Analysis of Using Sigmoid Loss for Contrastive Learning

Chungpa Lee
Yonsei University

Joonhwan Chang
Yonsei University

Jy-yong Sohn
Yonsei University

Abstract

Contrastive learning has emerged as a prominent branch of self-supervised learning for several years. Especially, CLIP, which applies contrastive learning to large sets of captioned images, has garnered significant attention. Recently, SigLIP, a variant of CLIP, has been proposed, which uses the sigmoid loss instead of the standard InfoNCE loss. SigLIP achieves the performance comparable to CLIP in a more efficient manner by eliminating the need for a global view. However, theoretical understanding of using the sigmoid loss in contrastive learning is underexplored. In this paper, we provide a theoretical analysis of using the sigmoid loss in contrastive learning, in the perspective of the geometric structure of learned embeddings. First, we propose the *double-Constant Embedding Model* (CCEM), a framework for parameterizing various well-known embedding structures by a single variable. Interestingly, the proposed CCEM is proven to contain the optimal embedding with respect to the sigmoid loss. Second, we mathematically analyze the optimal embedding minimizing the sigmoid loss for contrastive learning. The optimal embedding ranges from simplex equiangular-tight-frame to antipodal structure, depending on the temperature parameter used in the sigmoid loss. Third, our experimental results on synthetic datasets coincide with the theoretical results on the optimal embedding structures.

1 Introduction

Contrastive learning (CL), which learns the embedding space based on the comparison between training data, is recently considered as one of the popular method for learning good representations (Liu et al., 2021; Jaiswal et al., 2020; Chen et al., 2020a; Tian et al., 2020). The basic idea is, to train an encoder such that similar training data (called positive pairs) are mapped to closer embedding vectors, while the embedding vectors for dissimilar training data (called negative pairs) are far apart. For example, in Chen et al. (2020a), different augmentations of the same image are labeled as a positive pair, while augmentations of different images are labeled as a negative pair. In Radford et al. (2021), the caption text and the image extracted from a single captioned image are considered as a positive pair, while those extracted from different captioned images are considered as a negative pair. In general, different *views* (augmentations or modalities) on a single entity form a positive pair, while views on different entities form a negative pair.

Interestingly, self-supervised learning trained by such comparison-based criterion alone, without any labeled data, is sufficient to learn meaningful features (He et al., 2020; Chen et al., 2020a; Radford et al., 2021; Jia et al., 2021; Girdhar et al., 2023), recently outperforming the features learned by supervised learning. Various ways of defining contrastive loss have been considered in the past years (Schroff et al., 2015; Oh Song et al., 2016; Sohn, 2016; Gutmann and Hyvärinen, 2010, 2012; Hyvarinen et al., 2019), see Appendix A for more details. Especially, one of the pioneering works on multi-modal CL called CLIP (Radford et al., 2021) uses the InfoNCE loss (Oord et al., 2018), which is theoretically analyzed in recent papers (Saunshi et al., 2019, 2022; Wang and Isola, 2020; Wang and Liu, 2021; Chen et al., 2021; Arora et al., 2019). Above all, Lu and Steinerberger (2022) shows that the optimal solution of the variational problem of minimizing cross-entropy loss, which coincides with the InfoNCE loss, is structured by the simplex equiangular-tight-frame (ETF) if feature vectors are in a sufficiently high-dimensional space.

Recently, a variant of CLIP called SigLIP is proposed (Zhai et al., 2023), which uses the sigmoid loss instead of the InfoNCE loss, motivated by the fact that

computing with the sigmoid loss is much more efficient than using the InfoNCE loss. More precisely, CL using the InfoNCE loss reveals remarkable performance only if the negative sample size, which is limited by its mini-batch size, is sufficiently large (Chen et al., 2020a; Zhang et al., 2022a). However, the necessity for a larger batch size imposes the computation burdens on CL using the InfoNCE loss. On the other hand, learning CL with SigLIP is independent of batch size, enabling comparable performance with significantly reduced computational demands. Unfortunately, there is no theoretical discussions on the behavior of SigLIP.

Our Contributions In this paper, we have taken the first step towards understanding SigLIP, or in general, using the sigmoid loss for contrastive learning. Specifically, we have the following contributions, in terms of analyzing the embeddings learned by the sigmoid loss.

- We define the *double-Constant Embedding Model* (CCEM), a framework that models various structures formed by multiple embedding vectors. This model contains well-known structures including the simplex ETF structure and the antipodal structure.

- We show that optimal embedding which minimizes the sigmoid loss can be found in the proposed CCEM. In other words, it is sufficient to find the optimal solution within the structures that can be represented in the CCEM framework.

