OPTIMAL VC DIMENSION OF CONTRASTIVE LEARNING WITH MARGIN

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive learning is a successful paradigm for learning d-dimensional geometric representations from a collection of "anchor–positive–negative" triplets (i, j^+, k^-), indicating that "item i is closer to j than to k." Despite its success, understanding why contrastive learning leads to representations of high *generalization* quality—beyond the often pessimistic predictions from PAC-learning—remains a central question. Recently, Alon, Avdiukhin, Elboim, Fischer, and Yaroslavtsev (2024) proved that, for PAC-learning d-dimensional Euclidean representations of n-point datasets, $\Theta(\min(nd, n^2))$ triplets are necessary and sufficient, while they posed as an open question whether their VC dimension bounds for the more realistic setting of *contrastive learning with a margin* can be improved. For a margin parameter $\alpha > 0$, a triplet $(i, j^+, k^-)_\alpha$ is satisfied by the embedding $\phi : [n] \to \mathbb{R}^d$, if $\|\phi(i)-\phi(k)\|_2 > (1+\alpha)\cdot\|\phi(i)-\phi(j)\|_2$. In this work, we resolve their question by proving that the VC dimension of contrastive learning under any margin $\alpha \in (0, 1)$ is in fact $O(n/\alpha^2)$, improving on the previous bound of $O(n \log(n)/\alpha^2)$. We also establish that the bounds are optimal up to constant factors, by providing a matching lower bound of $\Omega(\frac{n}{\alpha^2})$ (the previously known lower bound was $\Omega(\frac{n}{\alpha})$), for $\alpha \geq \max(n^{-1/2}, d^{-1/2})$.

1 INTRODUCTION

Contrastive learning is a powerful paradigm in representation learning, where the goal is to find meaningful representations, encoded as features in d-dimensional Euclidean space, allowing us to replace data point x by its feature vector $\phi(x) \in \mathbb{R}^d$ in various downstream tasks. At the heart of contrastive learning lies the notion of a *triplet* comparison (i, j^+, k^-) indicating that "item i is closer to j than to k," also known as "anchor–positive–negative" (Schroff et al., 2015). For example, modern pipelines are trained with a contrastive loss, e.g., Triplet or InfoNCE (with one or more negatives), mapping data into d-dimensional vectors by pulling anchor-positive pairs closer, while pushing anchor-negative pairs further apart, enabling a unified approach for diverse types of data, including text, audio, graph-based tasks, and images (Schroff et al., 2015; Grover & Leskovec, 2016; Oord et al., 2018; Saunshi et al., 2019; Chen et al., 2020).

Unfortunately, despite its empirical success, many theoretical aspects of contrastive learning are not well-understood. In this work, we focus on *sample complexity* and *generalization bounds*—a central question in PAC-learning—making progress towards understanding when a representation learned from anchor–positive–negative samples generalizes well to unseen (test) samples.

1.1 SAMPLE COMPLEXITY OF CONTRASTIVE LEARNING

Following the framework by (Alon, Avdiukhin, Elboim, Fischer, and Yaroslavtsev, 2024), let $\mathcal{X}$ be an n-point dataset and let $\phi : \mathcal{X} \to \mathbb{R}^d$ be an embedding that maps points into d-dimensional Euclidean space (we do not make any assumptions on ϕ). Motivated by the anchor–positive–negative paradigm, typically we assume access to a collection of m triplets; these are usually generated from data, such as nearby words in a document, or neighboring video frames, that serve as the anchor-positive pair, and then a random other data point plays the role of the negative example. For more on how to pick negative samples, and their role in downstream performance, please refer to Saunshi et al. (2019); Robinson et al. (2020); Awasthi et al. (2022).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Definition 1.1. *We say that the embedding* ϕ satisfies *or* is consistent with *input triplet* (i, j^+, k^-) *if:*

$$||\phi(i) - \phi(k)||_2 > ||\phi(i) - \phi(j)||_2.$$

More generally, it is common to enforce a small margin in the above distance comparison Vankadara et al. (2023), i.e., for a margin parameter $\alpha > 0$*, the embedding* ϕ *satisfies the triplet* $(i, j^+, k^-)_\alpha$ *if:*

$$||\phi(i) - \phi(k)||_2 > (1 + \alpha)||\phi(i) - \phi(j)||_2.$$

The basic question motivating our work concerns the sample complexity of contrastive learning: *for a set of* n *items, what is the smallest number* $m^*(\varepsilon, \delta)$ *of anchor–positive–negative samples required to PAC-learn an underlying ground-truth embedding* $\phi(x) \in \mathbb{R}^d$*?*

In particular, we focus on the following problem in PAC-learning studied by Alon, Avdiukhin, Elboim, Fischer, and Yaroslavtsev (2024):

Problem 1.2 (PAC-Learning an embedding $\phi(\cdot)$ from triplets Alon et al. (2024))**.** *Fix error parameters* $\varepsilon, \delta > 0$*. Given i.i.d. samples* (i, j, k) *drawn from a distribution* $\mathcal{D}$ *over* $[n]^3$*, and labeled according to embedding* ϕ *as either* (i, j^+, k^-) *or* (i, k^+, j^-)*, based on whether* $||\phi(i)-\phi(j)||_2 \lesseqgtr ||\phi(i)-\phi(k)||_2$*, the goal is to find an embedding* $\hat{\phi}$[1] *that generalizes well to unseen triplets, i.e., with probability* $1 - \delta$*, for a subsequent sample* (i, j, k) *sampled from* $\mathcal{D}$*:*

$$\mathbb{P}_{(i,j,k)\sim\mathcal{D}}[\ \hat{\phi} \text{ disagrees with } \phi \text{ on } (i, j, k)] \leq \varepsilon.$$

In other words, for a test triplet query (i, j, k), the learned embedding $\hat{\phi}$ must classify whether the anchor i is closer to j or to k according to the (unknown) ϕ, i.e., whether the triplet is (i, j^+, k^-) or (i, k^+, j^-). Similarly, Problem 1.2 can be studied in the agnostic case (where samples are not necessarily consistent with some embedding), or with quadruplet samples (i, j, k, l) indicating whether $||\phi(i) - \phi(j)||_2 \lesseqgtr ||\phi(k) - \phi(l)||_2$ from the literature of ordinal embeddings Bilu & Linial (2005); Alon et al. (2008); Terada & Luxburg (2014); Vankadara et al. (2023).

Dimensionality-vs-Generalization. Recently, Alon, Avdiukhin, Elboim, Fischer and Yaroslavtsev (Alon et al., 2024) showed that the VC-dimension of the learning task in Problem 1.2 is $O(\min(nd, n^2))$, i.e., they proved that $m^*(\varepsilon, \delta) = O_{\varepsilon,\delta}(\min(nd, n^2))$ labeled triplets are sufficient[2] for PAC-learning the embedding $\phi : \mathcal{X} \to \mathbb{R}^d$. Typically, ϕ stems from the output layer of a neural network, so they focused on the representation dimensionality d, a key parameter of interest often ranging in the thousands Radford et al. (2018); Kusupati et al. (2022); Comanici et al. (2025).

1.2 Our Contribution: Generalization from $O(n)$ samples, under *any* margin

For the special case when the margin $\alpha = 0$ between distances $||\phi(i) - \phi(j)||_2, ||\phi(i) - \phi(k)||_2$, Alon et al. (2024) managed to show a matching lower bound of $\Omega(\min(nd, n^2))$ for Problem 1.2. However, it is unclear what happens for the well-motivated case of contrastive learning with margin $\alpha > 0$. Observe that $\alpha > 0$ provides some gap between distances, so the dimension d can be directly replaced by $O(\log n/\alpha^2)$ using standard tools, e.g., Johnson-Lindenstrauss (JL) lemma: as noted in (Alon et al., 2024), JL yields $d = O(\log n/\varepsilon^2)$, where ε is the distortion Dasgupta & Gupta (2003), so $O(\frac{n \log n}{\alpha^2})$ triplets suffice for generalization (set distortion $\varepsilon \approx \alpha$). Improving this bound was raised as an open question by Alon et al. (2024):

For contrastive learning with large enough values of margin $\alpha > 0$*,*
can the VC-dimension bound be improved beyond $O(\frac{n \log n}{\alpha^2})$*?*

Our Main Result. We go beyond the above approach based on JL, and we obtain optimal $O(n)$ bounds in the VC dimension, thus resolving the above question for *all* (constant) values of margin $\alpha > 0$ (not necessarily large values as originally stated in Alon et al. (2024)).

Our main result is a stronger, linear in n VC-bound of $O(n/\alpha^2)$ for contrastive learning. This is coupled with a matching lower bound of $\Omega(\frac{n}{\alpha^2})$, so when put together the two results give a tight characterization of the VC dimension in this setting. To make things formal, we need the following definition related to the notion of *partial* concept classes from Alon et al. (2022):

[1] In the contrastive learning with margin setting studies below, learning might require an improper predictor.

[2] The bound is $O(nd)$, since $d \leq n$ without affecting pairwise distance comparisons (Bilu & Linial, 2005).

