_{\ell=1}^{T}(-1)^{A_{j,i_\ell}+1}V_{i_\ell}\|_2 \ge m$. Hence,

$$\|\widetilde{x}\|_2 \ge \|\frac{1}{T}\sum_{\ell=1}^{T}(-1)^{A_{j,i_\ell}+1}\widetilde{V_{i_\ell}}\|_2 - \|\widetilde{x} - \frac{1}{T}\sum_{\ell=1}^{T}(-1)^{A_{j,i_\ell}+1}\widetilde{V_{i_\ell}}\|_2 \ge m(1-\epsilon/2) - m\epsilon/2 = m(1-\epsilon).$$

Applying Lemma 1, there exist unit vectors $\{\widetilde{U}_j\}_{j=1}^N$ in dimension $O(\epsilon^{-4}m^{-2}\log n)$ such that $(-1)^{A_{ji}+1}\langle\widetilde{V}_i, \widetilde{U_j}\rangle \ge m(1-\epsilon) \quad \forall i,j$. Clearly, this property remains true even if we replace $\{\widetilde{V}_i\}_{i=1}^n$ by $\{\widetilde{V}_i/\|\widetilde{V}_i\|_2\}_{i=1}^n$ as $\|\widetilde{V}_i\|_2 \le 1 \quad \forall i$. □

## 3 Impossibility Results: Proof of Theorem 1.5

We need the following technical ingredients toward constructing large sphere packings.

**Lemma 2** ([FPRU10])**.** *Let $n, s \in \mathbb{N}$ with $s < n$. Then there exists a family $\mathcal{F}$ of subsets of $[n]$ with the following properties. (1) Every set in $\mathcal{F}$ consists of exactly $s$ elements. (2) For all $T, T' \in \mathcal{F}$ with $T \ne T'$, $|T \cap T'| < \frac{s}{2}$. (3) $|\mathcal{F}| \ge (\frac{n}{4s})^{\frac{s}{2}}$.*

**Lemma 3.** *Suppose that $n \ge 3k$. Let $\{V_i\}_{i=1}^n$, $\{U_j\}_{j=1}^N$ be any dimension-$d$ margin-$m$ embedding of $S_{n,k}$. Then, for any $x \in \mathbb{R}^n$ with $\|x\|_0 \le k$, it holds that $\|Vx\|_2 \ge m\|x\|_1$.*

*Proof.* This follows immediately from Lemma 1. By homogeneity, it is enough to prove the fact when $\|x\|_1 = 1$. Now, take some $j$ such that $(-1)^{A_{ji}+1}$ and $x_i$ have the same sign whenever $x_i \ne 0$. Then, $Vx \in \mathsf{conv}(\{(-1)^{A_{ji}+1}V_i\}_{i=1}^n)$ and the result follows. □

*Proof of Theorem 1.5.* Let $V$ be as in Lemma 3. We first show that $m\sqrt{k} \le 1$. We already know even the sharper inequality $m \le \frac{1+o_k(1)}{2\sqrt{k}}$ from Corollary 1 but will nevertheless give a proof as the proof technique will be useful later on. Fix any set $T \subset [n]$ with $|T| = k$, and let $\epsilon$ be a uniformly random vector in $\{\pm 1\}^T$. Define $x = \sum_{i\in T}\epsilon_i e_i$. By Lemma 3, we have $\|Vx\|_2 = \|\sum_{i\in T}\epsilon_i V_i\|_2 \ge m\|x\|_1 = mk$. On the other hand,

$$\mathbb{E}\Big[\|\sum_{i\in T}\epsilon_i V_i\|_2^2\Big] = \mathbb{E}\Big[\sum_{i\in T}\|V_i\|_2^2 + \sum_{\substack{i,\ell\in T\\ i\ne\ell}}\epsilon_i\epsilon_\ell\langle V_i, V_\ell\rangle\Big] = \sum_{i\in T}\|V_i\|_2^2 = k,$$

where we used that $\epsilon_i$ and $\epsilon_\ell$ are independently chosen from $\pm 1$ for $i \ne \ell$. Thus, there exists an $\epsilon \in \{\pm 1\}^T$ such that $\|\sum_{i\in T}\epsilon_i V_i\|_2^2 \le k$. Combining with $\|\sum_{i\in T}\epsilon_i V_i\|_2 \ge mk$, we get $m^2k^2 \le k$, and therefore $m\sqrt{k} \le 1$. Assume $k \ge 8$ as otherwise the theorem holds after decreasing $C$.

Now, we will construct the large sphere packing. Set $s \coloneqq \lfloor \frac{k}{8} \rfloor$. By Lemma 2 there exists a set $\mathcal{F} \subset \binom{[n]}{s}$ such that for any $T \neq T'$ with $T, T' \in \mathcal{F}$, $|T \cap T'| < \frac{s}{2}$ and $|\mathcal{F}| \geq (\frac{n}{4s})^{\frac{s}{2}}$.

For a $T \in \mathcal{F}$, define the sign vector $\sigma^T \in \{\pm 1\}^T$ such that $\| \sum_{i \in T} \sigma_i^T V_i \|_2$ is minimized. Denote $y_T \coloneqq \sum_{i \in T} \sigma_i^T V_i$. By the same argument as above, $\|y_T\|_2^2 = \| \sum_{i \in T} \sigma_i^T V_i \|_2^2 \leq s$. In other words, $y_T \in B_d(0, \sqrt{s})$ for all $T \in \mathcal{F}$ where $B_d(0, r)$ is the $L_2$ ball of radius $r$ centered at 0.

Define $x_T \coloneqq \sum_{i \in T} \sigma_i^T e_i$ so that $Vx_T = y_T$. Consider any $T, T' \in \mathcal{F}$ with $T \neq T'$. Since $|T \cap T'| < \frac{s}{2}$ and $s \leq \frac{k}{8}$, it follows that $\|x_T - x_{T'}\|_0 \leq k$. Hence, by Lemma 3 we get

$$\|V(x_T - x_{T'})\|_2 = \|Vx_T - Vx_{T'}\|_2 \geq m \|x_T - x_{T'}\|_1.$$

Moreover, as every coordinate of $x_T, x_{T'}$ is in $\{0, \pm 1\}$,

$$\|x_T - x_{T'}\|_1 \geq |T \triangle T'| = |T| + |T'| - 2|T \cap T'| > 2s - s = s.$$

Combining the last two inequalities, we obtain

$$\|y_T - y_{T'}\|_2 = \|V(x_T - x_{T'})\|_2 = \|Vx_T - Vx_{T'}\|_2 \geq m \|x_T - x_{T'}\|_1 \geq ms.$$

Thus, the vectors $\{y_T\}_{T \in \mathcal{F}}$ are pairwise $ms$ separated in the Euclidean ball $B_d(0, \sqrt{s})$. So, the vectors $\tilde{y}_T \coloneqq \frac{y_T}{\sqrt{s}}$ are pairwise $m\sqrt{s}$ separated in the ball $B_d(0, 1)$. By a standard volume bound (for example [Ver26, Corollary 4.2.11]), it follows that $|\mathcal{F}| \leq (1 + \frac{2}{m\sqrt{s}})^d$. Using $|\mathcal{F}| \geq (\frac{n}{4s})^{\frac{s}{2}}$,

$$(\frac{n}{4s})^{\frac{s}{2}} \leq |\mathcal{F}| \leq (1 + \frac{2}{m\sqrt{s}})^d \implies d \log(1 + \frac{2}{m\sqrt{s}}) \geq \frac{s}{2} \log \frac{n}{4s}.$$

By compactness, there exists an embedding attaining margin $\mathsf{m}^{rd}(d, S_{n,k})$. Taking $m = \mathsf{m}^{rd}(d, S_{n,k})$ in the preceding argument and using $s = \Theta(k)$, the conclusion follows. □

## 4 Construction: Proof of Theorem 1.6

The construction will be based on lifts of random spherical vectors. To this end, we will need the following standard facts about the distribution of random spherical vectors.

**Fact 1** ([FM90]). *Let $z = (z_1, \ldots, z_r)$ be a uniformly random vector on $\mathbb{S}^{r-1}$. Then, for any $1 \leq s \leq r$, the distribution of $z_1^2 + \ldots + z_s^2$ is the distribution $\mathsf{Beta}(s/2, (r-s)/2)$ with density*

$$f_{s/2,(r-s)/2}(x) = \frac{\Gamma(\frac{r}{2})}{\Gamma(\frac{s}{2})\Gamma(\frac{r-s}{2})} x^{\frac{s}{2}-1}(1-x)^{\frac{r-s}{2}-1}.$$

An immediate corollary is the following tail estimate which we prove in Appendix B.

**Proposition 1.** *Let $X \sim \mathsf{Beta}(s/2, (r-s)/2)$ for some $2 \leq s \leq r-2$ and $\delta \in (0, 1]$. Then,*

$$\mathbb{P}[X \leq \delta] \leq \frac{\Gamma(\frac{r}{2})}{\Gamma(\frac{s}{2})\Gamma(\frac{r-s}{2})} \delta^{s/2}.$$

*Proof of Theorem 1.6.* Let $r = \lfloor \frac{\sqrt{8d-7}-1}{2} \rfloor$ so that $d \geq \frac{r(r+1)}{2} + 1$ and $r \geq k+2$. Define the map $\mathsf{vec} : \mathbb{R}^{r \times r} \longrightarrow \mathbb{R}^{r(r+1)/2}$ as the vector $\mathsf{vec}(M)$ with entries $(M_{ii})_{i=1}^r$ and $(\sqrt{2}M_{ij})_{i<j}$. We use the fact that for two symmetric matrices $M_1, M_2$, it holds that $\langle \mathsf{vec}(M_1), \mathsf{vec}(M_2) \rangle = \langle M_1, M_2 \rangle_F$.

Now, we can get to the actual construction. Let $a_1, a_2, \ldots, a_n$ be i.i.d. vectors uniformly drawn from $\mathbb{S}^{r-1}$. Let $\widetilde{V}_i = \mathsf{vec}(a_i a_i^T)$. Let $P_j \in \mathbb{R}^{r \times r}$ be the symmetric matrix projecting onto $\mathsf{span}(\{a_i : A_{ji} = 1\})$. Let $\widetilde{U}_j = \mathsf{vec}(P_j)$. Note that for every $i, j$, $\|\widetilde{V}_i\|_2 = \|a_i a_i^T\|_F = 1$ and $\|\widetilde{U}_j\|_2 = \|P_j\|_F = \sqrt{k}$ as it is a projection matrix on a subspace of dimension $k$. The actual $U_j, V_i$ that we will construct are simple linear transformations of $\widetilde{U}_j, \widetilde{V}_i$. First, however, we analyze the inner products between the constructed $\widetilde{U}_j$ and $\widetilde{V}_i$.

If $A_{ji} = 1$, then $\langle \widetilde{V}_i, \widetilde{U}_j \rangle = \langle P_j, a_i a_i^T \rangle = \langle P_j a_i, a_i \rangle = \langle a_i, a_i \rangle = 1$ as $a_i \in \mathsf{span}(\{a_i : A_{ji} = 1\})$.

If $A_{ji} = 0$, $\langle \widetilde{V}_i, \widetilde{U}_j \rangle = \langle P_j, a_i a_i^T \rangle = \langle P_j a_i, a_i \rangle = \langle P_j a_i, P_j a_i \rangle = \|P_j a_i\|_2^2 \sim \mathsf{Beta}(k/2, (r-k)/2)$ by Fact 1 since $P_j$ is a projection to a $k$-dimensional subspace of $\mathbb{R}^r$ independent of $a_i$. Therefore, $1 - \|P_j a_i\|_2^2 \sim \mathsf{Beta}((r-k)/2, k/2)$.