- The optimal embedding structure that minimizes sigmoid loss is specified. The optimal embedding forms simplex ETF when the temperature parameter is sufficiently large, while it forms an antipodal structure when the temperature is smaller than a threshold.

- Our experimental results on synthetic datasets support our theoretical findings on the behavior of optimal embeddings.

2 Related Works

Contrastive Learning Self-supervised learning has gained significant attention as it enables learning from large-scale data without the need for laborious supervision labeling (Goyal et al., 2019; Chen et al., 2020a; He et al., 2020; Chen et al., 2020b). Among various techniques, CL has been widely utilized in computer vision (Caron et al., 2020; Chen and He, 2021; Grill et al., 2020) and natural language processing (Giorgi et al., 2020; Wu et al., 2020) domains due to the excellent performance of learned embeddings. CL is a similarity-based training method, aiming to position similar data close to each other in the embedding space

while pushing dissimilar data far apart. Moreover, it is known that CL can be applied in multi-modal scenarios (Radford et al., 2021; Jia et al., 2021; Ma et al., 2020; Elizalde et al., 2023; Li et al., 2021; Mu et al., 2022; Yao et al., 2021; Goel et al., 2022; Zhai et al., 2023), which leverage data with multiple views from different modalities such as images paired with text captions. In this paper, we provide analysis on the CL, in the perspective of the learned embedding structure, which can be applied for both uni-modal and multi-modal scenarios.

Analysis on Geometry of Embeddings Several researchers focused on the geometric structure of optimal embeddings learned by CL such as antipodal and simplex ETF (Lu and Steinerberger, 2022; Awasthi et al., 2022; Sreenivasan et al., 2023; Graf et al., 2021). Especially, Lu and Steinerberger (2022) shows that the optimal feature vector solution of a variational problem related to cross-entropy loss forms a simplex ETF, based on previous results on the ETFs (Sustik et al., 2007; Fickus and Mixon, 2015; Strohmer and Heath Jr, 2003). On this basis, Sreenivasan et al. (2023) demonstrated that optimal representations for minimizing the contrastive loss form a simplex ETF when $d \geq N - 1$. Our paper is similar to this work in terms of finding optimal embeddings for minimizing contrastive losses. While existing works solve the optimization problem for softmax-based losses, our paper solves the problem for the sigmoid loss. In addition, we define a novel embedding structure model which includes the simplex ETF and antipodal structure in a single framework, which helps find the embeddings minimizing the sigmoid loss.

Analysis on Temperature Parameter of CL Various theoretical studies have explored on the temperature parameter of CL (Zhang et al., 2021, 2022a,b). Especially, Wang and Liu (2021) shows that the temperature parameter of CL controls the strength of penalties on the hard negative sample by using gradient analysis. Kukleva et al. (2023) introduces a schedule for the temperature parameter of CL to improve the representation quality. However, previous research focused on the role of the temperature parameter during the learning process without considering the optimal embeddings from different temperature parameters. Particularly, we demonstrated that the use of the sigmoid loss results in embeddings that are highly dependent on the temperature parameter.

3 Problem Setup

Here we first provide backgrounds on CL problem we are focusing on, and introduce several notations necessary for defining our problem setup. Then, we provide

the formal description of our target problem of optimizing the embedding with the sigmoid loss, and describe the skeleton of our paper.

3.1 Backgrounds

CL uses N paired instances $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N$, where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ for the same index i are two different views or modalities of the same object. The goal of CL is to train encoders f_x and f_y in a way that the output of the encoders for different views ($\boldsymbol{x}_i$ and $\boldsymbol{y}_i$) of the same object are mapped into similar embedding vectors, *i.e.*, $f_x(\boldsymbol{x}_i) \approx f_y(\boldsymbol{y}_i)$.

Denote the collections of embedding vectors by $\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_N]$ and $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_N]$ where $\boldsymbol{u}_i = f_x(\boldsymbol{x}_i) \in \mathbb{R}^d$ and $\boldsymbol{v}_i = f_y(\boldsymbol{y}_i) \in \mathbb{R}^d$ for $i \in [N]$. For mathematical simplicity, we use the normalized embedding space where $\boldsymbol{u}_i^\top \boldsymbol{u}_i = \boldsymbol{v}_i^\top \boldsymbol{v}_i = 1$ for all $i \in [N]$, which is commonly used in related works (Awasthi et al., 2022; Sreenivasan et al., 2023; Huang et al., 2023).