Definition 1.3 (Concept class $\mathcal{H}(n, d, \alpha)$ for embeddings with margin)**.** *Fix $\alpha > 0$, and let $\phi : [n] \to \mathbb{R}^d$ be an embedding. Associated with embedding $\phi(\cdot)$ is its triplet classifier $h_\phi : [n]^3 \to \{+1, -1, \star\}$, labeling[3] an input triplet as follows:*

$$h_\phi(i,j,k) = \begin{cases} +1, & \text{if } \|\phi(i) - \phi(k)\|_2 > (1+\alpha)\|\phi(i) - \phi(j)\|_2, \\ -1, & \text{if } \|\phi(i) - \phi(j)\|_2 > (1+\alpha)\|\phi(i) - \phi(k)\|_2, \\ \star, & \text{otherwise.} \end{cases}$$

The concept class containing all possible triplet classifiers is denoted[4] as $\mathcal{H}(n, d, \alpha)$. We only consider embeddings such that for every i, j, k with $j \neq k$, $\|\phi(i) - \phi(j)\| \neq \|\phi(i) - \phi(k)\|$.

Theorem 1.4 (Linear VC-dimension for *any* margin)**.** *The VC-dimension of $\mathcal{H}(n, d, \alpha)$ is $O(\frac{n}{\alpha^2})$ for $\alpha \in (0, 1)$ and $O(n)$ for $\alpha \geq 1$. Specifically, there exists absolute constant C such that:*

$$\mathrm{VCdim}(\mathcal{H}(n, d, \alpha)) \leq C \max\Big\{\frac{n}{\alpha^2}, n\Big\}.$$

Furthermore, assuming that $\alpha \geq \max(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{d}})$, the VC dimension is $\Omega(\frac{n}{\alpha^2})$, for $\alpha \leq 1$ and $\Omega(n)$ for $\alpha > 1$. In particular there exists an absolute constant c such that:

$$\mathrm{VCdim}(\mathcal{H}(n, d, \alpha)) \geq c \cdot \max\{\frac{n}{\alpha^2}, n\}.$$

We prove the upper bound of Theorem 3.1 in Section 3, the proof of the lower bound follows the construction for linear classification with margin and is proven in Appendix B.

A few remarks are in order. First, our result characterizes the VC-dimension of concept class $\mathcal{H}(n, d, \alpha)$ in the regime where $\alpha \geq \max(n^{-1/2}, d^{-1/2})$ as $\mathrm{VCdim}(\mathcal{H}(n, d, \alpha)) = \Theta(n \cdot \min(\max(1, \frac{1}{\alpha^2}), n, d))$. Second, it is surprising that under a margin α, dimensionality does not play a role: the larger the margin, the easier the learning task becomes, since the sought-after embedding ϕ is less refined, and a priori, one might expect that dimension d plays a role on the samples required to avoid misclassification of test samples (perhaps this is why Alon et al. (2024) originally asked the question for large margins as a simplification).

Learning Bounds via Geometric Approximation. A novel aspect in our analysis that allows us to obtain the $O(n)$ VC-bound, is the use of geometric approximation algorithms aiming to satisfy as many of the input triplets as possible. In contrast to the approach taken in Alon et al. (2024), where they used combinatorial bounds on the number of sign-patterns of system of polynomials, here we essentially reduce the d-dimensional learning problem to a 1-dimensional setting, whose sample complexity we can bound. On the one hand, we exploit the margin $\alpha > 0$ to show that a simple one-dimensional projection satisfies $\frac{1}{2} + \Omega(\alpha)$ fraction of triplets, independently of dimension d, where $\frac{1}{2}$-approximation is a trivial baseline achieved by a random embedding. On the other hand, we use the probabilistic method to show that no set of size $m = \Omega(\frac{n}{\varepsilon^2})$ can be shattered on the line $d = 1$, and more precisely, there exists a labeling of the triplet samples such that no embedding satisfies more than $\frac{1}{2} + \varepsilon$ of the triplets; since ε can be taken to be sufficiently smaller than α this leads to a contradiction, giving us Theorem 1.4. For formal definitions and a detailed technical overview see Section 2 below.

1.3 Further Related Work

There are intense efforts to understand various aspects of contrastive learning, both from a theoretical and a practical perspective. For example, many well-known works study relations between transfer learning, inductive biases and contrastive learning (Saunshi et al., 2019; 2022), posing specific assumptions on latent classes or how contrastive triplets are generated, in order to try to explain the observed success of such methods (Schroff et al., 2015; Oord et al., 2018; Chen et al., 2020). Other

[3] Symbol $\star$ is arbitrary, meaning classifiers may not answer certain queries (e.g., if distances are not separated).

[4] For the case of no margin, we simply write $\mathcal{H}(n, d)$. Note that for embeddings ϕ, as in the definition, $\mathcal{H}(n, d)$ is a total concept class.

works focus more on optimization loss functions such as spectral contrastive loss with good downstream performance (HaoChen et al., 2021), understanding the role of negative samples (Robinson et al., 2020; Awasthi et al., 2022; Lei et al., 2023), adversarial training Zou & Liu (2023) and more. Here, the work closest to ours is Alon et al. (2024), which served as the main inspiration towards our sample complexity results.

Moreover, the literature of non-metric embeddings, often called *ordinal* embeddings or *monotone* maps (Bilu & Linial, 2005; Alon et al., 2008; Terada & Luxburg, 2014; Emamjomeh-Zadeh & Kempe, 2018; Ghosh et al., 2019; Ghoshdastidar et al., 2019; Fan et al., 2020; Avdiukhin et al., 2023; Chatziafratis & Indyk, 2024; Avdiukhin et al., 2024), is also closely related to our setting, as their main algorithmic task is, given a collection of comparisons between pairwise distances, to find an embedding into a suitable geometric space, e.g., tree metric, ultrametric, or Euclidean space, so that as many of the given pairwise comparisons are satisfied. We note that, because humans are much faster at comparisons between objects, in contrast to providing numerical scores, such triplet comparisons are also widely-deployed in metric learning and crowdsourcing (Schultz & Joachims, 2003; Tamuz et al., 2011; Kulis, 2013). Other types of comparisons include the quadruplet setting (i, j, k, l) mentioned earlier, where comparisons are of the form "item i is closer to j, than item k is to l," see survey Vankadara et al. (2023).

Finally, there are several results on the computational complexity and approximation algorithms literature associated with the empirical risk minimizer on a given sample. Given m (non-contradictory) triplet constraints on n items, Bilu & Linial (2005) shows that $n-1$ dimensions always suffice to satisfy them, while $\Omega(n)$ are necessary. In Avdiukhin et al. (2024), they present an algorithm that achieves dimension $\sqrt{m}$ and further improve it via the notion of graph arboricity. For the case of non-realizable instances, recent work by Arvanitakis et al. (2026) showed that it is computational intractable to obtain a better than $\frac{1}{2}$-approximation (assuming the Unique Games Conjecture (Khot, 2002)), showing that the problem of maximizing satisfied triplets is approximation resistant (Håstad, 2001; Hast, 2005; Guruswami et al., 2008).

2 Technical Overview

We present the main proof strategy and technical ideas behind our improved $O(n/\alpha^2)$ bound on the VC-dimension for the concept class $\mathcal{H}(n, d, \alpha)$ of triplet classifiers with margin, and in the next Section 3, we give the proof of Theorem 1.4.

2.1 Our approach for $\alpha > 0$: Algorithmic Advantage over Random

To give some intuition and set notation, we first describe the approach taken in Alon et al. (2024) for the case of zero margin, and why it fails in our case when margin $\alpha > 0$.

Sign-Patterns of Polynomials vs. margin $\alpha > 0$. Given that each triplet (i, j^+, k^-) imposes a *geometric* constraint on how the representation ϕ embeds the n data points, a first attempt would be to analyze how ϕ partitions d-dimensional Euclidean space, and count the regions wherein triplet classifiers do not switch their label. Indeed, Alon et al. (2024) observed that each triplet is satisfied, if and only if the following polynomial $Q_{(i,j,k)}$ on $3 \cdot d$ number of variables, and of degree 2 is positive:

$$Q_{(i,j,k)}(\phi(i), \phi(j), \phi(k)) = \|\phi(i) - \phi(k)\|_2^2 - \|\phi(i) - \phi(j)\|_2^2$$

In total, this leads to m such polynomials (as there are m input triplets), with the total number of variables being $n \cdot d$. Then, Alon et al. (2024) used known results from algebraic combinatorics to upper bound the different sign-patterns of the resulting system $\left\{Q_{(i_t,j_t,k_t)}\right\}_{t=1}^{m}$ of m polynomials on $n \cdot d$ variables (Warren, 1968; Alon et al., 1985). This limits the size of a set that can be shattered giving an upper bound on the VC-dimension (see definitions below).

While the above approach worked for the case of no margin $\alpha = 0$, and yielded the bound $\Theta(nd)$, a similar approach for $\alpha > 0$ would still yield the same (dimension-dependent) bound $\Theta(nd)$: notice that if instead we defined the polynomial

$$Q'_{(i,j,k)}(\phi(i), \phi(j), \phi(k)) = \|\phi(i) - \phi(k)\|_2^2 - (1+\alpha)^2\|\phi(i) - \phi(j)\|_2^2,$$

the number of sign-patterns does not change, as those depend solely on m, the degree of the polynomials and the number of variables.