Now, let $\Delta = \left( \frac{\Gamma(\frac{k}{2})\Gamma(\frac{r-k}{2})}{2N(n-k)\Gamma(\frac{r}{2})} \right)^{\frac{2}{r-k}}$. By Proposition 1,

$$\mathbb{P}[\langle \widetilde{V}_i, \widetilde{U}_j \rangle \geq 1 - \Delta] = \mathbb{P}[1 - \|P_j a_i\|_2^2 \leq \Delta] \leq \frac{\Gamma(\frac{r}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{r-k}{2})} \Delta^{\frac{r-k}{2}} = \frac{1}{2N(n-k)}.$$

By union bound, it follows that there exists a realization of $a_1, a_2, \ldots, a_n$ such that $\langle \widetilde{V}_i, \widetilde{U}_j \rangle \leq 1 - \Delta$ for every $j, i$ such that $A_{ji} = 0$.

The final construction in dimension $\frac{r(r+1)}{2} + 1 \leq d$ is formed by setting $\beta = \sqrt{\frac{2\sqrt{k}}{2(\sqrt{k}+1)-\Delta}}$ and

$$V_i = (\beta \widetilde{V}_i, \sqrt{1-\beta^2}), U_j = (\frac{\beta}{\sqrt{k}} \widetilde{U}_j, -\sqrt{1-\beta^2}).$$

As $\|\widetilde{V}_i\|_2 = 1$, $\|\widetilde{U}_j\|_2 = \sqrt{k}$, it follows that $\|U_j\|_2 = 1$, $\|V_i\| = 1$. Using $\langle \widetilde{V}_i, \widetilde{U}_j \rangle = 1$, we can check that whenever $A_{ji} = 1$, $\langle U_j, V_i \rangle = \frac{\Delta}{2(\sqrt{k}+1)-\Delta}$. Whenever $A_{ji} = 0$, using that $\langle \widetilde{V}_i, \widetilde{U}_j \rangle \leq 1 - \Delta$, we conclude that $\langle U_j, V_i \rangle \leq -\frac{\Delta}{2(\sqrt{k}+1)-\Delta}$. All that is left to show is the asymptotics of $\Delta$. This is a standard corollary of Stirling's approximation which we prove for completeness in Appendix B. □

## 5 Limitations, Broader Impact, and LLM Usage Statement

One gap in our work is that in the regime $d = o(k \log(n/k))$, the lower bound on $\mathsf{m}^{rd}(d, S_{n,k})$ in Theorem 1.6 and upper bound in Theorem 1.5 do not match. A second limitation comes from the fact that our current experiments are in the free embedding model rather than on practical datasets and architectures. We believe that both of these are exciting directions for future study.

We are not aware of any direct negative societal impact of the work.

We used LLMs for completing some proofs, finding related works, generating code, and editing. We wrote the current manuscript without any LLM usage and only then used LLMs for editing. We have verified all code produced by LLMs. More detail about LLM usage is given in Appendix I.

## Acknowledgements

KB is thankful to Pravesh Kothari and Sidhanth Mohanty for insightful conversations at different stages of the current work.

## References

[AFR85] Noga Alon, Peter Frankl, and Vojtěch Ródl. Geometrical realization of set systems and probabilistic communication complexity. In *26th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 277–280. IEEE, 1985.

[AMY16] Noga Alon, Shay Moran, and Amir Yehudayoff. Sign rank versus vc dimension. In *Conference on Learning Theory*, pages 47–80. PMLR, 2016.

[AV06] Rosa I. Arriaga and Santosh S. Vempala. An algorithmic theory of learning: Robust concepts and random projection. *Machine Learning*, 63(2):161–182, 2006.

[Bar98] Peter L. Bartlett. The sample complexity of pattern classification with neural networks: The size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.

[BBH21] Matthew Brennan, Guy Bresler, and Brice Huang. De finetti-style results for wishart matrices: Combinatorial structure and phase transitions, 2021.

[BBNP25] Kiril Bangachev, Guy Bresler, Iliyas Noman, and Yury Polyanskiy. Global minimizers of sigmoid contrastive loss, 2025.

[BBS08] Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. A theory of learning with similarity functions. *Machine Learning*, 72(1–2):89–112, 2008.

[BBV06] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. Kernels as features: On kernels, margins, and low-dimensional mappings. *Machine Learning*, 65(1):79–94, 2006.

[BDDW08] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive approximation*, 28(3):253–263, 2008.

[BDER16] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, 49(3):503–532, 2016.

[BDES02] Shai Ben-David, Nadav Eiron, and Hans Ulrich Simon. Limitations of learning via embeddings in euclidean half spaces. *Journal of Machine Learning Research*, 3(Nov):441–461, 2002.

[BFS86] Laszlo Babai, Peter Frankl, and Janos Simon. Complexity classes in communication complexity theory. In *27th Annual Symposium on Foundations of Computer Science (sfcs 1986)*, pages 337–347, 1986.

[BFT17] Peter L. Bartlett, Dylan J. Foster, and Matus J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6240–6249, 2017.

[BHS+25] Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, et al. Perception encoder: The best visual embeddings are not at the output of the network. *arXiv preprint arXiv:2504.13181*, 2025.

[CKNH20] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[CS02] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292, March 2002.

[CSSS13] Andrew Cotter, Shai Shalev-Shwartz, and Nathan Srebro. Learning optimally sparse support vector machines. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page I–266–I–274. JMLR.org, 2013.

[CT05] Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[CTM+21] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

[CXZ+24] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, 2024.

[DAT+21] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9588–9597, 2021.

[DGXZ19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[EKM+18] Gamaleldin F. Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In Samy Bengio, Hanna Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 842–852, 2018.

[FM90] Peter Frankl and Hiroshi Maehara. Some geometric applications of the beta distribution. *Annals of the Institute of Statistical Mathematics*, 42(3):463–474, 1990.

[For02] Jürgen Forster. A linear lower bound on the unbounded error probabilistic communication complexity. *Journal of Computer and System Sciences*, 65(4):612–625, 2002.

[FPRU10] Simon Foucart, Alain Pajor, Holger Rauhut, and Tino Ullrich. The gelfand widths of lp-balls for 0<p<1. *Journal of Complexity*, 26(6):629–640, 2010.

[FR13] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York, 2013.

[FS99] Yoav Freund and Robert E. Schapire. Large margin classification using the perceptron algorithm. *Machine Learning*, 37(3):277–296, 1999.

[FSSS03] Jürgen Forster, Niels Schmitt, Hans Ulrich Simon, and Thorsten Suttorp. Estimating the optimal margins of embeddings in euclidean half spaces. *Machine Learning*, 51(3):263–281, 2003.

[Gam20] David Gamarnik. Explicit construction of rip matrices is ramsey-hard. *Communications on Pure and Applied Mathematics*, 73(9):2043–2048, 2020.

[GKPS67] Branko Grünbaum, Victor Klee, Micha A Perles, and Geoffrey Colin Shephard. *Convex polytopes*, volume 16. Springer, 1967.

[HCL06] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006.

[HCX+22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[JDJ19] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

[JDS11] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011.

[JL+84] William B Johnson, Joram Lindenstrauss, et al. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

[KM18] Saurabh Khanna and Chandra R. Murthy. On the restricted isometry of the columnwise khatri–rao product. *IEEE Transactions on Signal Processing*, 66(5):1170–1183, 2018.

[KMS20] Pritish Kamath, Omar Montasser, and Nathan Srebro. Approximate is good enough: Probabilistic variants of dimensional and margin complexity. In *Conference on Learning Theory*, pages 2236–2262. PMLR, 2020.

[KN96] Eyal Kushilevitz and Noam Nisan. *Communication complexity*. Cambridge University Press, USA, 1996.

[KOM+20] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, 2020.

[KR68] Chinubhai G Khatri and C Radhakrishna Rao. Solutions to some functional equations and their applications to characterization of probability distributions. *Sankhyā: the Indian journal of statistics, series A*, pages 167–180, 1968.

[LMSY23] Siqi Liu, Sidhanth Mohanty, Tselil Schramm, and Elizabeth Yang. Local and global expansion in random geometric graphs. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 817–825, 2023.

[LS09] Nati Linial and Adi Shraibman. Learning complexity vs communication complexity. *Combinatorics, Probability and Computing*, 18(1-2):227–245, 2009.

[LZZ+23] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023.

[MSX25] Jie Ma, Wujie Shen, and Shengjie Xie. An exponential improvement for ramsey lower bounds. *arXiv preprint arXiv:2507.12926*, 2025.

[MY20] Yu. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020.

[Nov62] Albert BJ Novikoff. On convergence proofs on perceptrons. In *Proceedings of the Symposium on the Mathematical Theory of Automata*, volume 12, pages 615–622. New York, NY, 1962.

[Nov95] Erich Novak. Optimal recovery and n-widths for convex classes of functions. *Journal of Approximation Theory*, 80(3):390–408, 1995.

[O+23] Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023.

[OLV18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Pis81] Gilles Pisier. Remarques sur un résultat non publié de b. maurey. *Séminaire d'Analyse fonctionnelle (dit" Maurey-Schwartz")*, pages 1–12, 1981.

[PS86] Ramamohan Paturi and Janos Simon. Probabilistic communication complexity. *Journal of Computer and System Sciences*, 33(1):106–123, 1986.

[RCSJ21] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021.

[RDS+15] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[RG19] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992, 2019.

[RKH+21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[SB00] Alexander J. Smola and Peter J. Bartlett. *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, USA, 2000.

[SGSR17] Jure Sokolić, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017.

[She09] Alexander Alexandrovich Sherstov. Lower bounds in communication complexity and learning theory via analytic methods. 2009.

[SHN+18] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[SVS+25] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[TGW+25] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025.

[VDZ+25] Henrique Schechter Vera, Sahil Dua, Biao Zhang, Daniel Salz, Ryan Mullins, Sindhu Raghuram Panyam, Sara Smoot, Iftekhar Naim, Joe Zou, Feiyang Chen, et al. Embeddinggemma: Powerful and lightweight text representations. *arXiv preprint arXiv:2509.20354*, 2025.

[Ver26] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2026.

[WBNL26] Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. In *The Fourteenth International Conference on Learning Representations*, 2026.

[WS09] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

[WYH+22] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022.

[WYL+26] Zihao Wang, Hang Yin, Lihui Liu, Hanghang Tong, Yangqiu Song, Ginny Wong, and Simon See. $\mathbb{R}^{2k}$ is theoretically large enough for embedding-based top-k retrieval. *arXiv preprint arXiv:2601.20844*, 2026.

[Yao79] Andrew Chi-Chih Yao. Some complexity questions related to distributive computing(preliminary report). In *Proceedings of the Eleventh Annual ACM Symposium on Theory of Computing*, STOC '79, page 209–213, New York, NY, USA, 1979. Association for Computing Machinery.

[Yao83] Andrew C. Yao. Lower bounds by probabilistic arguments. In *24th Annual Symposium on Foundations of Computer Science (sfcs 1983)*, pages 420–428, 1983.

[ZMKB23] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952, 2023.

[ZRHT03] Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'03, page 49–56, Cambridge, MA, USA, 2003. MIT Press.

# A Proofs of Preliminary Claims

## A.1 Relative Bias and Margin

**Proposition 2** (On Relative Bias and Margin)**.** *Suppose that $A \in \{0,1\}^{N\times n}$ and there exists a margin-$m$ relative bias-$\tau$ embedding of $A$ in dimension $d$. Then, there exists a relative bias-$0$ margin-$\frac{m}{1+|\tau|} \geq \frac{m}{2}$ embedding of $A$ in dimension $d+1$.*

*Proof.* Suppose that $\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n$ is a margin-$m$, relative bias-$\tau$ embedding of $A$. Then, $\langle U_j, V_i\rangle \geq \tau + m$ whenever $A_{ji} = 1$ and $\langle U_j, V_i\rangle \leq \tau - m$ otherwise.