Here we focus on two popular contrastive losses: the InfoNCE loss

$$
\begin{aligned}
\mathcal{L}^{\text{InfoNCE}}(\boldsymbol{U}, \boldsymbol{V}) := &-\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\boldsymbol{u}_i^\top \boldsymbol{v}_i / t)}{\sum_{j=1}^N \exp(\boldsymbol{u}_i^\top \boldsymbol{v}_j / t)} \\
&-\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\boldsymbol{u}_i^\top \boldsymbol{v}_i / t)}{\sum_{j=1}^N \exp(\boldsymbol{u}_j^\top \boldsymbol{v}_i / t)}
\end{aligned}
\tag{1}
$$

considered in Oord et al. (2018); Radford et al. (2021), and the sigmoid loss

$$
\begin{aligned}
\mathcal{L}^{\text{sig}}(\boldsymbol{U}, \boldsymbol{V}) := &-\frac{1}{N} \sum_{i=1}^N \log \frac{1}{1 + \exp(-t\boldsymbol{u}_i^\top \boldsymbol{v}_i + b)} \\
&-\frac{1}{N} \sum_{i=1}^N \sum_{j \in [N] \setminus \{i\}} \log \frac{1}{1 + \exp(t\boldsymbol{u}_i^\top \boldsymbol{v}_j - b)}
\end{aligned}
\tag{2}
$$

proposed by Zhai et al. (2023) as an alternative to the InfoNCE loss $\mathcal{L}^{\text{InfoNCE}}$ in Eq. 1 for the purpose of efficient computation. Here, $t > 0$ and $b \geq 0$ are temperature and bias, respectively.

Under this setting, one important question is about finding the optimal embedding vectors

$$
(\boldsymbol{U}^\star, \boldsymbol{V}^\star) := \arg\min_{\boldsymbol{U}, \boldsymbol{V}} \mathcal{L}(\boldsymbol{U}, \boldsymbol{V})
$$

learned by a pre-defined contrastive loss $\mathcal{L}$. A recent work (Lu and Steinerberger, 2022) shows that the optimal embedding vectors for the InfoNCE loss form a simplex ETF defined below:

Definition 1. *Suppose $d \geq N - 1$. A set of N vectors $\{\boldsymbol{w}_i\}_{i=1}^N$ in $\mathbb{R}^d$ is called $(N-1)$-simplex ETF, if*

$$
\begin{aligned}
\boldsymbol{w}_i^\top \boldsymbol{w}_i &= 1, \quad \forall i \in [N], \\
\boldsymbol{w}_i^\top \boldsymbol{w}_j &= -\frac{1}{N-1}, \quad \forall i,j \in [N] \text{ with } i \neq j.
\end{aligned}
$$

Corollary 1 (Theorem 1 of Lu and Steinerberger (2022)). *Let $(\boldsymbol{U}^\star, \boldsymbol{V}^\star) = \arg\min_{\boldsymbol{U}, \boldsymbol{V}} \mathcal{L}^{\text{InfoNCE}}(\boldsymbol{U}, \boldsymbol{V})$. Suppose $d \geq N - 1$, where d is the dimension of embedding vectors $\boldsymbol{u}_i^\star, \boldsymbol{v}_i^\star$. Then, $\{\boldsymbol{u}_i^\star\}_{i=1}^N$ forms a $(N-1)$-simplex ETF and $\boldsymbol{u}_i^\star = \boldsymbol{v}_i^\star$ holds for all $i \in [N]$.*

Note that Corollary 1 slightly differs from Theorem 1 of Lu and Steinerberger (2022), which demonstrates that the optimal embedding vectors for the InfoNCE loss $\mathcal{L}^{\text{InfoNCE}}$ *without any temperature parameter t* form a simplex ETF. However, even with the inclusion of a temperature parmeter t in the InfoNCE loss $\mathcal{L}^{\text{InfoNCE}}$, the result remains unchanged. The proof is on Appendix B.3.

We define the antipodal structure as follows.

Definition 2. *A set of $2N$ vectors $\{\boldsymbol{u}_i\}_{i=1}^N$ and $\{\boldsymbol{v}_i\}_{i=1}^N$ in $\mathbb{R}^d$ is called antipodal, if*

$$
\boldsymbol{u}_i = -\boldsymbol{v}_j, \quad \forall i,j \in [N].
$$

Note that if vectors $\{\boldsymbol{u}_i, \boldsymbol{v}_i\}_{i=1}^N$ form the antipodal structure, then the positive pairs are heading to the opposite direction, while the negative pairs are aligned, which is achieving the opposite of what contrastive learning is aiming at.