We take a different approach to exploit the margin through geometric approximation algorithms that estimate triplet labels induced by $\phi(i), \phi(j), \phi(k) \in \mathbb{R}^d$. On the one hand, the margin allows a 1-dimensional random projection algorithm to get slightly better-than-random approximations, specifically satisfying $\approx \frac{1}{2} + \Omega(\alpha)$ fraction of the triplets. This is crucially *independent* of what the dimension d is, and is the reason why our final bound in Theorem 1.4 does not have a dependence on d. On the other hand, if the VC-dimension was substantially larger than $O(n)$, and the labels of the training set were chosen at random to be either $+1$ or -1, then no embedding could fit the labels of the input triplets, better than random chance, which leads to a contradiction.

2.2 Rounding-vs-Noisy Labels in 1D

To make the above precise, we remind the reader of the standard definitions of *VC-dimension* and *Shattering* of a set (Shalev-Shwartz & Ben-David, 2014), adapted to our setting. The appropriate framework for our results is that of Alon et al. (2022), where hypothesis functions can be *undefined* on certain inputs (this corresponds to the asterisk case in Definition 1.3).

Definition 2.1 (Shattering). *We say that a set $S = \{(i_t, j_t, k_t)\}_{t=1}^m$ is shattered by $\mathcal{H}(n, d, \alpha)$ if for every $y : [m] \to \{-1, 1\}$ there exists a concept function $h_\phi \in \mathcal{H}(n, d, \alpha)$ such that for every $t \in [m]$, $h_\phi(i_t, j_t, k_t) = y(t)$.*

Notice that in the definition, the concept is not allowed to output the special symbol $\star$ on any of the points. We can now give the definition of the VC dimension of the hypothesis class $\mathcal{H}(n, d, \alpha)$.

Definition 2.2 (VC-dimension). *The VC-dimension of hypothesis class $\mathcal{H}$ is the largest number m such that there exists a set of size m that is shattered. We will use $\mathrm{VCdim}(\mathcal{H})$ to denote the VC dimension of hypothesis class $\mathcal{H}$.*

We directly apply the definition of the VC dimension and shattering to prove our main theorem. What we need to show is that for $m \approx \frac{n}{\alpha^2}$, no set of m unlabeled triplets $\{(i_t, j_t, k_t)\}_{i=1}^m$ is shattered. Equivalently, starting with m arbitrary triplets $\{(i_t, j_t, k_t)\}_{i=1}^m$ (one can think of them as being selected by an adversary) we need to show that there is a labeling of the triplets such that no embedding can satisfy all of them. We let $\varepsilon = \Theta(\alpha)$ be a sufficiently small constant fraction of α and let $m = \frac{n}{\varepsilon^2} = \Theta(\frac{n}{\alpha^2})$: we show that for a set $\{(i_t, j_t, k_t)\}_{t=1}^m$ of unlabeled triplets there exists a labeling (i.e., every unlabeled triplet is labeled as either $(i_t, j_t^+, k_t^-)_\alpha$ or $(i_t, k_t^+, j_t^-)_\alpha$) such that no embedding $\phi : [n] \to \mathbb{R}^n$ is consistent with all the labeled triplets.

Rounding to 1D. Our proof, perhaps surprisingly, begins by an observation drawing ideas from rounding semidefinite programs in the approximation algorithms literature. Suppose that we have a set of triplet constraints with margin α, $\left\{(i_t, j_t^+, k_t^-)_\alpha\right\}_{t=1}^m$ and an embedding $\phi : [n] \to \mathbb{R}^n$ that satisfies all the constraints. Then, as we show, one can *round* the embedding ϕ to a *one-dimensional embedding* $\hat{\phi} : [n] \to \mathbb{R}$ so that a fraction of at least $\frac{1}{2} + \Omega(\alpha)$ of the triplet constraints without margin $\left\{(i_t, j_t^+, k_t^-)\right\}_{t=1}^m$ are satisfied by $\hat{\phi}$. A priori this result seems weak: not only the constraints that $\hat{\phi}$ satisfies are without margin, but also only a fraction of $\frac{1}{2} + \Omega(\alpha)$ of them are satisfied.[5] The fact, however, that there always exists such an embedding in one dimension that is strictly better than the random solution ends up being strong enough to give us the result. Our rounding algorithm projects all the embedded vectors along a random direction in a way similar to the MAX-CUT rounding algorithm of Goemans & Williamson (1995), see Lemma 3.2 for the formal statement and proof.

No 1D embedding can fit noise on n/ε^2 triplets. We now go back to our setting where we are given a set of $m = \frac{n}{\varepsilon^2}$ unlabeled triplets, adversarially selected, and we want to show that there is a labeling of the triplets such that no embedding, even in n dimensions, can satisfy *all* of them. The key idea is that our rounding algorithm allows us to focus on the real line and forget about the margin parameter. Recall that Alon et al. (2024) show that the VC-dimension of embeddings in dimension d is $\Theta(nd)$, and so for $d = 1$, we get that the VC dimension is $\approx n$. What this means in particular, is that no embedding on the real line is able to fit $\approx \frac{n}{\varepsilon^2}$ triplets labeled by random noise: if we select the labels of a set of $m = n/\varepsilon^2$ triplets randomly then with constant probability, no embedding will be

[5] For comparison, a trivial one-dimensional embedding that maps every item to a uniform number in $[0, 1]$ satisfies half of the constraints on expectation.

able to satisfy more than $\frac{1}{2}+\varepsilon$ of them. This is formalized by the notion of Rademacher complexity and its bound in terms of the VC-dimension of a hypothesis class (Shalev-Shwartz & Ben-David, 2014; Bousquet et al., 2003). Using the probabilistic method, we get that for every set of $m = \frac{n}{\varepsilon^2}$ unlabeled triplets, there exists a labeling such that no embedding on the real line satisfies more than $(\frac{1}{2}+\varepsilon)m$ of them. The formal statement and proof can be found in Lemma 3.4.

2.3 Putting things together

We now discuss how to combine the ideas from the above paragraphs to get the result. Let $\{(i_t, j_t, k_t)\}_{t=1}^m$ be the set of adversarially selected triplets and suppose, for the sake of contradiction, that the set is shattered by the hypothesis class $\mathcal{H}(n, d, \alpha)$. On the one hand, we get that since m is large enough and linear in n, there exists a labeling of the triplets, such that no embedding on the real line (no $h \in \mathcal{H}(n, 1)$) satisfies more than a fraction of $\frac{1}{2}+\varepsilon$ of them. Consider that labeling for the triplets. By our assumption that the set is shattered there exists an embedding (potentially in n dimensions) such that all the triplets (with margin) are satisfied. Applying our rounding algorithm on that embedding we get that there exists an embedding on the real line satisfying a fraction of $\frac{1}{2}+\Omega(\alpha)$ of them (without margin). This leads to a contradiction that:

$$\frac{1}{2}+\Omega(\alpha) \leq \frac{1}{2}+\varepsilon,$$

yet we can choose $\varepsilon = \Theta(\alpha)$ to be sufficiently small compared to α. The above are made formal in our proof of Theorem 3.1 in Section 3.

3 Improved VC-dimension: Upper bound of Theorem 1.4

In this section, we prove the following theorem.

Theorem 3.1 (Upper bound for $\mathrm{VCdim}(\mathcal{H}(n, d, \alpha))$)**.** *The VC-dimension of $\mathcal{H}(n, d, \alpha)$ is linear in n and independent of the dimensionality d. In particular, there exists an absolute constant C such that:*

$$\mathrm{VCdim}(\mathcal{H}(n, d, \alpha)) \leq C \max(1, \frac{1}{\alpha^2}) \cdot n.$$

Note that the VC dimension is monotonically decreasing in α, thus it suffices to show that for $\alpha \in (0, 1]$ it is bounded by $C\frac{n}{\alpha^2}$. From now on we assume that $\alpha \in (0, 1]$. Our proof uses the following lemma which, roughly speaking, states that if a set of contrastive learning constraints with margin α, $\left\{(i_t, j_t^+, k_t^-)_\alpha\right\}_{t=1}^m$ are satisfiable (potentially in n dimensions) then there exists an embedding on the real line satisfying a $\frac{1}{2}+\Omega(\alpha)$ fraction of the contrastive constraints $\left\{(i_t, j_t^+, k_t^-)\right\}_{t=1}^m$. Note here that the constraints that $\hat{\phi}$ satisfies are without margin.

Lemma 3.2 (Rounding to 1D)**.** *Let $\phi : [n] \to \mathbb{R}^n$ be such that the constraints $\left\{(i_t, j_t^+, k_t^-)_\alpha\right\}_{t=1}^m$ are satisfied by ϕ. Then there exists a $\hat{\phi} : [n] \to \mathbb{R}$ such that at least $\left(\frac{1}{2}+\frac{\alpha}{5}\right) \cdot m$ constraints of $\left\{(i_t, j_t^+, k_t^-)\right\}_{t=1}^m$ are satisfied by $\hat{\phi}$.*

Proof. Let $g \sim \mathcal{N}(0, I_n)$ be a standard Gaussian vector. We will consider the random embedding $\hat{\phi}_g : [n] \to \mathbb{R}$ given by $\hat{\phi}(i) = \langle \phi(i), g \rangle$ and show that on expectation, the embedding satisfies a fraction of $\frac{1}{2}+\frac{\alpha}{5}$ of constraints in $\left\{(i_t, j_t^+, k_t^-)\right\}_{t=1}^m$. Note that by virtue of the fact that $\|\phi(i)-\phi(k)\| \neq \|\phi(i)-\phi(j)\|$ for every i, j, k with $j \neq k$, we get that the with probability one, $|\hat{\phi}_g(i)-\hat{\phi}_g(j)| \neq |\hat{\phi}_g(i)-\hat{\phi}_g(k)|$. The above combined imply that there exists a v such that the embedding $\hat{\phi}$ given by $\hat{\phi}(i) = \langle \phi(i), v \rangle$ satisfies a fraction of at least $\frac{1}{2}+\frac{\alpha}{5}$ of constraints and furthermore $h_{\hat{\phi}} \in \mathcal{H}(n, 1)$.