Now, let $\widetilde{V_i} \coloneqq (\frac{1}{\sqrt{1+|\tau|}}V_i, \frac{\sqrt{|\tau|}}{\sqrt{1+|\tau|}})$ and $\widetilde{U_j} \coloneqq (\frac{1}{\sqrt{1+|\tau|}}U_j, -\mathsf{sign}(\tau)\frac{\sqrt{|\tau|}}{\sqrt{1+|\tau|}})$. One can check that $\{\widetilde{U_j}\}_{j=1}^N, \{\widetilde{V_i}\}_{i=1}^n$ are unit norm and $\langle\widetilde{U_j}, \widetilde{V_i}\rangle \geq \frac{m}{1+|\tau|}$ whenever $A_{ji} = 1$ and $\langle\widetilde{U_j}, \widetilde{V_i}\rangle \leq -\frac{m}{1+|\tau|}$ otherwise. Finally, note that $\tau \leq \tau + m \leq \langle U_j, V_i\rangle \leq \|U_j\|_2 \times \|V_i\|_2 = 1$ whenever $A_{ji} = 1$. Similarly, $\tau \geq -1$. □

## A.2 Proof of Lemma 1

($\Longrightarrow$) Take any $x \in \mathsf{conv}(\{\sigma_i V_i\}_{i=1}^n)$. By definition, for some non-negative $\{\alpha_i\}_{i=1}^n$ with sum 1, $x = \sum_{i=1}^n \alpha_i\sigma_i V_i$. As $U$ is unit and $\sigma_i\langle U, V_i\rangle \geq m \quad \forall i$, it follows that

$$\|x\|_2 \geq \langle U, x\rangle = \langle U, \sum_{i=1}^n \alpha_i\sigma_i V_i\rangle = \sum_{i=1}^n \alpha_i\sigma_i\langle U, V_i\rangle \geq \sum_{i=1}^n \alpha_i m = m.$$

($\Longleftarrow$) Take $U_1$ to be the vector of minimal norm inside $\mathsf{conv}(\{\sigma_i V_i\}_{i=1}^n)$. We know that $\|U_1\|_2 = m$. Let $U' \coloneqq \frac{U_1}{\|U_1\|_2}$. We claim that the unit vector $U'$ satisfies the desired property. For the sake of contradiction, suppose that this is not true. Then, there exists some $i$ such that $\langle \sigma_i V_i, U' \rangle < m$. Equivalently, $\langle \sigma_i V_i, U_1 \rangle < m^2 = \|U_1\|_2^2$. Thus, $\langle \sigma_i V_i - U_1, U_1 \rangle < 0$. This, however, means that for some sufficiently small $\alpha > 0$, it holds that the vector $\alpha\sigma_i V_i + (1-\alpha)U_1$, which is inside $\mathsf{conv}(\{\sigma_i V_i\}_{i=1}^n)$, has smaller norm than $\|U_1\|_2 = m$, which is a contradiction. Indeed, note that

$$\frac{d}{d\alpha}\bigg|_{\alpha=0} \|\alpha\sigma_i V_i + (1-\alpha)U_1\|_2^2 = 2\langle \sigma_i V_i - U_1, U_1 \rangle < 0.$$

# B Missing Details from Section 4

## B.1 Proof of Proposition 1

$$\mathbb{P}[X \le \delta] = \int_0^\delta f_{s/2,(r-s)/2}(x)dx = \frac{\Gamma(\frac{r}{2})}{\Gamma(\frac{s}{2})\Gamma(\frac{r-s}{2})} \int_0^\delta x^{\frac{s}{2}-1}(1-x)^{\frac{r-s}{2}-1}dx$$
$$\le \frac{\Gamma(\frac{r}{2})}{\Gamma(\frac{s}{2})\Gamma(\frac{r-s}{2})} \int_0^\delta x^{\frac{s}{2}-1}dx \le \frac{\Gamma(\frac{r}{2})}{\Gamma(\frac{s}{2})\Gamma(\frac{r-s}{2})} \int_0^\delta \delta^{\frac{s}{2}-1}dx \le \frac{\Gamma(\frac{r}{2})}{\Gamma(\frac{s}{2})\Gamma(\frac{r-s}{2})}\delta^{s/2}.$$

## B.2 Gamma Function Asymptotics

To finish the asymptotics of $\Delta$ from Section 4, all we need to show is that for real numbers $x, y > 1/2$, with $x = \frac{k}{2}, y = \frac{r-k}{2}$, the following binomial-type approximation holds.

**Lemma 4.** *Suppose that $x, y \ge 1/2$ are real numbers. Then,*

$$\left(\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}\right)^{1/y} = \Theta\Big(\frac{y}{x+y}\Big).$$

*Proof.* By Stirling's formula, uniformly for $z \ge \frac{1}{2}$, $\Gamma(z) = \Theta\big(z^{z-\frac{1}{2}}e^{-z}\big)$. Hence

$$\left(\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}\right)^{1/y} = \Theta\left(\frac{x^{x-\frac{1}{2}}y^{y-\frac{1}{2}}}{(x+y)^{x+y-\frac{1}{2}}}\right)^{1/y} = \frac{y}{x+y}\left(1+\frac{y}{x}\right)^{-x/y}\left(\frac{x+y}{xy}\right)^{1/(2y)}.$$

Denote $t = x/y > 0$. Then,

$$\left(1+\frac{y}{x}\right)^{-x/y} = (1+t^{-1})^{-t} = \Theta(1),$$

since $1 \le (1+t^{-1})^t \le e$. Also, because $x, y \ge \frac{1}{2}$,

$$\frac{1}{2y}\log\left(\frac{x+y}{xy}\right) = \frac{1}{2y}\log\left(\frac{1}{x}+\frac{1}{y}\right)$$

is bounded above by $\log 4$, and below by

$$\frac{1}{2y}\log\left(\frac{1}{y}\right) = -\frac{\log y}{2y} \ge -\frac{1}{2e},$$

so

$$\left(\frac{x+y}{xy}\right)^{1/(2y)} = \Theta(1).$$

Therefore

$$\left(\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}\right)^{1/y} = \Theta\Big(\frac{y}{x+y}\Big),$$

as desired. □

## C Margin of the Construction of [AFR85]

### C.1 Proof of Theorem 1.1

*Proof of Lower Bound in Theorem 1.1, following [AMY16].* The goal is to show that $\mathsf{sr}(S_{n,k}) \geq 2k+1$. Towards this, we use the notion of dual sign rank.

Consider any matrix $A \in \{0,1\}^{N \times n}$. The set of columns indexed by $C$ is antipodally shattered if for every $v \in \{0,1\}^{|C|}$, either $v$ or $\neg v$ (in which all coordinates are flipped) appears as a row in $A$ restricted to $C$. Now, according to [AMY16, Corollary 1.6], the dual sign rank $\mathsf{dsr}(A)$ coincides with the largest possible $c$ for which a set of antipodally shattered columns $C$ with size $c$ exists.

Clearly, for $S_{n,k}$ then, $\mathsf{dsr}(S_{n,k}) = 2k+1$ since among the first $2k+1$ columns, every vector of Hamming weight at most $k$ exists. Finally, the definition of dual sign rank immediately implies that for any $A$, $\mathsf{dsr}(A) \leq \mathsf{sr}(A)$. It follows that $\mathsf{sr}(S_{n,k}) \geq \mathsf{dsr}(S_{n,k}) \geq 2k+1$. □

*Proof of Upper Bound in Theorem 1.1, following [AFR85].* The construction is based on a Vandermonde embedding of the document embeddings. We will not assume for now that $\{U_j\}_{j=1}^{\binom{n}{k}}$ and $\{V_i\}_{i=1}^{n}$ are unit norm since rescaling the norms will not change the signs.

Take any real numbers $t_1 < t_2 < \cdots < t_n$. Let $V_i = (1, t_i, t_i^2 \cdots, t_i^{2k})$.

Now, take any $j \in [\binom{n}{k}]$. We want to find a $U_j = (u_{j,0}, u_{j,1}, \ldots, u_{j,2k}) \in \mathbb{R}^{2k+1}$ such that $\langle U_j, V_i \rangle > 0$ whenever $(S_{n,k})_{ji} = 1$ and $\langle U_j, V_i \rangle < 0$ otherwise. Toward this end, observe that

$$\langle U_j, V_i \rangle = \sum_{\ell=0}^{2k} u_{j,\ell} t_i^{\ell} = P_j(t_i)$$

for the polynomial of degree $2k$ with coefficient vector $(u_{j,0}, u_{j,1}, \ldots, u_{j,2k})$.

Hence, it is enough to find a polynomial of degree $2k$ $P_j$ such that $P_j(t_i) > 0$ whenever $(S_{n,k})_{ji} = 1$ and $P_j(t_i) < 0$ otherwise.

Note, however, that the polynomial $P_j$ only needs to take $k$ positive values on the set $\{t_1, t_2, \ldots, t_n\}$ as $S_{n,k}$ has $k$-sparse rows. Hence, the polynomial needs to change its sign on at most $2k$ of the intervals $[t_1, t_2], [t_2, t_3], \ldots, [t_{n-1}, t_n]$. As for any $2k$ real numbers, there exists a polynomial of degree $2k$ with roots those numbers, we are done. □

### C.2 Proof of Corollary 2

We make a concrete choice of $t_1, t_2, \ldots, t_n$ and the polynomials from Appendix C.1.

*Proof of Corollary 2.* In the construction above, choose $t_i := -1 + \frac{2(i-1)}{n-1}$ for each $i \in [n]$ so that $\tilde{V}_i = (1, t_i, t_i^2 \cdots, t_i^{2k})$ and $V_i = \tilde{V}_i / \|\tilde{V}_i\|_2$. Note that as $t_i \in [-1, 1]$ for all $i$, it holds that $1 \leq \|\tilde{V}_i\|_2 \leq \sqrt{2k+1}$ for all $i$.

Now, take some $j \in [\binom{n}{k}]$. Our goal is to show that there exists a unit $U_j$ such that $\langle U_j, V_i \rangle \geq (2k)^{-1/4}(2n)^{-2k}$.

Let $1 \leq i_1 < i_2 < \cdots < i_r \leq n-1$ be the indices at which the sign changes, so that $A_{ji_\ell} \neq A_{ji_{\ell+1}}$ for $\ell = 1, \ldots, r$. Again, $r \leq 2k$.

For each such index, set $s_\ell := \frac{t_{i_\ell} + t_{i_\ell+1}}{2}$. Now define

$$p_j(s) := (-1)^{A_{ji_1}+1} \prod_{\ell=1}^{r} (s_\ell - s).$$

Then $p$ has degree $r \leq 2k$. Moreover, since the roots $s_1, \ldots, s_r$ lie exactly in the gaps where the sign pattern changes, the sign of $p_j(t_i)$ flips exactly when $A_{ji}$ flips. Since also $p_j(t_1)$ has sign $(-1)^{A_{j1}+1}$, it follows that $p_j(t_i)$ has sign $(-1)^{A_{ji}+1}$ for every $i \in [n]$. All we have to do is analyze the margin.

Denote $\Delta := \frac{2}{n-1}$. Because the points $t_1, \ldots, t_n$ are equally spaced and each $s_\ell$ is the midpoint of two consecutive points, we have $|t_i - s_\ell| \geq \frac{\Delta}{2} \geq \frac{1}{n}$ for all $i \in [n], \ell \in [r]$. Therefore, as $r \leq 2k$,

$$|p_j(t_i)| = \prod_{\ell=1}^{r} |t_i - s_\ell| \geq \left(\frac{1}{n}\right)^r \geq \left(\frac{1}{n}\right)^{2k} \qquad \text{for all } i \in [n],$$

Now, if the coefficient vector corresponding to $p_j$ is $(u_{j,0}, u_{j,1}, \ldots, u_{j,2k})$, we will set $\tilde{U}_j = (u_{j,0}, u_{j,1}, \ldots, u_{j,2k})$ and $U_j = \tilde{U}_j / \|\tilde{U}_j\|_2$. As the final margin will depend on $\|\tilde{U}_j\|_2$ we proceed to bound the norm.