3.2 Skeleton of This Paper

In this paper, we analyze the embedding $\boldsymbol{U}^\star, \boldsymbol{V}^\star$ optimized by the sigmoid loss $\mathcal{L}^{\text{sig}}$ in Eq. 2, denoted by

$$
(\boldsymbol{U}^\star, \boldsymbol{V}^\star) = \arg\min_{\boldsymbol{U}, \boldsymbol{V}} \mathcal{L}^{\text{sig}}(\boldsymbol{U}, \boldsymbol{V}).
$$

Before solving this problem, in Sec. 4, we first propose the *double-Constant Embedding Model* (CCEM), which models the evolution of different embedding vectors using a single parameter $\delta \geq 0$. Note that this model contains various embedding structures including simplex ETF. Interestingly, under some mild assumption (which holds in practice) on the contrastive loss, our proposed model is shown to contain the optimal embedding $(\boldsymbol{U}^\star, \boldsymbol{V}^\star)$.

Based on this observation, in Sec. 5, we use the proposed CCEM to explore the optimal embedding for sigmoid loss. Our mathematical results specify the optimal embedding for various hyperparameter settings. Furthermore, our theoretical findings are supported by experimental results on toy datasets.

Remark 1. *Note that our analysis focuses on the case of $d \geq N$, similar to related works (Lu and Steinerberger, 2022; Sreenivasan et al., 2023). This is because when $d < N - 1$, the ETF solution, which is connected with the optimal solution for our target CL, does not have a standard format for all N, d values (Sustik et al., 2007; Fickus et al., 2012; Fickus and Mixon, 2015; Azarija and Marc, 2018).*

4 Double-Constant Embedding Model

In this section, we propose a framework that models various types of geometric structures formed by embedding vectors $\{(\boldsymbol{u}_i, \boldsymbol{v}_i)\}_{i=1}^N$. Since it is not trivial to model arbitrary structures of $2N$ embedding vectors, here we restrict ourselves to model a set of vectors satisfying

$$\boldsymbol{u}_i^\top \boldsymbol{v}_i = c_1 \quad \text{for all } i \in [N], \tag{3}$$
$$\boldsymbol{u}_i^\top \boldsymbol{v}_j = c_2 \quad \text{for all } i, j \in [N] \text{ with } i \neq j$$

for some $c_1, c_2 \in [-1, 1]$, *i.e.*, the matrix M containing elements $M_{ij} = \boldsymbol{u}_i^\top \boldsymbol{v}_j$ is *double-constant* (O'Neill, 2021), where all diagonal elements are c_1 and all off-diagonal elements are c_2. Interestingly, as shown later in Corollary 2, the optimal embedding $(\boldsymbol{U}^\star, \boldsymbol{V}^\star)$ minimizing the sigmoid loss satisfies the double-constant property in Eq. 3, which shows that we can safely search the embedding vectors represented by our framework, without losing the optimality. With this in mind, we now introduce the *double-Constant Embedding Model* (CCEM).

Definition 3 (Double-Constant Embedding Model)**.** *Suppose $d \geq N$. Let $\{\boldsymbol{w}_i\}_{i=1}^N$ be the $(N-1)$-simplex ETF in $\mathbb{R}^{d-1}$. The double-Constant Embedding Model (CCEM) constructs $2N$ vectors $(\{\boldsymbol{u}_i^\delta, \boldsymbol{v}_i^\delta\}_{i=1}^N)$ as*

$$\boldsymbol{u}_i^\delta := \begin{pmatrix} \frac{1}{\sqrt{1+\delta^2}} \boldsymbol{w}_i \\ \frac{1}{\sqrt{1+\delta^2}} \delta \end{pmatrix} \in \mathbb{R}^d, \quad \boldsymbol{v}_i^\delta := \begin{pmatrix} \frac{1}{\sqrt{1+\delta^2}} \boldsymbol{w}_i \\ -\frac{1}{\sqrt{1+\delta^2}} \delta \end{pmatrix} \in \mathbb{R}^d,$$

which are parameterized by $\delta \geq 0$. The corresponding embedding matrix is denoted by

$$\boldsymbol{U}(\delta) = [\boldsymbol{u}_1^\delta, \cdots, \boldsymbol{u}_N^\delta], \quad \boldsymbol{V}(\delta) = [\boldsymbol{v}_1^\delta, \cdots, \boldsymbol{v}_N^\delta].$$

4.1 Properties and Examples of CCEM

Our proposed model has properties summarized below:

Proposition 1. *Let $\{\boldsymbol{u}_i^\delta, \boldsymbol{v}_i^\delta\}_{i=1}^N$ be $2N$ embedding vectors formed by the CCEM in Def. 3, for a given $\delta \geq 0$. The inner products between embedding vectors are*

$$(\boldsymbol{u}_i^\delta)^\top \boldsymbol{v}_i^\delta = \frac{1 - \delta^2}{1 + \delta^2} \quad \forall i \in [N], \tag{4}$$

$$(\boldsymbol{u}_i^\delta)^\top \boldsymbol{v}_j^\delta = -\frac{\frac{1}{N-1} + \delta^2}{1 + \delta^2} \quad \forall i, j \in [N] \text{ with } i \neq j,$$