For $i \in [n]$ let $\bar{i} = \phi(i)$. Observe that the embedding $\hat{\phi}$ satisfies a constraint (i, j^+, k^-) if:

$$\left|\langle \bar{i}-\bar{k}, g \rangle\right| > \left|\langle \bar{i}-\bar{j}, g \rangle\right|.$$

By linearity of expectation, the expected number of constraints that $\hat{\phi}_g$ satisfies is

$$\sum_{t=1}^m \mathbb{P}\left(\left|\langle \bar{i}_t-\bar{k}_t, g \rangle\right| > \left|\langle \bar{i}_t-\bar{j}_t, g \rangle\right|\right),$$

where the probability is over the randomness in g. It therefore suffices to show that for every $t \in [m]$:

$$\mathbb{P}\left(\left|\langle \bar{i}_t - \bar{k}_t, g\rangle\right| > |\langle \bar{i}_t - \bar{j}_t, g\rangle|\right) \geq \frac{1}{2} + \frac{\alpha}{5},$$

where for vectors $u = \bar{i}_t - \bar{k}_t$ and $v = \bar{i}_t - \bar{j}_t$ it holds that $\|u\| > (1+\alpha)\|v\|$. To that end, we prove the following claim.

Claim 3.3. *Let $u, v \in \mathbb{R}^d$ and let g be a standard Gaussian vector. If $\|u\| > (1+\alpha)\|v\|$ for some $\alpha > 0$, then*

$$\mathbb{P}\left(|\langle u, g\rangle| > |\langle v, g\rangle|\right) \;\geq\; \frac{2}{\pi}\arctan(1+\alpha).$$

In particular, for $\alpha \in [0,1]$:

$$\mathbb{P}\left(|\langle u, g\rangle| > |\langle v, g\rangle|\right) \;\geq\; \frac{1}{2} + \frac{\alpha}{5}.$$

Claim 3.3, which we prove in Appendix A gives us the result.[6] □

In the next lemma we use the upper bound on the VC dimension of contrastive learning on the real line, proven in Alon et al. (2024), to prove that for every set of $\approx \frac{n}{\varepsilon^2}$ unlabeled triplets (i_t, j_t, k_t), there exists a labeling of them such that no embedding is consistent with more than a fraction of $\frac{1}{2} + \varepsilon$ of them.

Lemma 3.4. *There exists an absolute constant c such that for every $\varepsilon > 0$, the following holds. For every $m \geq \frac{c}{\varepsilon^2} n$ and set $\{(i_t, j_t, k_t)\}_{t=1}^m$ there exists a labeling $y : [m] \to \{-1, 1\}$ such that no embedding $\phi : [n] \to \mathbb{R}$ satisfies more than a fraction of $\frac{1}{2} + \varepsilon$ of the geometric triplet constraints induced by the labeling. In particular, for every $h \in \mathcal{H}(n, 1)$ it holds that:*

$$\frac{1}{m}\sum_{t=1}^{m} \mathbf{1}\left\{h(i_t, j_t, k_t) = y(t)\right\} \leq \frac{1}{2} + \varepsilon$$

Proof. We will use the probabilistic method. In particular, we will pick the label y randomly and show that

$$\mathbb{E}_y\left[\sup_{h \in \mathcal{H}(n,1)} \frac{1}{m}\sum_{t=1}^{m} \mathbf{1}\left\{h(i_t, j_t, k_t) = y(t)\right\}\right] \leq \frac{1}{2} + \varepsilon, \tag{1}$$

using bounds for the worst case empirical Rademacher complexity of a hypothesis class in terms of the VC-dimension, and then the result directly follows. In particular, we let $y(t) = 1$ with probability $1/2$ and $y(t) = -1$ with the remaining probability, independently of everything else. We can equivalently write Equation 1 as:

$$\mathbb{E}_y\left[\sup_{h \in \mathcal{H}(n,1)} \frac{1}{m}\sum_{t=1}^{m} h(i_t, j_t, k_t) \cdot y(t)\right] \leq 2\varepsilon, \tag{2}$$

and observe that the left hand side exactly corresponds to the empirical Rademacher complexity of hypothesis class $\mathcal{H}(n, 1)$ on points $\{(i_t, j_t, k_t)\}_{t=1}^m$. We now use (see e.g. Bousquet et al. (2003)) that there exists an absolute constant c' such that the empirical Rademacher complexity of a hypothesis class $\mathcal{H}$ on m points can be upper bounded in terms of the VC dimension of the class. In particular, if $\mathcal{R}_m$ is the worst-case[7] empirical Rademacher complexity on m points then

$$\mathcal{R}_m \leq c'\sqrt{\frac{\text{VCdim}(\mathcal{H})}{m}}. \tag{3}$$

[6] Observe this claim is for two vectors u, v as long as they have a gap α, so it also allows us to capture the case of quadruplet comparisons (i, j, k, l) (Bilu & Linial, 2005; Vankadara et al., 2023), where we compare the distance between i, j to the distance between k, l.

[7] Supremum over sets of size m.

We apply this bound to the left hand side of Equation 2 and use that $\mathrm{VCdim}(\mathcal{H}(n,1)) = O(n)$ to get that for c being a sufficiently large constant:

$$\mathbb{E}_y \left[\sup_{h \in \mathcal{H}(n,1)} \frac{1}{m} \sum_{t=1}^{m} h(i_t, j_t, k_t) \cdot y(t) \right] \leq c' \sqrt{\frac{c_0 n \varepsilon^2}{cn}} \leq 2\varepsilon,$$

which gives us the result. □

We can now prove our main Theorem 3.1.

Proof of Theorem 3.1. Let $\varepsilon < \frac{\alpha}{5}$ and let c be a large enough constant. We will show that no set of size $m \geq c\frac{n}{\varepsilon^2} = \Omega(\frac{n}{\alpha^2})$ can be shattered, which gives the result. Consider a set of triplets $\{(i_t, j_t, k_t)\}_{t=1}^m$ and assume, for the sake of contradiction, that the set is shattered by $\mathcal{H}(n, d, \alpha)$, meaning that for every labeling of the triplets, there is an embedding $\phi : [n] \to \mathbb{R}^n$ that is consistent with all the corresponding triplet constraints.

Formally, by Lemma 3.4, taking c bigger than the constant in the lemma we get that there exists a labeling y^* such that for every $h \in \mathcal{H}(n, 1)$:

$$\frac{1}{m} \sum_{t=1}^{m} \mathbf{1} \left\{ h(i_t, j_t, k_t) = y_t^* \right\} \leq \frac{1}{2} + \varepsilon. \tag{4}$$

On the other hand, by our assumption that the set $\{(i_t, j_t, k_t)\}_{t=1}^m$ is shattered, we get that there exists an embedding ϕ satisfying all the corresponding constraints: if $y_t^* = 1$ then $\|\phi(i_t) - \phi(k_t)\| > (1+\alpha)\|\phi(i_t) - \phi(j_t)\|$, and if $y_t^* = -1$ then $\|\phi(i_t) - \phi(j_t)\| > (1+\alpha)\|\phi(i_t) - \phi(k_t)\|$. Without loss of generality and for notational convenience, we can assume that the labeled triplets are $\left\{(i_t, j_k^+, k_t^-)_\alpha\right\}_{t=1}^m$; by Lemma 3.2, there exists a one dimensional embedding $\hat{\phi} : [n] \to \mathbb{R}$ such that a fraction of at least $\frac{1}{2} + \frac{\alpha}{5}$ constraints in $\left\{(i_t, j_t^+, k_t^-)\right\}$ are satisfied.[8] For the function $h_{\hat{\phi}} \in \mathcal{H}(n, 1)$ this implies that:

$$\frac{1}{m} \sum_{t=1}^{m} \mathbf{1} \left\{ h_{\hat{\phi}}(i_t, j_t, k_t) = y_t^* \right\} \geq \frac{1}{2} + \frac{\alpha}{5}. \tag{5}$$

Combining Equations 4 and 5 we get that:

$$\frac{1}{2} + \frac{\alpha}{5} \leq \frac{1}{2} + \varepsilon,$$

which is a contradiction since we have taken $\varepsilon < \frac{\alpha}{5}$. □

4 Experiments

To complement our theoretical bounds, we empirically explore how the generalization error $\hat{\varepsilon}$ is affected as we vary the dimension d, the number of points n, the margin α and the number m of given triplet samples. We test both on synthetic data generated from a Gaussian distribution and on real datasets (CIFAR-100). Our goal is to test the dimension-independence of our $O(n/\alpha^2)$ bound.