Because each $s_\ell$ lies in $[-1, 1]$, every coefficient of $p$ is a sum of products of numbers of absolute value at most 1. Hence the coefficient of $s^\ell$ has absolute value at most $\binom{r}{\ell} \leq \binom{2k}{\ell}$, and so

$$\|\tilde{U}_j\|_2^2 = \sum_{\ell=0}^{2k} u_{j,\ell}^2 \leq \sum_{\ell=0}^{2k} \binom{2k}{\ell}^2 = \binom{4k}{2k} \leq \frac{2^{4k}}{\sqrt{3k}}.$$

Altogether, this implies that for every $i, j$, we have that

$$\begin{aligned}(-1)^{A_{ji}+1}\langle U_j, V_i\rangle &= (-1)^{A_{ji}+1} \frac{\langle \tilde{U}_j, \tilde{V}_i\rangle}{\|\tilde{U}_j\|_2 \times \|\tilde{V}_i\|_2} = \frac{(-1)^{A_{ji}+1} p_j(t_i)}{\|\tilde{U}_j\|_2 \times \|\tilde{V}_i\|_2} \\ &\geq \frac{n^{-2k}}{\sqrt{2k+1}\sqrt{2^{4k}/\sqrt{3k}}} = (2k)^{-1/4} 2^{-2k} n^{-2k}.\end{aligned}$$

□

# D Spectral Bound on Margin Complexity

## D.1 Proof of Theorem 1.2

*Proof of Theorem 1.2.* Suppose that $\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n$ is a margin-$m$ relative-bias-0 embedding of $A$. Take any matrix $J \in \mathbb{R}^{N \times n}$ such that $J_{ji} \geq 0$ for all $i, j$ such that $A_{ji} = 1$, $J_{ji} \leq 0$ for all $i, j$ such that $A_{ji} = 0$, and $\sum_{j,i} |J_{ji}| = 1$.

Now, let $\alpha$ be any non-negative number. We have that

$$\begin{aligned} 0 &\leq \|U - \alpha JV\|_F^2 \\ &= \|U\|_F^2 - 2\alpha\langle U, JV\rangle + \langle JV, JV\rangle \\ &= \|U\|_F^2 - 2\alpha\langle UV^T, J\rangle + \alpha^2\langle VV^T, J^TJ\rangle \implies \\ 2\alpha\langle UV^T, J\rangle &\leq \|U\|_F^2 + \alpha^2\langle VV^T, J^TJ\rangle \end{aligned} \tag{3}$$

Now, observe that $\|U\|_F^2 = N$ since $U$ is composed of $N$ unit-norm embeddings. Furthermore, as $VV^T$ is PSD,

$$\langle VV^T, J^TJ\rangle \leq \langle VV^T, \|J^TJ\|_{op} I\rangle = \|J^TJ\|_{op} \times \|V\|_2^2 = \|J\|_{op}^2 \times n.$$

Finally, $(UV^T)_{ji} J_{ji} \geq m|J_{ji}|$ for any $ji$. Indeed, this is the case since if $A_{ji} = 1$, both values are positive and $(UV^T)_{ji} \geq m$. Otherwise, both values are negative and $(UV^T)_{ji} \leq -m$. Altogether, we obtain

$$\begin{aligned} &2\alpha m \sum_{ji} |J_{ji}| \leq N + \alpha^2 \|J\|_{op}^2 \times n \iff \\ &2\alpha m \leq N + \alpha^2 \|J\|_{op}^2 \times n. \end{aligned} \tag{4}$$

We used that $\sum_{ji} |J_{ji}| = 1$. Taking $\alpha = \sqrt{\frac{N}{n\|J\|_{op}^2}}$, we arrive at the desired conclusion. □

### D.2 Applications of Theorem 1.2

We now use Theorem 1.2 to show Corollary 1.

*Proof of Corollary 1.* First, we prove that $\mathsf{m}^{rd}(+\infty, S_{n,k}) \leq \frac{1+o_k(1)}{2\sqrt{k}}$. We will do so by choosing an appropriate matrix $J$ in Theorem 1.2. Concretely, let $y = \frac{1}{2} - \frac{1}{2}\frac{n-2k}{n-2k+2\sqrt{(n-1)k(n-k)}}$ (so $y \approx \frac{1}{2} - \frac{1}{4\sqrt{k}}$), $N = \binom{n}{k}$, and set

$$J_{ji} = \begin{cases} \frac{y}{Nk} \text{ if } (S_{n,k})_{ji} = 1, \\ -\frac{1-y}{N(n-k)} \text{ if } (S_{n,k})_{ji} = 0. \end{cases}$$

One can check that $\sum_{ji} |J_{ji}| = 1$ and $J^T J = I\frac{1}{4Nnk}(1+o_k(1))$. Hence, $\|J\|_{op} = \frac{1}{2\sqrt{Nnk}}(1+o_k(1))$ and we obtain the bound $m \leq \frac{1+o(1)}{2\sqrt{k}}$.

Now, we prove that $\mathsf{m}^{rd}(+\infty, S_{n,k}) \geq \frac{1+o_k(1)}{2\sqrt{k}}$ by exhibiting a construction in dimension $n+1$. Take $V_i = (\sqrt{1-(4k)^{-1/2}}e_i, (4k)^{-1/4}) \in \mathbb{R}^{n+1}$ and $U_j = (\sqrt{1-(4k)^{-1/2}}\sum_{i\,:\,A_{ji}=1}\frac{1}{\sqrt{k}}e_i, -(4k)^{-1/4})$ where $e_i$ form the standard basis. One can easily check that $\langle U_j, V_i\rangle \geq \frac{1+o_k(1)}{2\sqrt{k}}$ whenever $A_{ji} = 1$ and $\langle U_j, V_i\rangle \leq -\frac{1+o_k(1)}{2\sqrt{k}}$ otherwise. Together with the previous result, this settles the margin complexity of $S_{n,k}$ at $\frac{1+o_k(1)}{2\sqrt{k}}$. □

We can also recover the exact bound for $k = 1$ from [BBNP25] using Theorem 1.2.

**Corollary 3.** $\mathsf{m}^{rd}(+\infty, I_n) \leq \frac{n}{3n-4}$.

*Proof.* Set $J = \frac{1}{3n-4}I - \frac{2}{n(3n-4)}11^T$. Then, $\|J\|_{op} = \frac{1}{3n-4}$ from which the bound follows. □

## E Maximal Margin of $S_{n,k}$ in dimension $O(k \log n)$ from Compressed Sensing

We begin by defining the restricted isometry property.

**Definition 2** ([CT05])**.** *Let $M \in \mathbb{R}^{d\times n}$ be a matrix. For $1 \leq k \leq n$ and $\delta_k \in (0,1)$ we say that $M$ satisfies $(\delta_k, k)$ restricted isometry property (RIP) if for any vector $v$ that is at most $k$-sparse,*

$$\|v\|_2^2(1-\delta_k) \leq \|Mv\|_2^2 \leq \|v\|_2^2(1+\delta_k).$$

**Remark 1.** *If $M$ satisfies $(\delta_k, k)$-RIP, it also satisfies $(\delta_m, m)$-RIP for any $m \leq k, \delta_m \geq \delta_k$.*

We now state a theorem of [CT05] in the language of our work.

**Theorem E.1** (Modification of [CT05])**.** *Suppose that the matrix $V = (V_1 V_2 \cdots V_n)$ satisfies $(\delta_{3k}, 3k)$-RIP such that $\delta_{3k} \leq 1/3$. Let $c$ be a vector with support $T$ of size at most $k$. Then, there exists a real vector $U_j$ such that:*

1. *$\langle U_j, V_i\rangle = c_j$ whenever $i \in T$.*
2. *$|\langle U_j, V_i\rangle| \leq \frac{\delta_{3k}}{(1-2\delta_{3k})\sqrt{|T|}}\|c\|_2$ whenever $i \notin T$.*
3. *$\|U_j\|_2 \leq C(\delta_{3k})$ for some constant $C$ that depends only on $\delta_{3k}$.*

*Proof.* Parts 1 and 2 appear directly in [CT05, Lemma 2.2] except that we have replaced orthogonality $\theta$ values by RIP $\delta$ values using [CT05, Lemma 1.2] and used the fact that $V$ also satisfies $(\delta_{3k}, k)$-RIP and $(\delta_{3k}, 2k)$-RIP as in the above remark.

Part 3 follows from the following fact (in the notation of [CT05], $U_j = w$).

$$\|U_j\|_2 = \|w\|_2 = \|\sum_{i=1}^{+\infty}(-1)^{i-1}w_i\|_2 \leq \sum_{i=1}^{+\infty}\|w_i\|_2 \leq K(\delta_{3k})\sum_{i=1}^{+\infty}\left(\frac{\delta_{3k}}{1-\delta_{3k}}\right)^{i-1} \leq 2K(\delta_{3k}),$$

where all inequalities but the last are in [CT05, Lemma 2.2]. The last follows from $3\delta_{3k} \leq 1$. □

Now, to apply this theorem, we need a matrix satisfying strong RIP-properties. It turns out that this is achieved by random Gaussian matrices.

**Proposition 3.** *[[CT05, BDDW08]] Suppose that $d \geq Ck\log(n/k)$ for some sufficiently large constant $C$ and let $\delta = 1/6$. Then, with high probability, the matrix $M \in \mathbb{R}^{d\times n}$ in which each entry is independent $\mathcal{N}(0, 1/d)$ satisfies $(1/6, 3k)$-RIP.*

*Proof.* The proof follows directly from [CT05, Lemma 3.1] by choosing the right parameters. Namely, $S = 3k, m = n, p = Ck\log(n/k)$. Then, $r = S/m = 3k/n$. Hence, for the binary entropy function $H$,

$$\begin{aligned} f(r) &= \sqrt{n/(Ck\log n)}(\sqrt{3k/n} + \sqrt{2H(3k/n)}) \\ &= O\left(\sqrt{n/(Ck\log(n/k))}\sqrt{3k/n}\sqrt{\log(n/3k)}\right) = O(C^{-1/2}). \end{aligned}$$

Therefore, for some large enough $C$, it holds that $f(r) \leq 1/128$. By [CT05, Lemma 3.1] with $\epsilon = 1$,

$$\begin{aligned} &\mathbb{P}[1 + \delta_{3k} \geq (1 + 2f(r))^2] \leq 2\exp(-mH(r)/2) \\ &= 2\exp(-n\Theta(k\log(n/k)/n)) = 2\exp(-\Theta(k\log(n/k))). \end{aligned}$$

Altogether, with high probability $1 - \exp(-\Omega(k\log(n/k)))$, $\delta_{3k} \leq (1 + 1/64)^2 - 1 \leq 1/6$. □

**Corollary 4.** *For some large enough $C$, there exists a margin $\Omega(1/\sqrt{k})$-embedding of $S_{n,k}$ in dimension $Ck\log(n/k)$.*

*Proof.* The proof follows immediately by Theorem E.1 and Proposition 3. Namely, let $\widetilde{V} = (\widetilde{V_1}, \widetilde{V_2}, \ldots, \widetilde{V_n})$ be a random Gaussian matrix of size $d \times n$ with $d = Ck\log(n/k)$. With high probability, the following two properties hold: $\widetilde{V}$ satisfies $(1/6, 3k)$-RIP and each column of $\widetilde{V}$ has norm at most 2 (apply the RIP with $v = e_i$).

As $\widetilde{V}$ satisfies the $(1/6, 3k)$-RIP, then we know there exists $\{\widetilde{U}_j\}_{j=1}^N$ such that $\langle\widetilde{U_j}, \widetilde{V_i}\rangle = \frac{1}{\sqrt{k}}$ if $A_{ji} = 1$ and $|\langle\widetilde{U_j}, \widetilde{V_i}\rangle| \leq \frac{1}{4\sqrt{k}}$ if $A_{ji} = 0$. This follows by setting $c$ to be the vector $c = \frac{1}{\sqrt{k}}A_{j,:}$ in Theorem E.1.