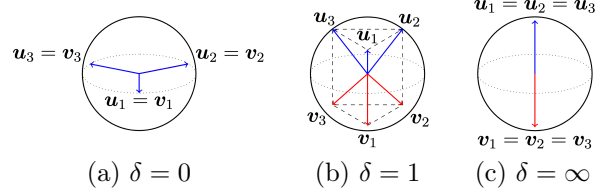(a) $\delta = 0$ (b) $\delta = 1$ (c) $\delta = \infty$

Figure 1: The proposed double-constant embedding model in Def. 3 for different δ values, when we have $N = 3$ embedding vector pairs $\{(\boldsymbol{u}_i^\delta, \boldsymbol{v}_i^\delta)\}_{i=1}^N$ in $d = 3$ dimensional space. (1) $\delta = 0$ (simplex ETF), the angle between positive pair ($\boldsymbol{u}_i^0$ and $\boldsymbol{v}_i^0$) is 0. (2) $\delta = 1$, the angle between $\boldsymbol{u}_i^1$ and $\boldsymbol{v}_i^1$ is $\frac{\pi}{2}$. (3) $\delta = \infty$, the angle between $\boldsymbol{u}_i^\infty$ and $\boldsymbol{v}_i^\infty$ is π, forming 'antipodal' embedding.

i.e., the double-constant property in Eq. 3 is satisfied. When $\delta = 0$, the embedding vectors satisfy

$$\{\boldsymbol{u}_i^0\}_{i=1}^N \text{ form } (N-1)\text{-simplex ETF},$$
$$\boldsymbol{u}_i^0 = \boldsymbol{v}_i^0 \quad \forall i \in [N].$$

When $\delta = \infty$, the embedding vectors satisfy

$$\boldsymbol{u}_i^\infty = -\boldsymbol{v}_j^\infty \quad \forall i, j \in [N]$$

which is 'antipodal' embedding defined in Def. 2.

Below we provide an example of embeddings formed by our model, when $d = 3$ and $N = 3$.

Example 1. *Fig. 1 illustrates an example of the CCEM proposed in Def. 3, where three embedding vector pairs ($N = 3$) are in three dimensional space ($d = 3$). When $\delta = 0$, the embedding structure model follows precisely the simplex ETF, which coincides with the result of Prop. 1. As δ increases, the angle between positive pairs ($\boldsymbol{u}_i, \boldsymbol{v}_i$) increases for all $i \in [N]$, and the distance between negative pairs ($\boldsymbol{u}_i, \boldsymbol{v}_j$) also increases for all $i \neq j$. This means that increasing δ is pushing both positive pairs and negative pairs apart. In the extreme case of $\delta = \infty$, we have $\boldsymbol{u}_i = -\boldsymbol{v}_j$ for all i, j, meaning that pairs with positive labels ($\boldsymbol{u}_i$ and $\boldsymbol{v}_i$) are ignored and pairs with negative labels ($\boldsymbol{u}_i$ and $\boldsymbol{v}_j \; \forall i \neq j$) dominate the behavior. Thus, we call this case as the 'antipodal' structure.*

4.2 Optimal Embeddings are in CCEM

Now we provide theoretical results supporting that the CCEM proposed in Def. 3 is sufficient to find the optimal embedding that minimizes the sigmoid loss. We start with a theorem stating that the optimal embedding of CL can be found in the CCEM, under a mild condition on the loss function.

Theorem 1. *Consider an arbitrary contrastive loss that follows the form of*

$$\mathcal{L}(\boldsymbol{U}, \boldsymbol{V}) = \sum_{i=1}^{N} l_{\mathrm{p}}(\boldsymbol{u}_i^\top \boldsymbol{v}_i) + \sum_{i=1}^{N} \sum_{j \in [N] \setminus \{i\}} l_{\mathrm{n}}(\boldsymbol{u}_i^\top \boldsymbol{v}_j), \quad (5)$$

where l_{p} is an arbitrary convex decreasing function for positive pairs, and l_{n} is an arbitrary convex increasing function for negative pairs.

Then, the optimal embeddings $(\boldsymbol{U}^\star, \boldsymbol{V}^\star) = \arg\min_{\boldsymbol{U}, \boldsymbol{V}} \mathcal{L}(\boldsymbol{U}, \boldsymbol{V})$ can be uniquely found in the CCEM, i.e.,

$$\exists! \delta^\star \geq 0 \ such\ that\ (\boldsymbol{U}(\delta^\star), \boldsymbol{V}(\delta^\star)) = (\boldsymbol{U}^\star, \boldsymbol{V}^\star).$$

Note that in the above theorem, we assume the loss function $\mathcal{L}$ satisfies Eq. 5. This is quite mild assumption, since various loss functions used in practice satisfy the condition. For example, the sigmoid loss $\mathcal{L}^{\mathrm{sig}}$ in Eq. 2 is in the form of Eq. 5 (as shown below), thus the above theorem implies the following.