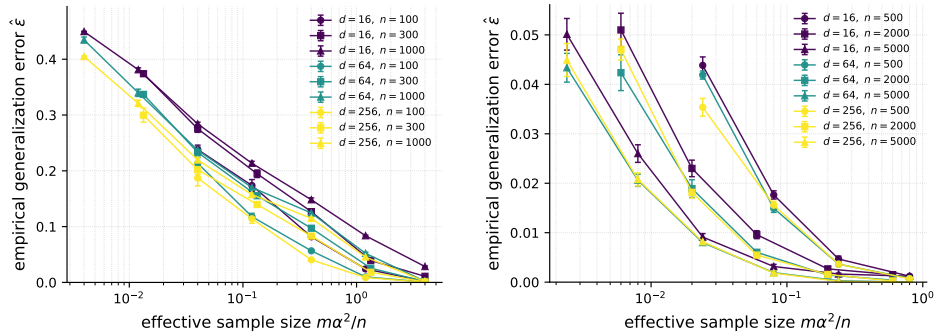
Experimental Setting. For each configuration (n, d, α), we sample m training triplets from $[n]^3$ uniformly at random, reject any that fail the $(1+\alpha)$-margin condition under a ground-truth metric ϕ^*, and fit an embedding $\hat{\phi} : [n] \to \mathbb{R}^d$ by minimizing the multiplicative-margin triplet loss

$$\mathcal{L}(\hat{\phi}) \;=\; \frac{1}{m} \sum_{t=1}^{m} \left[(1+\alpha)^2 \left\|\hat{\phi}(i_t) - \hat{\phi}(j_t)\right\|_2^2 - \left\|\hat{\phi}(i_t) - \hat{\phi}(k_t)\right\|_2^2 \right]_+ ,$$

which is the natural surrogate for the $(1+\alpha)$-separation defining $\mathcal{H}(n, d, \alpha)$ (see Schroff et al. (2015) and survey Vankadara et al. (2023)). We optimize with Adam, with learning rate 0.01 and 1000 steps, and report the empirical generalization error $\hat{\varepsilon}$ on $5{,}000$ held-out $(1+\alpha)$-separated test triplets.

[8] These are constraints without margin.

We perform experiments for two types of ground-truth embeddings. In the *synthetic* setting, ϕ^* is the undelying embedding of n points drawn i.i.d. from $\mathcal{N}(0, I_{50})$ and the learner directly parameterizes the n embedding rows in $\mathbb{R}^d$. In the *real-data* setting, ϕ^* is the L2-normalized penultimate feature (this provides a vector in $\mathbb{R}^{512}$) of an ImageNet-pretrained ResNet18 He et al. (2016) on 5,000 CIFAR-100 test images Krizhevsky (2009), frozen during training, and the learner is a linear head $\mathbb{R}^{512} \to \mathbb{R}^d$ on top of these features.



(a) Synthetic, $\alpha = 0.2$.

(b) CIFAR-100 / ResNet18, $\alpha = 0.2$.

Figure 1: Dimension-independence: empirical generalization error $\hat{\varepsilon}$ against $m\alpha^2/n$ at $\alpha = 0.2$. **(a)** Synthetic setup with underlying embedding $\phi^* \in \mathbb{R}^{50}$. **(b)** CIFAR-100 / ResNet18 setup with a linear head over 512-dimensional features.

Dimension-independence. As Theorem 3.1 does not depend on the embedding dimension d, for fixed $m\alpha^2/n$, the empirical error should not change as we increase d. To test this we fix $\alpha = 0.2$ and vary $d \in \{16, 64, 256\}$. On the synthetic side we keep the ground-truth ambient dimension at 50 so that the data distribution itself is unchanged, and only the learner's capacity changes. We sweep $n \in \{100, 300, 1000\}$ and $m \in \{100, 300, 10^3, 3{\cdot}10^3, 10^4, 3{\cdot}10^4, 10^5\}$ on the synthetic setup, and $n \in \{500, 2000, 5000\}$ and $m \in \{300, 10^3, 3{\cdot}10^3, 10^4, 3{\cdot}10^4, 10^5\}$ on the real-data setup, over 5 random seeds. Figure 1 agrees with the theoretical predictions on both setups, as it shows the three curves corresponding to different dimensions $d = 16, 64, 256$ (encoded by different colors) lie on top of each other within each n-trajectory. Interestingly, the plots show that on real-data (Fig. 1 (b)), dimension-independence is quite prominent, since as d varies, the error $\hat{\varepsilon}$ changes only slightly (notice the small values of $\hat{\varepsilon}$ on the y-axis). Appendix C provides further experiments.

Conclusion

We resolved an open question from Alon et al. (2024) related to provable generalization guarantees and the VC-dimension of contrastive learning, in the presence of a margin parameter separating the distance between an anchor-positive pair from an anchor-negative pair. Enforcing a margin is not only a practically well-motivated approach in the literature of contrastive learning and ordinal embeddings, but as we show it also leads to interesting theoretical insights. Our main result improves upon the direct application of standard low-distortion metric embedding techniques, such as the Johnson-Lindenstrauss lemma, and provides an optimal $\Theta(n/\alpha^2)$ bound for the VC-dimension for learning a ground-truth embedding on n data points of interest. Surprisingly, this is independent of the underlying dimensionality d of the ground-truth embedding, which is often large. Given that our bound is optimal in n, our work makes steps towards understanding the empirical success of contrastive learning. Technically, our approach uses ideas from approximation algorithms and geometric projections, coupled with learning theory bounds for learning triplet comparisons.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing, in particular for typos and grammatical errors as well as for feedback on exposition AI tools were also used to help with the experiments. Generative AI tools were used for literature search, in particular to find a reference of the Rademacher complexity bound in terms of VC dimension. We have not used generative AI tools for technical ideas or proofs. We have reviewed all AI-assisted work. We checked that the reference provided to us indeed gives the claim that we use. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This is a work of theoretical nature, we don't see any potential ethical concerns regarding our work.

REFERENCES

Noga Alon, Peter Frankl, and Vojtech Rodl. Geometrical realization of set systems and probabilistic communication complexity. In *26th Annual Symposium on Foundations of Computer Science (sfcs 1985)*, pp. 277–280. IEEE, 1985.

Noga Alon, Mihai Bădoiu, Erik D Demaine, Martin Farach-Colton, MohammadTaghi Hajiaghayi, and Anastasios Sidiropoulos. Ordinal embeddings of minimum relaxation: general properties, trees, and ultrametrics. *ACM Transactions on Algorithms (TALG)*, 4(4):1–21, 2008.

Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A theory of pac learnability of partial concept classes. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 658–671. IEEE, 2022.

Noga Alon, Dmitrii Avdiukhin, Dor Elboim, Orr Fischer, and Grigory Yaroslavtsev. Optimal sample complexity of contrastive learning. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

Dionysis Arvanitakis, Vaggos Chatziafratis, and Yiyuan Luo. Provable accuracy collapse in embedding-based representations under dimensionality mismatch. *43rd International Conference on Machine Learning (ICML)*, 2026.

Dmitrii Avdiukhin, Grigory Yaroslavtsev, Danny Vainstein, Orr Fischer, Sauman Das, and Faraz Mirza. Tree learning: optimal sample complexity and algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 6701–6708, 2023.

Dmitrii Avdiukhin, Vaggos Chatziafratis, Orr Fischer, and Grigory Yaroslavtsev. Embedding dimension of contrastive learning and k-nearest neighbors. *Advances in Neural Information Processing Systems*, 37:41359–41393, 2024.

Pranjal Awasthi, Nishanth Dikkala, and Pritish Kamath. Do more negative samples necessarily hurt in contrastive learning? In *International conference on machine learning*, pp. 1101–1116. PMLR, 2022.

Yonatan Bilu and Nati Linial. Monotone maps, sphericity and bounded second eigenvalue. *Journal of Combinatorial Theory, Series B*, 95(2):283–299, 2005.

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Summer school on machine learning*, pp. 169–207. Springer, 2003.

Vaggos Chatziafratis and Piotr Indyk. Dimension-accuracy tradeoffs in contrastive embeddings for triplets, terminals & top-k nearest neighbors. In *2024 Symposium on Simplicity in Algorithms (SOSA)*, pp. 230–243. SIAM, 2024.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

Ehsan Emamjomeh-Zadeh and David Kempe. Adaptive hierarchical clustering using ordinal queries. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 415–429. SIAM, 2018.

Bohan Fan, Diego Ihara Centurion, Neshat Mohammadi, Francesco Sgherzi, Anastasios Sidiropoulos, and Mina Valizadeh. Learning lines with ordinal constraints. *arXiv preprint arXiv:2004.13202*, 2020.

Nikhil Ghosh, Yuxin Chen, and Yisong Yue. Landmark ordinal embedding. *Advances in neural information processing systems*, 32, 2019.

Debarghya Ghoshdastidar, Michaël Perrot, and Ulrike von Luxburg. Foundations of comparison-based hierarchical clustering. *Advances in neural information processing systems*, 32, 2019.

Michel X Goemans and David P Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6): 1115–1145, 1995.

Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 855–864, 2016.