Now, let $\bar{V}_i = (\widetilde{V}_i, (5/8)^{1/2}k^{-1/4}))$ and $\bar{U}_j = (\widetilde{U_j}, -(5/8)^{1/2}k^{-1/4})$. We can easily check that $\langle\bar{U}_j, \bar{V}_i\rangle \geq \frac{3}{8\sqrt{k}}$ if $A_{ji} = 1$ and $\langle\bar{U}_j, \bar{V}_i\rangle \leq -\frac{3}{8\sqrt{k}}$ if $A_{ji} = 0$. Finally, we know that $\|\bar{V}_i\|_2 \leq 2, \|\bar{U}_j\|_2 \leq C(1/6) = O(1)$. Thus, setting $V_i = \frac{\bar{V}_i}{\|\bar{V}_i\|_2}$ and $U_j = \frac{\bar{U}_j}{\|\bar{U}_j\|_2}$, we obtain $\langle U_j, V_i\rangle \geq \Omega(1/\sqrt{k})$ if $A_{ji} = 1$ and $\langle U_j, V_i\rangle \leq -\Omega(1/\sqrt{k})$ if $A_{ji} = 0$. □

**Remark 2.** *We note that Corollary 4 falls short of our main results Theorem 1.4 and Theorem 1.6 in two ways. First, it only provides margin $\Omega(1/\sqrt{k})$ instead of the optimal $(1 + o_k(1))/(2\sqrt{k})$. It might be possible to fix this by more carefully keeping track of the constants. The biggest challenge seems to be the norm of $C(\delta)$ from Theorem E.1.*

*Much more important, however, seems the limitation that in the $k$-sparse setting, only results for margin $\Theta(1/\sqrt{k})$ seem possible from Theorem E.1. Concretely, note that if $c$ is unit norm in Theorem E.1 (which is without loss of generality as the inequalities are homogeneous in $(U, c)$), there exists some $V_i$ such that $\langle U_j, V_i\rangle \leq 1/\sqrt{k}$. Yet, in the current form of the theorem, it might be the case that $\|U_j\|_2 \times \|V_i\|_2 \geq 1$ as evidenced by the case $V = I$ when $U_j = c_j$. Thus, $U_j$ needs to be unit norm. Altogether, $\langle \frac{U_j}{\|U_j\|_2}, \frac{V_i}{\|V_i\|_2}\rangle \leq 1/\sqrt{k}$.*

*To compare, our Theorem 1.4 applies also to $k$-sparse $A$ when a larger margin is possible, say $m = 1/k^{1/4}$. In that case, it will give dimension $O(k^{1/2}\log n)$.*

*At the same time, our Theorem 1.6 gives constructions of smaller margin, say $1/n$ but also in much smaller dimensions. Concretely, $d = \Theta(k^2)$ for margin $1/n$.*

## F Adapting Theorem 1.5 to general $k$-sparse matrices

We note that the same proof method of Theorem 1.5 can be applied to an arbitrary relevance matrix. We first need a modified version of Lemma 3. Towards this end, we need the following definition. Let $\mathcal{S}$ be a family of subsets of $[n]$. We say that $B \subseteq [n]$ is shattered if for any $C \subseteq B$, there exists some $S \in \mathcal{S}$ such that $S \cap B = C$.

**Lemma 5.** *Let $A \in \{0,1\}^{N\times n}$, and let $\{V_i\}_{i=1}^n, \{U_j\}_{j=1}^N \in \mathbb{R}^d$ be a margin-$m$ embedding of $A$. Define the linear map $V : \mathbb{R}^n \to \mathbb{R}^d$ by $Ve_i = V_i$. Suppose that $S \subseteq [n]$ is shattered by the family of row supports*

$$\mathcal{C}_A \coloneqq \big\{\{i \in [n] : A_{ji} = 1\} : j \in [N]\big\}.$$

*Then for any $x \in \mathbb{R}^n$ supported on $S$, it holds that $\|Vx\|_2 \geq m\|x\|_1$.*

*Proof.* By homogeneity, it is enough to prove the claim when $\|x\|_1 = 1$. Since $S$ is shattered, there exists some $j \in [N]$ such that for every $i \in S$, $A_{ji} = 1 \iff x_i > 0$. Therefore,

$$\left\langle U_j, \sum_{i\in S} x_i V_i \right\rangle = \sum_{i\in S} x_i \langle U_j, V_i\rangle \geq m \sum_{i\in S} |x_i| = m.$$

Indeed, if $x_i > 0$, then $A_{ji} = 1$, so $\langle U_j, V_i\rangle \geq m$, while if $x_i < 0$, then $A_{ji} = 0$, so $\langle U_j, V_i\rangle \leq -m$, and hence again $x_i\langle U_j, V_i\rangle \geq m|x_i|$. Since $\|U_j\|_2 = 1$, it follows that

$$\|Vx\|_2 = \Big\|\sum_{i\in S} x_i V_i\Big\|_2 \geq \Big\langle U_j, \sum_{i\in S} x_i V_i\Big\rangle \geq m. \quad \square$$

Then we can use a nearly identical proof as in Theorem 1.5.

**Theorem F.1** (Lower bound for $k$-sparse matrices)**.** *Let $A \in \{0,1\}^{N\times n}$ be such that every row of $A$ has exactly $k$ ones, and let*

$$\mathcal{C}_A \coloneqq \big\{\{i \in [n] : A_{ji} = 1\} : j \in [N]\big\} \subseteq 2^{[n]}$$

*denote the family of row supports. Suppose $\{V_i\}_{i=1}^n, \{U_j\}_{j=1}^N \subset \mathbb{R}^d$ form a margin-$m$ embedding of $A$ with $m > 0$. Fix $\alpha \in (0, 1/2]$ and set $s = \lfloor \alpha k \rfloor$. If $s \geq 1$, then*

$$d \geq \frac{\log |\mathcal{F}_s(A)|}{\log\Big(1 + \frac{2}{m\sqrt{s}}\Big)},$$

*where $\mathcal{F}_s(A) \subseteq \binom{[n]}{s}$ is a family such that for all distinct $T, T' \in \mathcal{F}_s(A)$, $|T \cap T'| < \frac{s}{2}$ and $T \cup T'$ is shattered by $\mathcal{C}_A$.*

*Proof of Theorem F.1.* Let $V$ be as in Lemma 5. Set $s = \lfloor \alpha k \rfloor$. By definition, there exists a family $\mathcal{F}_s(A) \subseteq \binom{[n]}{s}$ such that for any distinct $T, T' \in \mathcal{F}_s(A)$, we have $|T \cap T'| < \frac{s}{2}$ and $T \cup T'$ is shattered by $\mathcal{C}_A$.

For each $T \in \mathcal{F}_s(A)$, define the sign vector $\sigma^T \in \{\pm 1\}^T$ so that $\big\|\sum_{i\in T} \sigma_i^T V_i\big\|_2$ is minimized, and set $y_T \coloneqq \sum_{i\in T} \sigma_i^T V_i$.

To bound $\|y_T\|_2$, let $\epsilon$ be uniformly random in $\{\pm 1\}^T$. Then

$$\mathbb{E}\left\|\sum_{i\in T} \epsilon_i V_i\right\|_2^2 = \mathbb{E}\left[\sum_{i\in T} \|V_i\|_2^2 + \sum_{\substack{i,\ell\in T\\ i\neq\ell}} \epsilon_i\epsilon_\ell\langle V_i, V_\ell\rangle\right] = \sum_{i\in T} \|V_i\|_2^2 = s,$$

since the cross terms vanish in expectation. Because $\sigma^T$ minimizes the norm, it follows that

$$\|y_T\|_2^2 = \Big\|\sum_{i\in T} \sigma_i^T V_i\Big\|_2^2 \leq s.$$

Hence $y_T \in B_d(0, \sqrt{s})$ for all $T \in \mathcal{F}_s(A)$.

Now define $x_T \coloneqq \sum_{i \in T} \sigma_i^T e_i$ so that $Vx_T = y_T$. Consider distinct $T, T' \in \mathcal{F}_s(A)$. Since $\mathrm{supp}(x_T - x_{T'}) \subseteq T \cup T'$ and $T \cup T'$ is shattered by $\mathcal{C}_A$, Lemma 5 gives

$$\|V(x_T - x_{T'})\|_2 \geq m\|x_T - x_{T'}\|_1.$$

Moreover, since every coordinate of $x_T$ and $x_{T'}$ lies in $\{0, \pm 1\}$,

$$\|x_T - x_{T'}\|_1 \geq |T \triangle T'| = |T| + |T'| - 2|T \cap T'| > 2s - s = s.$$

Therefore,

$$\|y_T - y_{T'}\|_2 = \|Vx_T - Vx_{T'}\|_2 = \|V(x_T - x_{T'})\|_2 \geq m\|x_T - x_{T'}\|_1 \geq ms.$$

Thus, the vectors $\{y_T\}_{T \in \mathcal{F}_s(A)}$ are pairwise $ms$-separated in the ball $B_d(0, \sqrt{s})$. When we rescale by $\tilde{y}_T \coloneqq y_T/\sqrt{s}$, we obtain a collection of points in $B_d(0, 1)$ that are pairwise $m\sqrt{s}$-separated. By the standard volume bound (for example [Ver26, Corollary 4.2.11]),

$$|\mathcal{F}_s(A)| \leq \left(1 + \frac{2}{m\sqrt{s}}\right)^d.$$

So, it immediately follows that when $s = \lfloor \alpha k \rfloor$, we have

$$d \geq \frac{\log |\mathcal{F}_{\lfloor \alpha k \rfloor}(A)|}{\log\left(1 + \frac{2}{m\sqrt{\lfloor \alpha k \rfloor}}\right)}$$

which concludes the proof. □

## G Proofs of Propositions on Margin Importance

**Proposition 4** (Compositional Generalization and Margin)**.** *Suppose $\{U_j\}_{j=1}^N$, $\{V_i\}_{i=1}^n$ form a margin-$m$ embedding in $\mathbb{R}^d$ of $S_{n,k}$ with $0 < m < 1$. Then for all $T \subset [n]$ with $k \leq |T| \leq \min(k + \frac{2km}{1-m}, n-1)$, there exists an $h \in \mathbb{S}^{d-1}$ and $c \in \mathbb{R}$ such that for all $i \in T$, it holds that $\langle h, V_i \rangle \geq c$ while for all $j \notin T$, $\langle h, V_j \rangle \leq c$.*

*Proof of Proposition 4.* For a set $S \subseteq [n]$ with $|S| = k$, let $U_S$ denote the vector such that $\langle U_S, V_i \rangle \geq m$ for all $i \in S$ and $\langle U_S, V_j \rangle \leq -m$ for all $j \notin S$. Define

$$h_T \coloneqq \sum_{S \subseteq T,\ |S|=k} U_S.$$

Then

$$\begin{aligned}\langle h_T, V_i \rangle &= \left\langle \sum_{S \subseteq T,\ i \in S,\ |S|=k} U_S + \sum_{S \subseteq T,\ i \notin S,\ |S|=k} U_S, V_i \right\rangle \\ &= \sum_{S \subseteq T,\ i \in S,\ |S|=k} \langle U_S, V_i \rangle + \sum_{S \subseteq T,\ i \notin S,\ |S|=k} \langle U_S, V_i \rangle.\end{aligned}$$

Notice that if $i \in S$, then by definition $\langle U_S, V_i \rangle \geq m$. If $i \notin S$, then since $U_S$ and $V_i$ have unit norm, we have $\langle U_S, V_i \rangle \geq -1$.

Now fix any $i \in T$. Then there are exactly $\binom{|T|-1}{k-1}$ sets of size $k$ containing $i$ and contained in $T$. Also, there are exactly $\binom{|T|}{k} - \binom{|T|-1}{k-1}$ sets of size $k$ contained in $T$ that do not contain $i$.