Corollary 2. *Consider the sigmoid loss $\mathcal{L}^{\mathrm{sig}}(\boldsymbol{U}, \boldsymbol{V})$ defined in Eq. 2. Then,*

$$\min_{\boldsymbol{U}, \boldsymbol{V}} \mathcal{L}^{\mathrm{sig}}(\boldsymbol{U}, \boldsymbol{V}) = \min_{\delta \geq 0} \mathcal{L}^{\mathrm{sig}}(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta)).$$

In other words, the optimal embedding vectors for the sigmoid loss can be found in the CCEM.

Proof. The sigmoid loss $\mathcal{L}^{\mathrm{sig}}$ in Eq. 2 can be reformulated as Eq. 5, when $l_{\mathrm{p}}(x) = \log(1 + \exp(-tx + b))$ and $l_{\mathrm{n}}(x) = \log(1 + \exp(tx - b))$ for a given temperature $t > 0$ and bias $b \geq 0$. Combining this with Theorem 1 completes the proof. □

Now we provide a proof of Theorem 1. In order to establish Theorem 1, we introduce the mathematical result, which demonstrates that the CCEM defined in Def. 3 is a structure that minimizes the sum of the negative pairs' inner product, for a given sum of the positive pairs' inner product:

Lemma 1. *Suppose $d \geq N$. Let $\{\boldsymbol{u}_i, \boldsymbol{v}_i\}_{i=1}^{N}$ be $2N$ distinct vectors in $\mathbb{R}^d$, satisfying $\boldsymbol{u}_i^\top \boldsymbol{u}_i = \boldsymbol{v}_i^\top \boldsymbol{v}_i = 1$ for all $i \in [N]$. Then, for every $c \in [-1, 1]$, there exists unique $\delta \geq 0$ such that*

$$(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta)) = \arg\min_{(\boldsymbol{U}, \boldsymbol{V}) \in S_{(c)}} \sum_{i=1}^{N} \sum_{j \in [N] \setminus \{i\}} \boldsymbol{u}_i^\top \boldsymbol{v}_j. \quad (6)$$

where $S_{(c)} = \left\{ (\boldsymbol{U}, \boldsymbol{V}) : \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i = c \right\}$

Proof. See Appendix B.1. □

It should be pointed out that there could be other embedding vectors that satisfy the minimization of Eq. 6. However, the unique existence of δ in Lemma 1 implies that there is a unique CCEM structure satisfying the minimization property of Eq. 6. This is crucial as we require the double-constant properties of the CCEM in Eq. 4 to prove Theorem 1 as below.

Proof of Theorem 1

Proof. Given two embedding matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, define $\boldsymbol{M}$ with $M_{ij} = \boldsymbol{u}_i^\top \boldsymbol{v}_j$ as the similarity matrix of two embeddings. Define the function h of $c \in [-1, 1]$ as

$$h(c) := \min_{\left\{ (\boldsymbol{U}, \boldsymbol{V}) : \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i = c \right\}} \sum_{i=1}^{N} \sum_{j \in [N] \setminus \{i\}} \boldsymbol{u}_i^\top \boldsymbol{v}_j,$$

which is the minimum summation of the off-diagonal terms of similarity matrix $\boldsymbol{M}$, among embedding matrices $\boldsymbol{U}, \boldsymbol{V}$ satisfying that the sum of diagonal terms is Nc.

To simplify the notation, define

$$l(x, y) := l_{\mathrm{p}}\left(\frac{x}{N}\right) + (N-1) l_{\mathrm{n}}\left(\frac{y}{N(N-1)}\right).$$

By using Jensen's inequality on the loss $\mathcal{L}$ in Eq. 5,

$$\frac{1}{N}\mathcal{L} \geq l_{\mathrm{p}}\left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i\right) + \frac{1}{N}\sum_{i=1}^{N}\sum_{j \in [N]\setminus\{i\}} l_{\mathrm{n}}(\boldsymbol{u}_i^\top \boldsymbol{v}_j)$$

$$\geq l_{\mathrm{p}}\left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i\right)$$

$$+ (N-1) \cdot l_{\mathrm{n}}\left(\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j \in [N]\setminus\{i\}} \boldsymbol{u}_i^\top \boldsymbol{v}_j\right)$$

$$= l\left(\sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i, \ \sum_{i=1}^{N}\sum_{j \in [N]\setminus\{i\}} \boldsymbol{u}_i^\top \boldsymbol{v}_j\right)$$

$$\geq l\left(\sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i, \ h\left(\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i\right)\right) \quad (7)$$

where the equality conditions of Jensen's inequalities (1st and 2nd inequalities in the above equation) are

$$\boldsymbol{u}_i^\top \boldsymbol{v}_i = c_1 \quad \forall i, \quad (8)$$
$$\boldsymbol{u}_i^\top \boldsymbol{v}_j = c_2 \quad \forall i, j \in [N] \text{ with } i \neq j.$$