Venkatesan Guruswami, Rajsekar Manokaran, and Prasad Raghavendra. Beating the random ordering is hard: Inapproximability of maximum acyclic subgraph. In *2008 49th Annual IEEE Symposium on Foundations of Computer Science*, pp. 573–582. IEEE, 2008.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in neural information processing systems*, 34:5000–5011, 2021.

Gustav Hast. Beating a random assignment. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pp. 134–145. Springer, 2005.

Johan Håstad. Some optimal inapproximability results. *Journal of the ACM (JACM)*, 48(4):798–859, 2001.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Subhash Khot. On the power of unique 2-prover 1-round games. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pp. 767–775, 2002.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Brian Kulis. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4): 287–364, 2013.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

Yunwen Lei, Tianbao Yang, Yiming Ying, and Ding-Xuan Zhou. Generalization analysis for contrastive representation learning. In *International Conference on Machine Learning*, pp. 19200–19227. PMLR, 2023.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International conference on machine learning*, pp. 5628–5637. PMLR, 2019.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pp. 19250–19286. PMLR, 2022.

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*, pp. 815–823, 2015.

Matthew Schultz and Thorsten Joachims. Learning a distance metric from relative comparisons. *Advances in neural information processing systems*, 16, 2003.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. Adaptively learning the crowd kernel. *28th International Conference on Machine Learning (ICML)*, 2011.

Yoshikazu Terada and Ulrike Luxburg. Local ordinal embedding. In *International Conference on Machine Learning*, pp. 847–855. PMLR, 2014.

Leena Chennuru Vankadara, Michael Lohaus, Siavash Haghiri, Faiz Ul Wahab, and Ulrike Von Luxburg. Insights into ordinal embedding algorithms: A systematic evaluation. *Journal of Machine Learning Research*, 24(191):1–83, 2023.

Hugh E Warren. Lower bounds for approximation by nonlinear manifolds. *Transactions of the American Mathematical Society*, 133(1):167–178, 1968.

Xin Zou and Weiwei Liu. Generalization bounds for adversarial contrastive learning. *Journal of Machine Learning Research*, 24(114):1–54, 2023.

A REDUCTION TO 1D

In this section we prove Claim 3.3.

Proof of Claim 3.3. If $v = 0$, the claim is immediate. Assume $v \neq 0$, which also implies $u \neq 0$. Let $\bar{u} = u/\|u\|$ and $\bar{v} = v/\|v\|$. Define $\beta = 1 + \alpha$. Then the inequality $|\langle u, g\rangle| > |\langle v, g\rangle|$ holds if and only if

$$|\langle \bar{u}, g\rangle| > \frac{\|v\|}{\|u\|} \cdot |\langle \bar{v}, g\rangle|.$$

Since $\|u\| > \beta\|v\|$, it suffices to show

$$\mathbb{P}\left(|\langle \bar{v}, g\rangle| < \beta|\langle \bar{u}, g\rangle|\right) \;\geq\; \frac{2}{\pi}\arctan(\beta).$$

If $\bar{v} = \pm\bar{u}$, then the event holds with probability 1, so we may assume that $\bar{u}$ and $\bar{v}$ are not collinear. Let S be the two-dimensional subspace spanned by $\bar{u}$ and $\bar{v}$. We decompose g as

$$g = \tilde{g} + g^{\perp},$$

where $\tilde{g}$ is the orthogonal projection of g onto S. Since both $\bar{u}$ and $\bar{v}$ lie in S, we have

$$\langle \bar{u}, g\rangle = \langle \bar{u}, \tilde{g}\rangle, \quad \langle \bar{v}, g\rangle = \langle \bar{v}, \tilde{g}\rangle.$$

Let $\theta \in (0, \pi)$ be the angle between $\bar{u}$ and $\bar{v}$, measured counterclockwise, and let φ be the angle from $\bar{u}$ to $\tilde{g}$, also measured counterclockwise. To visualize the configuration, align the first basis vector e_1 of S with $\bar{u}$. Then $\bar{u}$ has coordinates $(1, 0)$, $\bar{v}$ has coordinates $(\cos\theta, \sin\theta)$, and $\tilde{g}$ has coordinates $(\cos\varphi, \sin\varphi)$. See Figure 2. We exclude the events $\tilde{g} = 0$ and $\varphi \in \{\pi/2, 3\pi/2\}$, which occur with probability 0. Then φ is uniformly distributed on

$$A = (-\pi/2, \pi/2) \cup (\pi/2, 3\pi/2).$$

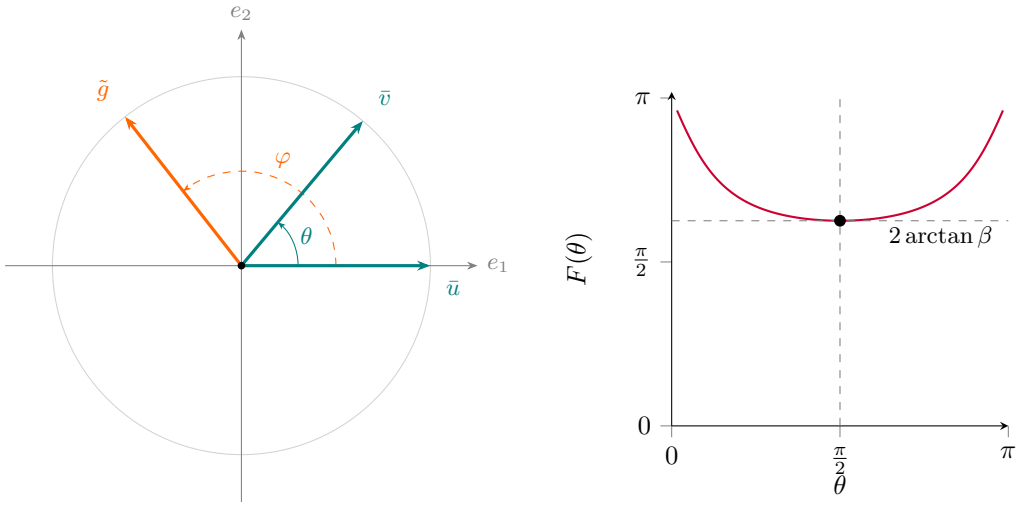


Figure 2: Left: vectors $\bar{u}$, $\bar{v}$, and $\tilde{g}$ in the plane, with angles θ and φ. Right: the function $F(\theta)$ for $\beta = 1.5$, illustrating that the minimum is attained at $\theta = \pi/2$.

We have

$$\langle \tilde{g}, \bar{u}\rangle = \|\tilde{g}\|\cos\varphi, \quad \langle \tilde{g}, \bar{v}\rangle = \|\tilde{g}\|\cos(\varphi - \theta).$$

Therefore,

$$|\langle \bar{v}, g\rangle| < \beta|\langle \bar{u}, g\rangle| \iff |\cos(\varphi - \theta)| < \beta|\cos\varphi|.$$

Define

$$\mathcal{E} = \{\varphi \in A : |\cos(\varphi - \theta)| < \beta|\cos\varphi|\}.$$

Since $\cos\varphi \neq 0$, this is equivalent to

$$\mathcal{E} = \Big\{\varphi \in A : -\beta < \frac{\cos(\varphi - \theta)}{\cos\varphi} < \beta\Big\}.$$

Using $\cos(\varphi - \theta) = \cos\varphi\cos\theta + \sin\varphi\sin\theta$, we get

$$\mathcal{E} = \Big\{\varphi \in A : -\beta < \cos\theta + \tan(\varphi)\sin\theta < \beta\Big\}.$$

Since $\theta \in (0, \pi)$, we have $\sin\theta > 0$, and hence

$$\mathcal{E} = \Big\{\varphi \in A : \frac{-\beta - \cos\theta}{\sin\theta} < \tan(\varphi) < \frac{\beta - \cos\theta}{\sin\theta}\Big\}.$$

Therefore, $\mathcal{E}$ consists of two intervals I and $I + \pi$, where

$$I = \Big(\arctan\Big(\frac{-\beta - \cos\theta}{\sin\theta}\Big), \arctan\Big(\frac{\beta - \cos\theta}{\sin\theta}\Big)\Big).$$

The probability of $\varphi \in \mathcal{E}$ equals

$$\frac{|\mathcal{E}|}{2\pi} = \frac{|I|}{\pi} = \frac{1}{\pi}\Big[\arctan\Big(\frac{\beta - \cos\theta}{\sin\theta}\Big) - \arctan\Big(\frac{-\beta - \cos\theta}{\sin\theta}\Big)\Big].$$

Denote the expression in the square brackets on the right hand side by $F(\theta)$ (see Figure 2):

$$F(\theta) = \arctan\Big(\frac{\beta - \cos\theta}{\sin\theta}\Big) - \arctan\Big(\frac{-\beta - \cos\theta}{\sin\theta}\Big).$$

Then the probability above is $F(\theta)/\pi$. The derivative of $F(\theta)$ equals

$$F'(\theta) = \frac{1 - \beta\cos\theta}{1 - 2\beta\cos\theta + \beta^2} - \frac{1 + \beta\cos\theta}{1 + 2\beta\cos\theta + \beta^2}.$$

Write it as

$$F'(\theta) = G(-\beta\cos\theta) - G(\beta\cos\theta), \quad \text{where } G(x) = \frac{x + 1}{2x + 1 + \beta^2}.$$