So it follows that

$$\begin{aligned}
\langle h_T, V_i\rangle &= \sum_{S\subseteq T,\ i\in S,\ |S|=k} \langle U_S, V_i\rangle + \sum_{S\subseteq T,\ i\notin S,\ |S|=k} \langle U_S, V_i\rangle \\
&\geq m\binom{|T|-1}{k-1} + (-1)\left(\binom{|T|}{k} - \binom{|T|-1}{k-1}\right) \\
&= (1+m)\binom{|T|-1}{k-1} - \binom{|T|}{k} \\
&= \frac{(1+m)k}{|T|}\binom{|T|}{k} - \binom{|T|}{k} \\
&= \frac{(1+m)k-|T|}{|T|}\binom{|T|}{k}.
\end{aligned}$$

Furthermore, if $i \notin T$, then $\langle U_S, V_i\rangle \leq -m$ for all $S \subseteq T$ with $|S| = k$. Since there are $\binom{|T|}{k}$ subsets of size $k$ contained in $T$, we get

$$\langle h_T, V_i\rangle = \sum_{S\subseteq T,\ |S|=k} \langle U_S, V_i\rangle \leq -m\binom{|T|}{k}.$$

We also know that, since $m < 1$,

$$\begin{aligned}
k + \frac{2km}{1-m} \geq |T| &\implies (1-m)k + 2mk \geq |T|(1-m) \\
&\implies (1+m)k - |T| \geq -m|T| \\
&\implies \frac{(1+m)k-|T|}{|T|} \geq -m.
\end{aligned}$$

Therefore, for every $i \in T$ and $j \notin T$, we have

$$\langle h_T, V_i\rangle \geq \frac{(1+m)k-|T|}{|T|}\binom{|T|}{k} \geq -m\binom{|T|}{k} \geq \langle h_T, V_j\rangle.$$

Therefore, there exists a real constant $c' \in \mathbb{R}$ satisfying

$$\langle h_T, V_i\rangle \geq c' \geq \langle h_T, V_j\rangle \qquad \text{for all } i \in T \text{ and } j \notin T.$$

Notice that $\|h_T\| \neq 0$ because, since $|T| < n$, for some $i \notin T$ we have $\langle V_i, h_T\rangle \leq -m\binom{|T|}{k} < 0$. So, if we set $h \coloneqq h_T/\|h_T\|$, it follows that there exists an $h \in \mathbb{S}^{d-1}$ and $c = c'/\|h_T\| \in \mathbb{R}$ such that

$$\langle h, V_i\rangle \geq c \qquad \forall i \in T, \quad \text{and} \quad \langle h, V_j\rangle \leq c \qquad \forall j \notin T. \quad \square$$

We note that even a small margin already suffices to force downward generalization. Indeed, once $m \geq \frac{k-1}{2n-k-1}$, the following proposition implies that every nonempty subset with $1 \leq |T| \leq k$ is retrievable. This is related to the notion of a $k$-neighborly polytope [GKPS67]: if $\{V_i\}_{i=1}^n$ forms a $k$-neighborly polytope, then every subset of at most $k$ vertices forms a face and is therefore exposed by a supporting hyperplane. Thus, every such subset can be separated from the remaining vertices and hence can be retrieved.

**Proposition 5** (Downward Generalization and Margin)**.** *Suppose $\{U_j\}_{j=1}^N$, $\{V_i\}_{i=1}^n$ form a margin-$m$ embedding in $\mathbb{R}^d$ of $S_{n,k}$ with $0 < m < 1$. Then for all $T \subset [n]$ with $\max\{1, k - \frac{2m(n-k)}{1-m}\} \leq |T| \leq k$, there exists an $h \in \mathbb{S}^{d-1}$ and $c \in \mathbb{R}$ such that for all $i \in T$, it holds that $\langle h, V_i\rangle \geq c$ while for all $j \notin T$, $\langle h, V_j\rangle \leq c$.*

*Proof of Proposition 5.* The case $|T| = k$ is immediate by taking $h = U_T$ and $c = 0$, so assume $|T| < k$.

We use the same notation as in the proof of Proposition 4. Define

$$h_T \coloneqq \sum_{S\supset T,\ |S|=k} U_S.$$

Then

$$\langle h_T, V_i\rangle = \left\langle \sum_{S\supset T,\ i\in S,\ |S|=k} U_S + \sum_{S\supset T,\ i\notin S,\ |S|=k} U_S, V_i \right\rangle$$
$$= \sum_{S\supset T,\ i\in S,\ |S|=k} \langle U_S, V_i\rangle + \sum_{S\supset T,\ i\notin S,\ |S|=k} \langle U_S, V_i\rangle.$$

Again, if $i \in S$, then by definition $\langle U_S, V_i\rangle \geq m$.

Now fix any $i \in T$. Then there are exactly $\binom{n-|T|}{k-|T|}$ sets of size $k$ containing $T$, and so

$$\langle h_T, V_i\rangle \geq m\binom{n-|T|}{k-|T|}.$$

For $j \notin T$, the number of sets of size $k$ containing $T$ and $j$ is $\binom{n-|T|-1}{k-|T|-1}$, while the number of sets of size $k$ containing $T$ but not $j$ is $\binom{n-|T|-1}{k-|T|}$. Hence, using $\langle U_S, V_j\rangle \leq 1$ when $j \in S$ and $\langle U_S, V_j\rangle \leq -m$ when $j \notin S$, we have

$$\langle h_T, V_j\rangle \leq \binom{n-|T|-1}{k-|T|-1} - m\binom{n-|T|-1}{k-|T|}.$$

Thus it suffices to show

$$\binom{n-|T|-1}{k-|T|-1} - m\binom{n-|T|-1}{k-|T|} \leq m\binom{n-|T|}{k-|T|}.$$

Since

$$\binom{n-|T|}{k-|T|} = \binom{n-|T|-1}{k-|T|-1} + \binom{n-|T|-1}{k-|T|},$$

this inequality is equivalent to

$$m \geq \frac{\binom{n-|T|-1}{k-|T|-1}}{\binom{n-|T|-1}{k-|T|-1} + 2\binom{n-|T|-1}{k-|T|}} = \frac{k-|T|}{k-|T|+2(n-k)} \iff |T| \geq k - \frac{2m(n-k)}{1-m}.$$

Therefore, for every $i \in T$ and $j \notin T$,

$$\langle h_T, V_i\rangle \geq c' \coloneqq m\binom{n-|T|}{k-|T|} \geq \langle h_T, V_j\rangle.$$

Since $c' > 0$, we have $h_T \neq 0$. Taking $h = \frac{h_T}{\|h_T\|}$ and $c = \frac{c'}{\|h_T\|}$ completes the proof. □

**Proposition 6** (Robustness and Margin)**.** *Suppose that the object space $\mathcal{O}$ and the query space $\mathcal{Q}$ are metric spaces with metrics $d_{\mathcal{O}}$ and $d_{\mathcal{Q}}$, respectively. Suppose that the corresponding encoders $f_\theta$ and $g_\phi$ have relative-bias-$\tau$ and margin $m$ on the datasets $\{Q_j\}_{j=1}^N \subset \mathcal{Q}$ and $\{O_i\}_{i=1}^n \subset \mathcal{O}$ with relevance matrix $A \in \{0,1\}^{N\times n}$. If $f_\theta$ and $g_\phi$ are both $L$-Lipschitz for some $L > 0$, then for every $j \in [N]$, every perturbation $Q'_j \in \mathcal{Q}$ satisfying $d_{\mathcal{Q}}(Q'_j, Q_j) < \frac{m}{L}$ preserves perfect retrieval. Namely, $\langle g_\phi(Q'_j), f_\theta(O_i)\rangle > \tau$ if $A_{ji} = 1$ and $\langle g_\phi(Q'_j), f_\theta(O_i)\rangle < \tau$ if $A_{ji} = 0$. Likewise, the same holds for small perturbations of the objects.*

*Proof of Proposition 6.* Since $d_{\mathcal{Q}}(Q'_j, Q_j) < \frac{m}{L}$ and $g_\phi$ is $L$-Lipschitz, we have

$$\|g_\phi(Q'_j) - g_\phi(Q_j)\| \leq L\, d_{\mathcal{Q}}(Q'_j, Q_j) < m.$$

Now fix $i \in [n]$. If $A_{ji} = 1$, then since $f_\theta(O_i)$ has norm 1, we have

$$\left|\langle g_\phi(Q'_j) - g_\phi(Q_j),\ f_\theta(O_i)\rangle\right| \leq \|g_\phi(Q'_j) - g_\phi(Q_j)\|\, \|f_\theta(O_i)\| < m.$$

Therefore,

$$\begin{aligned}\langle g_\phi(Q'_j), f_\theta(O_i)\rangle &= \langle g_\phi(Q'_j) - g_\phi(Q_j), f_\theta(O_i)\rangle + \langle g_\phi(Q_j), f_\theta(O_i)\rangle \\ &> -m + (\tau + m) = \tau.\end{aligned}$$

Similarly, if $A_{ji} = 0$, then

$$\begin{aligned}\langle g_\phi(Q'_j), f_\theta(O_i)\rangle &= \langle g_\phi(Q'_j) - g_\phi(Q_j), f_\theta(O_i)\rangle + \langle g_\phi(Q_j), f_\theta(O_i)\rangle \\ &< m + (\tau - m) = \tau.\end{aligned}$$

□

**Proposition 7** (Maximal-Margin Embeddings and Sigmoid Loss)**.** *Consider the sigmoid loss with inverse-temperature $t > 0$ and bias $b$ over embeddings $\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n$ :*

$$\mathcal{L}_{\textit{sigmoid}}(\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n; t, b) \coloneqq \sum_{j\in[N], i\in[n]} \log\Big(1 + \exp\big((-1)^{A_{ji}}(t\langle U_j, V_i\rangle - b)\big)\Big).$$

*Then*

1. *If $\mathcal{L}_{\textit{sigmoid}}(\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n; t, b) < \log 2$, then there exists $m > 0$ such that $\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n$ is a margin-$m$, relative-bias-$\frac{b}{t}$ embedding of $A$.*
2. *If $\{U_j\}_{j=1}^N$ and $\{V_i\}_{i=1}^n$ form a margin-$m_*$, relative-bias-$\tau$ embedding of $A$ where*

$$m_* \coloneqq \min\left\{\min_{A_{ji}=1}\left(\langle U_j, V_i\rangle - \tau\right), \min_{A_{ji}=0}\left(\tau - \langle U_j, V_i\rangle\right)\right\} > 0.$$

*Then*

$$\lim_{T\to\infty}\frac{1}{T}\log \mathcal{L}_{\textit{sigmoid}}(\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n; T, T\tau) = -m_*.$$

*Proof of Proposition 7.* We begin with part (1). Suppose that $\mathcal{L}_{\text{sigmoid}}(\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n; t, b) < \log 2$. For each fixed pair $(j, i)$, the term $\log\Big(1 + \exp\big((-1)^{A_{ji}}(t\langle U_j, V_i\rangle - b)\big)\Big)$ is nonnegative. So, it follows that for every fixed pair $(j, i)$,

$$\log\Big(1 + \exp\big((-1)^{A_{ji}}(t\langle U_j, V_i\rangle - b)\big)\Big) < \log 2.$$

Since the function $x \mapsto \log(1 + e^x)$ is strictly increasing, we obtain

$$(-1)^{A_{ji}}(t\langle U_j, V_i\rangle - b) < 0.$$

If $A_{ji} = 1$, then $-(t\langle U_j, V_i\rangle - b) < 0$, so $t\langle U_j, V_i\rangle - b > 0$. If $A_{ji} = 0$, then $t\langle U_j, V_i\rangle - b < 0$.
Since there are finitely many pairs, set

$$m \coloneqq \min\left\{\min_{A_{ji}=1}\left(\langle U_j, V_i\rangle - \frac{b}{t}\right), \min_{A_{ji}=0}\left(\frac{b}{t} - \langle U_j, V_i\rangle\right)\right\} > 0.$$

Then $\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n$ form a margin-$m$, relative-bias-$\frac{b}{t}$ embedding of $A$.