The last inequality holds since $l(x, y)$ is an increasing function with respect to y. For every $c \in [-1, 1]$, by Lemma 1, there is a unique $\delta = \delta(c) \geq 0$ such that

$$(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta)) = \arg\min_{(\boldsymbol{U}, \boldsymbol{V}) \in S_{(c)}} \sum_{i=1}^{N}\sum_{j \in [N]\setminus\{i\}} \boldsymbol{u}_i^\top \boldsymbol{v}_j$$

where $S_{(c)} = \left\{ (\boldsymbol{U}, \boldsymbol{V}) : \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i = c \right\}$. It implies

$$\sum_{i=1}^{N} \sum_{j \in [N] \setminus \{i\}} \left(\boldsymbol{u}_i^\delta \right)^\top \boldsymbol{v}_j^\delta = h \left(\frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{u}_i^\delta \right)^\top \boldsymbol{v}_i^\delta \right). \quad (9)$$

Moreover, every CCEM fulfills the equality conditions given in Eq. 8 , due to its double-constant properties in Eq. 4. Thus, combining these facts with Eq. 9, the inequalities of Eq. 7 reduces to equality as

$$\mathcal{L}(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta)) \hspace{3cm} (10)$$
$$= N \cdot l \left(\sum_{i=1}^{N} \left(\boldsymbol{u}_i^\delta \right)^\top \boldsymbol{v}_i^\delta, h \left(\frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{u}_i^\delta \right)^\top \boldsymbol{v}_i^\delta \right) \right).$$

Since $\delta = \delta(c)$ is a function of $c = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i$, from Eq. 7, we have

$$\mathcal{L} \geq N \cdot l \left(\sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i, \ h \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i \right) \right)$$
$$\geq N \min_{\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i} l \left(\sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i, \ h \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i \right) \right)$$
$$= N \min_{\delta \geq 0} l \left(\sum_{i=1}^{N} \left(\boldsymbol{u}_i^\delta \right)^\top \boldsymbol{v}_i^\delta, \ h \left(\frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{u}_i^\delta \right)^\top \boldsymbol{v}_i^\delta \right) \right)$$
$$= \min_{\delta \geq 0} \mathcal{L}(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta))$$

where the last equality holds from Eq. 10.

Therefore, the optimal embeddings $(\boldsymbol{U}(\delta^\star), \boldsymbol{V}(\delta^\star))$ found from the CCEM minimize the contrastive loss over all possible embeddings.

$\square$

Remark 2. *The proposed* CCEM *can be effectively utilized not only for the loss in Eq. 5 format, but also for many other contrastive losses, e.g., InfoNCE loss. Similar to Corollary 2, it can be shown that the optimal embedding vectors for InfoNCE loss can be found in* CCEM*, as formally stated below:*

Corollary 3. *Consider InfoNCE loss* $\mathcal{L}^{\mathrm{InfoNCE}}(\boldsymbol{U}, \boldsymbol{V})$ *defined in Eq. 1. Then,*

$$\min_{\boldsymbol{U}, \boldsymbol{V}} \mathcal{L}^{\mathrm{InfoNCE}}(\boldsymbol{U}, \boldsymbol{V}) = \min_{\delta \geq 0} \mathcal{L}^{\mathrm{InfoNCE}}(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta)).$$

Proof. See Appendix B.2. $\square$

5 Embedding Learned by Sigmoid Loss

In this section, we explore the optimal embedding learned by the sigmoid loss using the suggested CCEM. In Sec. 5.1, we provide theoretical results explaining

how the structure of optimal embedding varies depending on the temperature and the bias term used in the sigmoid loss, assuming $d \geq N$. The behavior proven in Sec. 5.1 is empirically confirmed in Sec. 5.2, not only for $d \geq N$ case but also for $d < N$ case.