Function $G(x)$ is increasing for $x > -(\beta^2 + 1)/2$ because

$$G(x) = \frac{1}{2}\Big[1 - \frac{\beta^2 - 1}{2x + 1 + \beta^2}\Big]$$

and $\beta^2 - 1 > 0$. Observe that $\beta\cos\theta > -(\beta^2 + 1)/2$ and $-\beta\cos\theta > -(\beta^2 + 1)/2$. Hence $F'(\theta) < 0$ when $\cos\theta > 0$ and $F'(\theta) > 0$ when $\cos\theta < 0$. Therefore, F is minimized at $\theta = \pi/2$. Consequently,

$$F(\theta) \geq F(\pi/2) = \arctan(\beta) - \arctan(-\beta) = 2\arctan(\beta).$$

Thus

$$\mathbb{P}(\varphi \in \mathcal{E}) = \frac{F(\theta)}{\pi} \geq \frac{2}{\pi}\arctan(\beta).$$

This proves the first part of the claim. We now use that the function $\arctan(x)$ is concave for $x \geq 0$ meaning that in any interval, it can be lower bounded by the secant line between the endpoints of the interval. In particular, we get that for $\alpha \in [0, 1]$:

$$\begin{aligned}\frac{2}{\pi}\arctan(\beta) &\geq \frac{2}{\pi}\arctan(1 + \alpha) \\ &\geq \frac{2}{\pi}\Big((\arctan(2) - \frac{\pi}{4})\alpha + \frac{\pi}{4}\Big) \\ &= \frac{1}{2} + \frac{2}{\pi}\Big(\arctan(2) - \frac{\pi}{4}\Big)\alpha,\end{aligned}$$

where $\frac{2}{\pi}\left(\arctan(2) - \frac{\pi}{4}\right) \approx 0.204833 \geq \frac{1}{5}$. □

B A MATCHING LOWER BOUND

In this section we prove a matching lower bound for the VC dimension of contrastive learning with margin parameter α. In particular we show that as long as $\alpha \geq \Omega(\max(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{d}}))$, then $\mathrm{VCdim}(\mathcal{H}(n,d,\alpha)) \geq \Omega(\max\{\frac{n}{\alpha^2}, n\})$.

Theorem B.1. *Suppose that $\alpha \geq c' \max(n^{-1/2}, d^{-1/2})$ for some large enough absolute constant c'. Then, there exists an absolute constant c such that:*

$$\mathrm{VCdim}(\mathcal{H}(n,d,\alpha)) \geq c\frac{n}{\alpha^2}.$$

Proof. We divide between two cases. We first consider the case where $\alpha \geq \frac{1}{4}$, in which case we show that the VC dimension is $\Omega(n)$. We make a distinction between the elements in $[n]$ we use i, j to denote the first two elements and $\{k_t\}_{t=1}^{n-2}$ for the rest. We consider the set of $\Theta(n)$ triplets given by $\{(i,j,k_t)\}_{t=1}^{n-2}$ and show that they are shattered. For a labeling of the triplets, let $A = \left\{t : (i,j,k_t) \text{ is labeled as } (i,k_t^+,j^-)_\alpha\right\}$ and $B = [n] \setminus A$. We define the embedding to be $\phi(i) = 0$, $\phi(j) = 1$, furthermore for $t \in A$ we let $\phi(k_t) = \kappa$ where $\kappa < \frac{1}{2+\alpha}$ and for $t \in B$, $\phi(k_t) = \tau$ where $\tau > 1 + \alpha$. The embedding satisfies all the triplets. In the above construction there exist i, j, k with $j \neq k$ such that $\|\phi(i) - \phi(j)\| = \|\phi(i) - \phi(k)\|$. A small perturbation of all points resolves this issue while maintaining the conclusion.

In the remaining we consider $\alpha < \frac{1}{4}$. We proceed to describe an instance of $\approx \frac{n}{\alpha^2}$ triplets that is shattered. Let for convenience $\gamma = 4\alpha$. We first make a distinction between the elements in $[n]$. We will use $k_1, \ldots, k_{\frac{1}{\gamma^2}}$ to denote the first $\frac{1}{\gamma^2}$ elements of $[n]$ and i_t, j_t with $t \in \left[\frac{n-\frac{1}{\gamma^2}}{2}\right]$ for the remaining elements. The set of triplets that we propose is inspired by the construction of Alon et al. (2022) for the lower bound on the VC dimension of linear classifiers with margin.

We consider the set of triplets $\left\{(i_t, j_t, k_l) : t \in \left[\frac{n-\frac{1}{\gamma^2}}{2}\right], l \in [\frac{1}{\gamma^2}]\right\}$ and show that the set is shattered. Note that the number of triplets is $\frac{1}{2\gamma^2}(n - \frac{1}{\gamma^2}) = \Theta(\frac{n}{\alpha^2})$ (here, we have used that $\frac{1}{\alpha^2} \leq c' \cdot n$), thus showing that they can be shattered gives us the result. For every labeling of the triplets, the embedding for points k_l will be the same. In particular, we will always map k_l to e_l, the l-th standard basis vector (here we have used that $\frac{1}{\alpha^2} \leq c' \cdot d$). Now consider a labeling of the triplets and observe that we can decompose the triplets as $\cup_t \{(i_t, j_t, k_l)\}$, where having fixed the mappings of k_l, the instances are decoupled for different t. It therefore suffices to show that for every labeling of triplets $\{(i,j,k_l)\}$ there is an embedding of points i and j that satisfies all the triplets.

Consider a labeling of the triplets $\{(i,j,k_l)\}_{l=1}^{\frac{1}{\gamma^2}}$ and let A be the subset of l in $[\frac{1}{\gamma^2}]$ such that the triplet (i,j,k_l) is labeled as $(i,k_l^+,j^-)_\alpha$ and $B = [\frac{1}{\gamma^2}] \setminus A$. Let $w = \sum_{l\in A} e_l - \sum_{l \in B} e_l$, w is exactly the vector that defines the hyperplane separating the points $\{e_l, l \in A\}$ from the points $\{e_l, l \in B\}$ with margin 2γ, see Proposition 17 in Alon et al. (2022). We let $\phi(i) = \gamma \cdot w$. We also select j to be a point at distance $\sqrt{2}$ from $\phi(i)$. We now consider $l \in A$, we have that

$$\begin{aligned}
\|\phi(i) - \phi(k_l)\| &= \|\phi(i) - e_l\| \\
&= \left((1-\gamma)^2 + \gamma^2\left(\frac{1}{\gamma^2} - 1\right)\right)^{\frac{1}{2}} \\
&= \sqrt{2 - 2\gamma}.
\end{aligned}$$

This in turn means that:

$$\begin{aligned}
\|\phi(i) - \phi(j)\| &= \sqrt{2} \\
&= \frac{1}{\sqrt{1-\gamma}} \cdot \|\phi(i) - e_l\| \\
&\geq (1 + \frac{1}{2}\gamma) \cdot \|\phi(i) - e_l\| \\
&= (1 + 2\alpha) \cdot \|\phi(i) - \phi(k_l)\|
\end{aligned}$$

where we have used that the function $f(x) = \frac{1}{\sqrt{1-x}}$ is convex and $f'(0) = \frac{1}{2}$. We therefore get the the $(i, k_l^+, j^-)_\alpha$ is satisfied by the embedding.

On the other hand, for $l \in B$, we have that

$$\begin{aligned}\|\phi(i) - \phi(k_l)\| &= \sqrt{1+\gamma} \cdot \sqrt{2} \\ &= \sqrt{1+\gamma} \cdot \|\phi(i) - \phi(j)\| \\ &\geq (1 + \frac{\gamma}{4}) \cdot \|\phi(i) - \phi(j)\| \\ &= (1+\alpha) \cdot \|\phi(i) - \phi(j)\|,\end{aligned}$$

meaning that the triplet $(i, j^+, k_l^-)_\alpha$ is satisfied. In the above, construction there exist i, j, k with $j \neq k$ such that $\|\phi(i) - \phi(j)\| = \|\phi(i) - \phi(k)\|$. A small perturbation of all points resolves this issue while maintaining the conclusion.

□

C Additional Experiments

In this section we provide the additional experiments referenced from Section 4. All experiments run on a standard CPU and can be completed within an hour.

C.1 Synthetic α-sweep

In Section 4, we fixed α and varied d. Here we do the opposite: on the synthetic setup, we fix $d = 50$ to match the ground-truth ambient dimension, and sweep $\alpha \in \{0.1, 0.2, 0.4\}$, $n \in \{100, 300, 1000\}$, $m \in \{100, 300, 10^3, 3{\cdot}10^3, 10^4, 3{\cdot}10^4, 10^5\}$ over 5 seeds. In Figure 3, we plot the empirical error $\hat{\varepsilon}$ against the effective sample size $m\alpha^2/n$ predicted by Theorem 3.1. All nine (n, α) curves collapse onto a single trajectory and drop from random-guess error to near zero once $m\alpha^2/n \gtrsim 1$. If we plot against m/n instead, the curves separate by margin, so the α^2 factor in our obtained bound is quite precise.