We now prove part 2. For each pair $(j, i) \in [N] \times [n]$, define

$$\gamma_{ji} \coloneqq \begin{cases}\langle U_j, V_i\rangle - \tau, & A_{ji} = 1, \\ \tau - \langle U_j, V_i\rangle, & A_{ji} = 0.\end{cases}$$

Then $\gamma_{ji} > 0$ for all $(j, i)$, and by definition $m_* = \min_{j,i}\gamma_{ji}$.
If $A_{ji} = 1$, then

$$(-1)^{A_{ji}}\big(T\langle U_j, V_i\rangle - T\tau\big) = -\big(T\langle U_j, V_i\rangle - T\tau\big) = -T(\langle U_j, V_i\rangle - \tau) = -T\gamma_{ji}.$$

If $A_{ji} = 0$, then

$$(-1)^{A_{ji}}\big(T\langle U_j, V_i\rangle - T\tau\big) = T\langle U_j, V_i\rangle - T\tau = -T(\tau - \langle U_j, V_i\rangle) = -T\gamma_{ji}.$$

Therefore,

$$\mathcal{L}_{\text{sigmoid}}(\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n; T, T\tau) = \sum_{j\in[N],\, i\in[n]} \log\bigl(1 + e^{-T\gamma_{ji}}\bigr).$$

Since $\gamma_{ji} \geq m_*$ for all $(i, j)$, we obtain the upper bound

$$\mathcal{L}_{\text{sigmoid}}(\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n; T, T\tau) \leq Nn \log(1 + e^{-Tm_*}) \leq Nne^{-Tm_*},$$

where we used $\log(1 + x) \leq x$ for all $x \geq 0$. For the lower bound, choose $(j_0, i_0)$ such that $\gamma_{j_0 i_0} = m_*$. Then $\mathcal{L}_{\text{sigmoid}}(\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n; T, T\tau) \geq \log(1 + e^{-Tm_*})$. Also, for $x \in [0, 1]$, $\log(1+x) \geq \frac{x}{2}$. Hence for all sufficiently large $T$ (so that $e^{-Tm_*} \leq 1$), $\log(1+e^{-Tm_*}) \geq \frac{1}{2}e^{-Tm_*}$. Thus for all sufficiently large $T$,

$$\frac{1}{2}e^{-Tm_*} \leq \mathcal{L}_{\text{sigmoid}}(\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n; T, T\tau) \leq Nne^{-Tm_*}.$$

Taking logarithms gives

$$-Tm_* - \log 2 \leq \log \mathcal{L}_{\text{sigmoid}}(\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n; T, T\tau) \leq -Tm_* + \log(Nn).$$

Dividing by $T$ and letting $T \to \infty$, we conclude that

$$\lim_{T\to\infty} \frac{1}{T} \log \mathcal{L}_{\text{sigmoid}}(\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n; T, T\tau) = -m_* \quad \square$$

**Proposition 8** (Large Margins and Quantization)**.** *Suppose that $\{U_j\}_{j=1}^N$, $\{V_i\}_{i=1}^n$ is a margin-$m$ embedding of $A$. Then there exists a codebook $\mathcal{C} \subset \mathbb{S}^{d-1}$ with size $|\mathcal{C}| \leq \left(1 + \frac{8}{m}\right)^d$ such that the quantized embeddings $\tilde{U}_j \in \arg\min_{x\in\mathcal{C}} \|x - U_j\|$ and $\tilde{V}_i \in \arg\min_{x\in\mathcal{C}} \|x - V_i\|$ form a margin-$\frac{m}{2}$ embedding.*

*Proof.* There exists a $\frac{m}{4}$-net $\mathcal{C}$ of size $|\mathcal{C}| \leq \left(1 + \frac{2}{\frac{m}{4}}\right)^d = \left(1 + \frac{8}{m}\right)^d$ [Ver26, Corollary 4.2.11]. For any $j \in [N]$ and $i \in [n]$, we have

$$|\langle \tilde{U}_j, \tilde{V}_i\rangle - \langle U_j, V_i\rangle| = |\langle \tilde{U}_j - U_j, \tilde{V}_i\rangle + \langle U_j, \tilde{V}_i - V_i\rangle|$$

Since $\|\tilde{V}_i\| = \|U_j\| = 1$, we know that $|\langle \tilde{U}_j, \tilde{V}_i\rangle - \langle U_j, V_i\rangle| \leq \frac{m}{2}$ because $\|\tilde{U}_j - U_j\|, \|\tilde{V}_i - V_i\| \leq \frac{m}{4}$. Therefore, the vectors $\{\tilde{U}_j\}_{j=1}^N$, $\{\tilde{V}_i\}_{i=1}^n$ form a margin-$\frac{m}{2}$ embedding of $A$. $\square$

# H InfoNCE and Sigmoid in The Free Embedding Model

## H.1 InfoNCE loss

Proposition 7 characterizes the global minimizers of the sigmoid loss. We give the corresponding characterization for InfoNCE.

Throughout, let $s_{ji} = \langle U_j, V_i\rangle$, $P_j = \{i \in [n] : A_{ji} = 1\}$, and $k_j = |P_j|$. The one-sided loss [OLV18] is

$$\mathcal{L}_{\text{InfoNCE}}(\{U_j\}_{j=1}^N, \{V_i\}_{i=1}^n; t) = -\sum_{j=1}^N \sum_{i\in P_j} \log \frac{\exp(ts_{ji})}{\sum_{\ell=1}^n \exp(ts_{j\ell})}. \tag{5}$$

We first hold the unit vectors $U_j, V_i$ fixed and let $t \to +\infty$. We use the convention $0 \log 0 = 0$.

**Proposition 9** (One-sided InfoNCE)**.** *For every $t > 0$, the loss in (5) is at least $\sum_{j=1}^N k_j \log k_j$. It converges to this bound as $t \to +\infty$ if and only if, for each $j$ with $k_j > 0$,*

$$\{i \in [n] : A_{ji} = 1\} = P_j = \bigl\{i \in [n] : s_{ji} = \max_{\ell\in[n]} s_{j\ell}\bigr\}.$$

*In particular, when every row has exactly $k$ ones, the bound is $Nk \log k$.*

*Proof.* Fix $j$ with $k_j > 0$, and let $p_{ji}(t)$ be the softmax probability in (5). By the arithmetic-geometric mean inequality,

$$\left(\prod_{i \in P_j} p_{ji}(t)\right)^{1/k_j} \leq \frac{1}{k_j} \sum_{i \in P_j} p_{ji}(t) \leq \frac{1}{k_j}.$$

Thus the contribution $L_j(t)$ of row $j$ is at least $k_j \log k_j$. If the row contains a zero, its positive entries have total probability strictly less than one at every finite $t$, so the bound for this row is strict.

Write $M_j = \max_{\ell \in [n]} s_{j\ell}$ and $h_j = |\{\ell : s_{j\ell} = M_j\}|$. Factoring $\exp(tM_j)$ out of the denominator gives

$$L_j(t) = t \sum_{i \in P_j} (M_j - s_{ji}) + k_j \log\left(\sum_{\ell=1}^{n} \exp\left(t(s_{j\ell} - M_j)\right)\right).$$

The second term converges to $k_j \log h_j$. If any positive entry has score below $M_j$, the first term diverges to $+\infty$. Otherwise it vanishes, $h_j \geq k_j$, and the limit equals $k_j \log k_j$ exactly when $h_j = k_j$. Summing over the nonempty rows proves the claim. □

For $A = I$, the condition reduces to a unique maximizer in each row, as in [BBNP25]. For $k_j > 1$, even an embedding with positive margin in the sense of Definition 1 has diverging loss if two positive entries in row $j$ have different scores.

**Symmetric InfoNCE.** Write $Q_i = \{j \in [N] : A_{ji} = 1\}$ and $r_i = |Q_i|$. Let $L_{\text{row}}(t)$ denote (5) and let $L_{\text{col}}(t)$ denote the same loss with $A$ transposed and the roles of $U$ and $V$ exchanged. We take the symmetric loss to be $L_{\text{sym}}(t) = (L_{\text{row}}(t) + L_{\text{col}}(t))/2$.

**Proposition 10** (Symmetric InfoNCE)**.** *For every $t > 0$,*

$$L_{\text{sym}}(t) \geq \frac{1}{2}\left(\sum_{j=1}^{N} k_j \log k_j + \sum_{i=1}^{n} r_i \log r_i\right).$$

*The loss converges to this bound if and only if the positives are exactly the maximizers in every nonempty row and every nonempty column. For $A = S_{n,k}$ and $2 \leq k < n$, this is equivalent to the existence of $c \in \mathbb{R}$ such that*

$$\langle U_j, V_i \rangle = c \quad \text{if } A_{ji} = 1, \qquad \langle U_j, V_i \rangle < c \quad \text{if } A_{ji} = 0.$$

*Proof.* Apply Proposition 9 to $A$ and $A^\top$. Each loss has nonnegative excess above its own lower bound, so the average approaches its bound exactly when both excesses tend to zero.

Consider the bipartite graph on the queries and objects with an edge $(j, i)$ whenever $A_{ji} = 1$. Two edges sharing a query have the same score by the row condition, and two edges sharing an object have the same score by the column condition. Thus the edge score is constant on each connected component. For $A = S_{n,k}$ and $k \geq 2$, any two objects belong to a common $k$-subset, and every query has a neighbor. Hence the graph is connected. All positive scores equal some $c$, and every negative score is strictly smaller by the row condition. Conversely, these inequalities imply both maximizer conditions. □

When $A = S_{n,k}$, every column has $r = \binom{n-1}{k-1}$ ones and $nr = Nk$. The symmetric lower bound is therefore $\frac{Nk}{2}(\log k + \log r)$. For $k = 1$, the graph is a matching and each matched pair may have its own positive score, provided it is the unique maximum in its row and column. Neither loss requires distinct negative entries in a row or column to have equal scores.

### H.2 Experiments

We choose to work with $S_{n,k}$ for $k \in \{2, 3\}$ and $n \in \{20, 40, 60, \ldots, 240\}$ when $k = 2$ and $n \in \{20, 40, 60, \ldots, 120\}$ when $k = 3$. We also carry out experiments for $k \in \{4, 5, 6\}$, but due to computational constraints we take a uniformly random submatrix of $S_{n,k}$ of a given size instead of working the entire $\binom{n}{k}$ rows.

We set the inverse temperature to be trainable for both InfoNCE and sigmoid, as common in practice, e.g. [ZMKB23], and we fix the bias to $b = 0$ for sigmoid loss. We initialize the embeddings $\{U\}_{j=1}^{N}, \{V\}_{i=1}^{n}$ as random i.i.d. points on $\mathbb{S}^{d-1}$. Then, we run the Adam optimizer with cosine annealing learning rate over $\mathcal{L}_{\text{sigmoid}}, \mathcal{L}_{\text{InfoNCE}}$ for 100000 steps to obtain the embeddings. We embed in each of the dimensions $d \in \{5, 6, 7, \ldots, 30\}$. We start from 5 since this is the minimal dimension for which an embedding of $S_{n,k}$ with $k = 2$ exists, see Theorem 1.1.

We ran our experiments on a single H200 GPU. The runs for each triplet $n, d, k$ took up to one hour (for our largest experiment, $n = 120, k = 3, d = 30$).

Then, we plot in Figure 1 the smallest dimension for which a positive margin was achieved when $k = 2$ and $n$ varies. This smallest dimension is an effective sign rank for the respective loss.

In Figure 2 we choose three dimensions and plot the respective margins obtained when $k = 2$ for the varying $n$. We reproduce the same results with $k = 3$ and obtain qualitatively similar behavior:

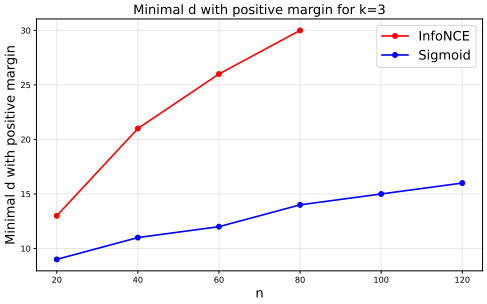



Figure 4: Minimal dimension needed to achieve a non-zero margin after 100000 training steps for the InfoNCE and sigmoid losses. The sigmoid succeeds in much smaller dimensions.

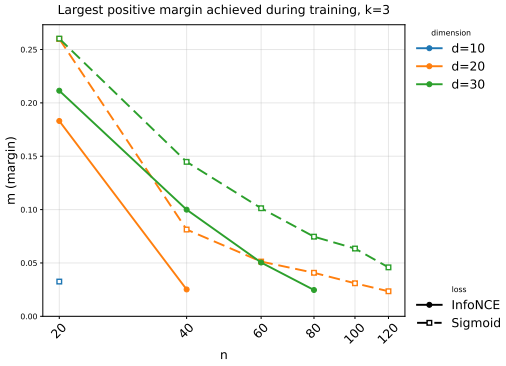



Figure 5: Maximal margin achieved during 100000 training steps for $d \in \{10, 20, 30\}$, using the InfoNCE and sigmoid losses showing clear advantage of the sigmoid loss.