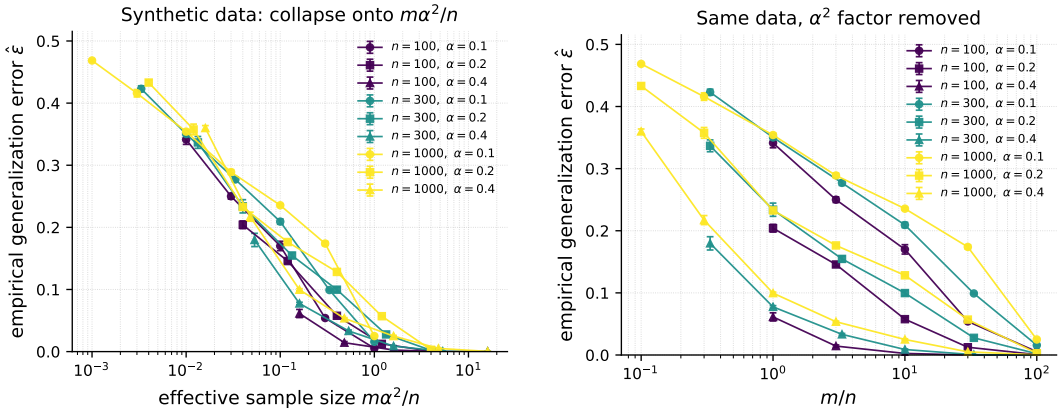


Figure 3: Synthetic α-sweep. Left: all nine (n, α) curves collapse when plotted against the predicted effective sample size $m\alpha^2/n$. Right: without the α^2 factor the curves separate by margin α.

C.2 Real data α-sweep

On the real-data setup, we fix $d = 64$, and sweep $\alpha \in \{0.1, 0.2, 0.4\}$, $n \in \{500, 2000, 5000\}$, $m \in \{300, 10^3, 3{\cdot}10^3, 10^4, 3{\cdot}10^4, 10^5\}$ over 5 seeds. Figure 4 reproduces both diagnostics: plotting against $m\alpha^2/n$ orders the curves and they decay together to zero error, while removing α^2 separates them by α. Compared to the synthetic data, two things look different: the empirical errors are well below their synthetic counterparts at the same $m\alpha^2/n$, and at fixed $m\alpha^2/n$ larger n gives *lower* rather than higher error. Both behaviors are what we would expect when the linear head

explores a more structured subspace of $\mathcal{H}(n, d, \alpha)$, rather than randomly generated data (it has only $O(d \cdot 512)$ parameters, independent of n). A smaller class generalizes better than our worst-case bound predicts, and at fixed $m\alpha^2/n$, a larger value of n means more triplets constraining the same fixed-size parameter matrix.

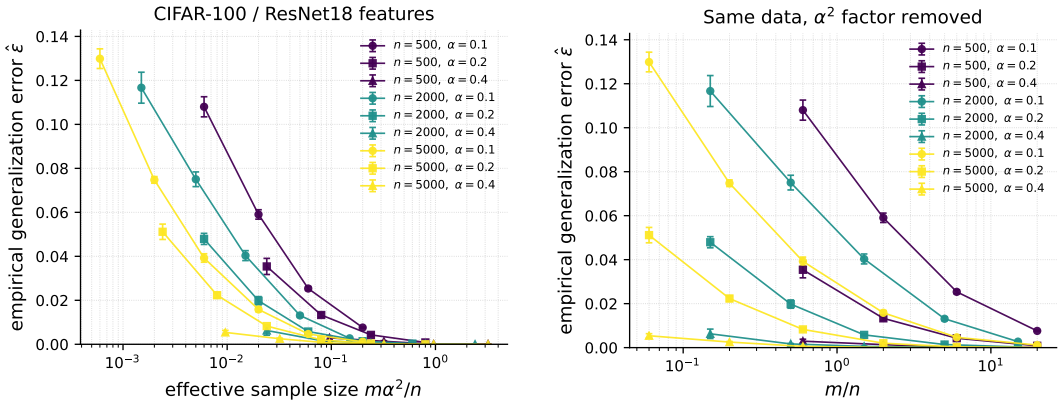


Figure 4: Real-data α-sweep. Left: collapse against $m\alpha^2/n$. Right: with the α^2 factor removed.

C.3 Dimension-independence across α

The dimension-independence experiment in Section 4 uses a single representative margin $\alpha = 0.2$. We repeat this experiment at $\alpha \in \{0.05, 0.1, 0.4\}$ on the synthetic and real-data setups (Figure 5). The three d-curves overlap within each n-trajectory across the full range in both cases. The one exception is mild underfitting at $d = 16$ when α is small and $m\alpha^2/n$ is large: 16 dimensions cannot faithfully embed the ground truth ($d^\star = 50$ synthetic, up to 512 on CIFAR-100), and tight margins are more sensitive to the resulting distortion.

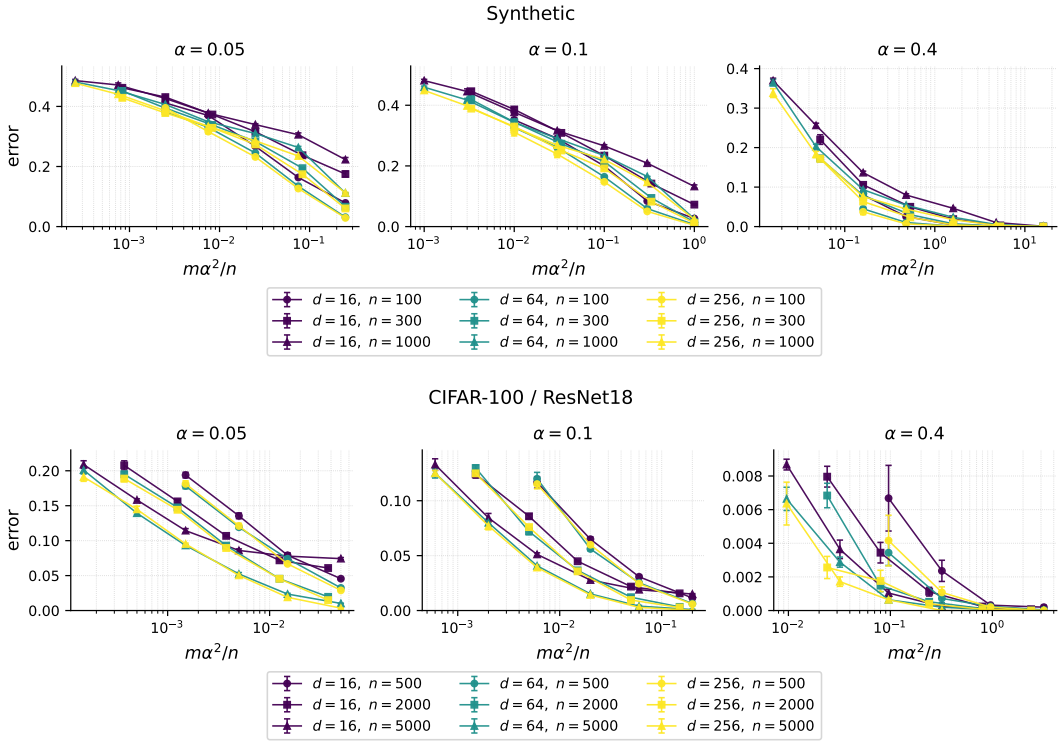


Figure 5: Dimension-independence across α. Top: synthetic setup. Bottom: CIFAR-100 / ResNet18.

C.4 MARGIN SANITY CHECK

The hypothesis class $\mathcal{H}(n, d, \alpha)$ is defined by embeddings whose output itself satisfies the $(1+\alpha)$-separation; the multiplicative-margin loss is a continuous surrogate for this constraint. As a post-hoc check that the learned $\hat{\phi}$ actually *lives in the hypothesis class*, we measure the fraction of held-out triplets on which the learned embedding satisfies $\|\hat{\phi}(i)-\hat{\phi}(k^-)\| \geq (1+\alpha)\|\hat{\phi}(i)-\hat{\phi}(j^+)\|$ (Figure 6). The fraction climbs from around 0.5 at the smallest m to above 0.95 once $m\alpha^2/n \gtrsim 1$. So whenever the empirical error is small, the learned embedding really is $(1+\alpha)$-separated on most test triplets.

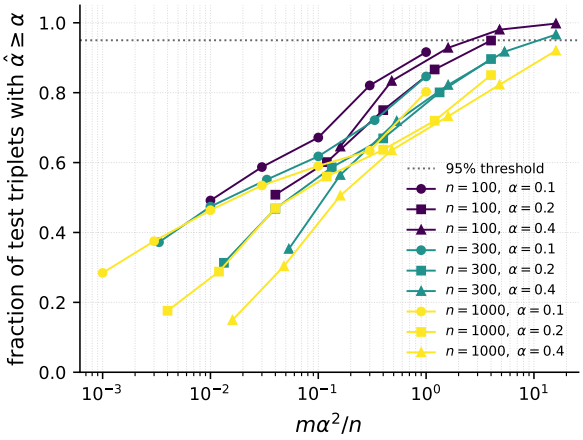


Figure 6: Margin sanity check. Once $m\alpha^2/n \gtrsim 1$, the learned embedding satisfies the hypothesis-class constraint on $>95\%$ of test triplets.