The results for $k \in \{4, 5, 6\}$ and a subsampled adjacency matrix are qualitatively similar.

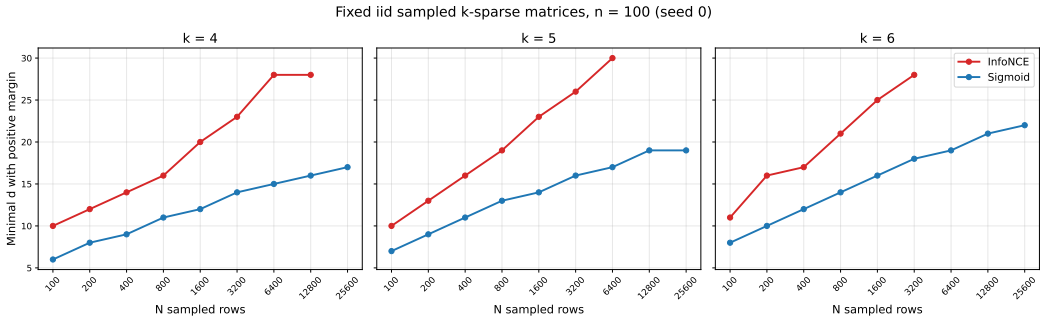



Figure 6: Minimal dimension needed to achieve non-zero margin after 100000 steps for the InfoNCE and Sigmoid losses. We fix $n = 100$ and vary the number of rows we subsample from $S_{n,k}$.

**Comparing The Geometries of Sigmoid and InfoNCE losses.** Finally, we experimentally check Proposition 9. Concretely, we measure the average variance across queries $j$ for of $\langle U_j, V_i \rangle$ for which $A_{ji} = 1$. According to Proposition 9, this variance should be zero for InfoNCE. Indeed, experiments show that the variance is much lower for InfoNCE than it is for the sigmoid loss.

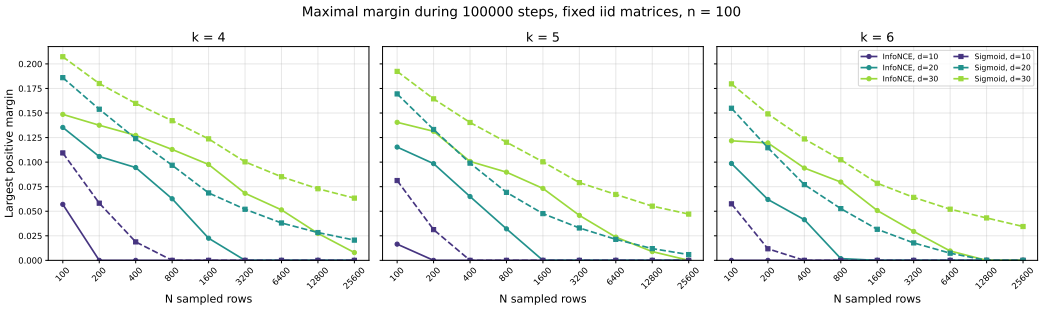



Figure 7: Maximal margin achieved during 100000 training steps for $d \in \{10, 20, 30\}$, using the InfoNCE and sigmoid losses. We fix $n = 100$ and vary the number of rows we subsample from $S_{n,k}$.

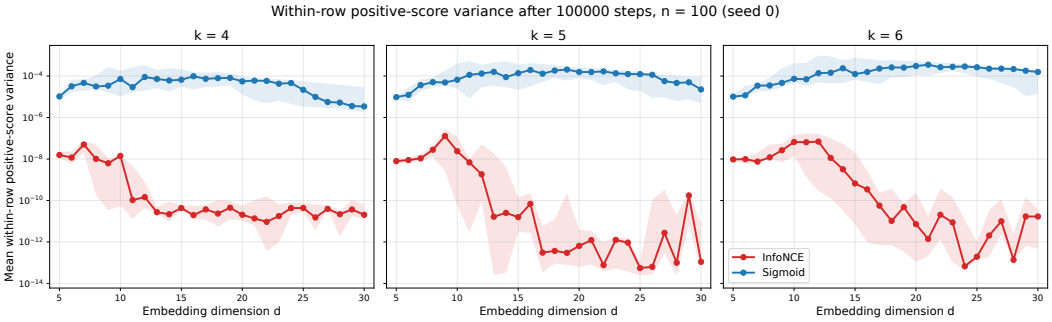



Figure 8: $\mathbb{E}_j[\mathbf{Var}[\langle U_j, V_i\rangle | A_{ji} = 1]]$ for embeddings produced by InfoNCE and Sigmoid losses. The solid lines are the means and the shade the standard deviation over 9 distinct runs. According to Proposition 9, this variance should be 0 for InfoNCE, while no such pressure exists for the Sigmoid loss according to Proposition 7. The experiments support the much lower variance in the case of InfoNCE.

# I LLM Usage

We used LLMs for completing some proofs, finding related works, generating code, and editing. We wrote the initial manuscript without any LLM usage and only then used LLMs for editing. We outline the theorem-proving process in more detail, as it could be valuable to the community.

**Proofs:** We used ChatGPT Pro 5.4 - Extended for proving some theorems. Here is the breakdown of the process, with prompts. We format the prompts in LaTeX and clean up typos, but otherwise there are no changes.

1. Corollary 4: We wanted to see if we could find a matching construction for the $O(k \log(n/k)/\log(k))$ lower bound on the dimension necessary for $\Theta(k^{-1/2})$ margin from [WBNL26]. We prompted ChatGPT with the following prompt to obtain Corollary 4:

> Hi! I want a construction for the following problem. Given is the following bipartite graph parametrized by integers $n$ and $k$. On the left side, there are n vertices. On the right side, there are $N = \binom{n}{k}$ corresponding to the subsets of $[n]$ of size $k$. For each subset of $k$ vertices on the left side, there is a unique vertex on the right side adjacent exactly to them. Now, I want to produce n unit vectors in dimension $d$ $V_1, V_2, \ldots, V_n$ for the vertices on the left and $N$ unit vectors in dimension $d$ $U_1, U_2, \ldots U_N$ for the vertices on the right such that $\langle V_i, U_j\rangle > \Omega(1/\sqrt{k})$ if and only if $(i, j)$ is an edge and $\langle V_i, U_j\rangle < -\Omega(1/\sqrt{k})$ otherwise. Give me such a construction in dimension $d = O(k \log n)$.

2. Theorem 1.4: We knew how to prove the weaker result for $d = O(\mathsf{m}^{rd}(+\infty, A)^{-2} \log(nN))$ using the Johnson-Lindenstrauss lemma. As stated, however, this weaker result only implies

dimension $O(k^2 \log n)$ for margin $\Theta(k^{-1/2})$-embeddings of $S_{n,k}$. We already knew that it is possible to improve this to $d = O(k \log n)$ via Corollary 4. We conjectured that Theorem 1.5 should be true. We spent some time devising a proof strategy and submitted the following sequence of prompts to ChatGPT. The first one is to check if it can come up with the result we already had:

> You are an expert researcher at graph theory, random matrix theory, compressed sensing! Consider the following problem.
> Given is the following bipartite graph parametrized by integers $n$ and $k$. On the left side, there are n vertices. On the right side, there are $N = \binom{n}{k}$ corresponding to the subsets of $[n]$ of size $k$. For each subset of $k$ vertices on the left side, there is a unique vertex on the right side adjacent exactly to them. Now, I want to produce n unit vectors $V_1, V_2, \ldots, V_n$ for the vertices on the left and $N$ unit vectors $U_1, U_2, \ldots U_N$ for the vertices on the right such that $\langle V_i, U_j \rangle > m$ if and only if $(i, j)$ is an edge and $\langle V_i, U_j \rangle < -m$ otherwise for some $m > 0$.
> Prove that there exist $n$ unit vectors in dimension $d$ $V'_1, V'_2, \ldots, V'_n$ for the vertices on the left and $N$ unit vectors in dimension $d$ $U'_1, U'_2, \ldots U'_N$ for the vertices on the right such that $\langle V_i, U_j \rangle > m/2$ if and only if $(i, j)$ is an edge and $\langle V_i, U_j \rangle < -m/2$ otherwise, where $d = O(\log(Nn)/m^2)$.

After it succeeded with this task, we prompted for a variant of Theorem 1.4.

> can you attempt to now make a construction in $O(m^{-2} \log n)$ even if $n \ll N$. Perhaps you need to consider the convex program for fixed $V_1, V_2, \ldots, V_n$ that guarantees the existence of $U_1, U_2, .., U_N$ with margin $m$. You know that this convex program for the given $V_1, V_2, \ldots, V_n$ has a solution. You can reduce the dimension of $V_1, V_2, \ldots, V_n$ via JL or some other projection to $d = O(m^{-2} \log n)$ and then consider the convex program again. Using duals may be useful.

This gave us essentially the proof of Theorem 1.4 with some minor bugs which we fixed.

3. Theorem 1.5: We asked ChatGPT to prove a bound of $d = \Omega(k \log n)$ for margin $\Theta(k^{-1/2})$ in the $S_{n,k}$ case. We knew that $d = O(k \log n)$ is sufficient in light of our main Theorem 1.4. This time, we generated the prompt with ChatGPT. The final prompt was:

> Problem
> Let $N = \binom{n}{k}$. For each subset $S \subset [n]$ with $|S| = k$, assign a unit vector $U_S \in \mathbb{S}^{d-1}$, and for each $j \in [n]$, assign a unit vector $V_j \in \mathbb{S}^{d-1}$, such that for some margin $m > 0$:
>
> $$\langle U_S, V_j \rangle \geq m \quad \text{if } j \in S, \qquad \langle U_S, V_j \rangle \leq -m \quad \text{if } j \notin S.$$
>
> Goal
> Upper bound $N = \binom{n}{k}$ in terms of $d$ and $m$.
> Your goal is to get matching upper and lower bounds for N Currently we have up to multiplicative constants that $d \geq k \log(n)/\log(1/m)$. We have $d$ is at least $k \log n/\log(k)$ and at most $k \log n$. So there is a gap. Try to fill it in.

4. Theorem 1.6: Next, we wanted constructions for a large margin when $d = o(k \log(n/k))$. Interestingly, prompts like the following one (preceded by the conversation for Theorem 1.4) failed:

> Suppose that I want to achieve a margin $= 1/n$ in dimension smaller than $o(k \log n)$. How can I do this?

ChatGPT would always either give the construction in dimension $2k + 1$ which has margin $n^{-\Omega(k)}$ [AFR85] and Corollary 2; or the construction where $V_1, V_2, \ldots, V_n$ are i.i.d. on the sphere which requires dimension $\Omega(k^2 \log n)$ [WYL+26].
What made ChatGPT succeed is giving it a much more concrete (even if imprecise) goal:

> Can you try to obtain margin $1/n$ in dimension $k\sqrt{\log n}$.

It produced the current Theorem 1.6 in the special case for $d = \Theta(k^2)$ and margin $1/n$, acknowledging that it only works when $k = o(\sqrt{\log n})$. Then, we asked ChatGPT to generalize this construction which is a rather mechanical change.

Interestingly, similar versions of the prompt which are, however, too strong and ask for a potentially incorrect result failed, e.g:

| Try to find a construction in dimension $O(k)$ with margin $1/n$. |
|---|

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We fully prove the stated theoretical results on $\mathsf{m}^{rd}(d, A)$ in Theorems 1.4 and 1.5. The experimental details are provided in Part 4 of Section 1.4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We prove our three main claims with all stated asumptions in the following sections. The proof of Theorem 1.4 is in Section 2, of Theorem 1.5 in Section 3, and of Theorem 1.6 in Section 4.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our experiments are described in Part 4 of Section 1.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: github/TopK.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: In Part 4 of Section 1.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Not applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research process does not involve experiments with human data. We are not aware of any direct negative impact of work.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: We are not aware of any potential negative impact of our work. We state this clearly in Section 5.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not utilize any real-world data or models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We do not borrow any code directly. Our experiments are, however, similar to the ones in [WBNL26] and we acknowledge this explicitly in Part 4 of Section 1.4.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: no new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: In Section 5 and Appendix I.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.