5.1 Theoretical Results

Recall that Corollary 2 implies that the optimal embedding vectors $\boldsymbol{U}^\star, \boldsymbol{V}^\star$ can be found by solving

$$\min_{(\boldsymbol{U}, \boldsymbol{V}) \in S} \mathcal{L}^{\mathrm{sig}}(\boldsymbol{U}, \boldsymbol{V}),$$

which is equivalent to finding the optimal parameter

$$\delta^\star := \arg \min_{\delta \geq 0} \mathcal{L}^{\mathrm{sig}}(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta)).$$

The below theorem provides the behavior of $\delta^\star$:

Theorem 2 (Optimal Embedding for sigmoid loss). *Consider the case when $t = b$, i.e., the temperature and the bias are identical.*

When $N = 3$, we have $\delta^\star = 0$, i.e., the optimal embedding for sigmoid loss forms a simplex ETF, regardless of t value.

On the other hand, when $N \geq 4$, the optimal $\delta^\star$ is a monotonic decreasing function of t, satisfying

$$\delta^\star = \begin{cases} 0, & \text{if and only if } t > \frac{N-1}{N} \log(N-3) \\ \infty, & \text{if and only if } t < \frac{1}{2} \log \frac{N-2}{2}. \end{cases}$$

Proof. See Appendix B.4. $\square$

Remark 3. *The above theorem implies that the optimal embeddings form a simplex ETF when the temperature is sufficiently large, while they reduce to the 'antipodal' embedding when the temperature is smaller than a threshold. Note that the threshold temperature value is in the order of $\Theta(\log N)$.*

Remark 4. *According to the empirical results reported in Zhai et al. (2023), SigLIP enjoys high performance only when the temperature parameter is set to a sufficiently large number. This coincides with our result in Theorem 2 stating that sufficiently large t is necessary to let the trained embeddings form a simplex ETF with $\delta^\star = 0$, instead of the antipodal structure with $\delta^\star = \infty$.*

Note that Theorem 2 implies that for a given N, the optimal $\delta^\star$ (representing the distance between positive pairs) decreases as the temperature parameter $t = b$ increases, leading to a tighter alignment of the positive pairs, i.e., $\boldsymbol{u}_i^\top \boldsymbol{v}_i$ approaches to 1. Now we provide a corollary relating the optimal embeddings obtained by InfoNCE loss and sigmoid loss.
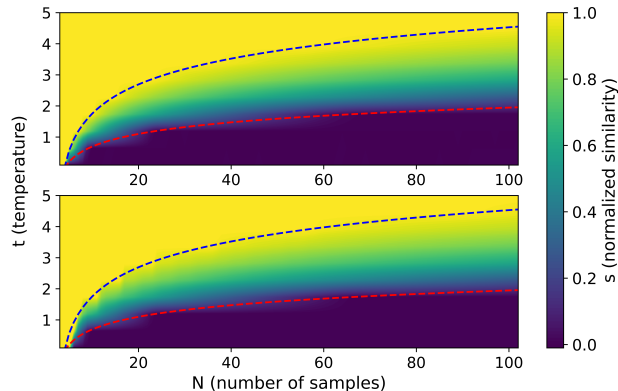
Figure 2: The normalized similarity s of positive pairs measured for the embeddings trained by the sigmoid loss $\mathcal{L}^{\text{sig}}$, for various N and t. We test for two cases: $d = N$ (upper) and $d = \frac{N}{2}$ (below). The blue dashed line satisfies $t = \frac{N-1}{N} \log(N-3)$, and the red dashed line satisfies $t = \frac{1}{2} \log \frac{N-2}{2}$, which are the threshold values obtained in Theorem 2. This plot shows that our theoretical results coincide with empirical observations.

Corollary 4. *Consider the case when $t = b$. When $N = 3$, the optimal solution of minimizing $\mathcal{L}^{\text{sig}}$ is equal to that of minimizing $\mathcal{L}^{\text{InfoNCE}}$, for any temperature parameter t. When $N \geq 4$, both solutions are identical, if and only if $t > \frac{N-1}{N} \log(N-3)$.*

Proof. Recall that the optimal embedding for InfoNCE loss $\mathcal{L}^{\text{InfoNCE}}$ always follows the simplex ETF irrespective of the temperature parameter t, as in Corollary 1. Combining this with Theorem 2 completes the proof. □

Therefore, by using an appropriate temperature parameter $t > \frac{N-1}{N} \log(N-3)$, the simplex ETF solution is attainable by training with the sigmoid loss (Zhai et al., 2023), which is shown to be more efficient than training with InfoNCE loss.

5.2 Experimental results

Now we provide empirical observations supporting our theoretical results. Specifically, we test the behavior of the optimal embedding trained by the sigmoid loss for various settings, and check whether it coincides with our result in Theorem 2. Although our theoretical results hold only when $d \geq N$, we test whether the behavior of optimal embeddings coincides with the result of Theorem 2 for two cases: $d = N$ and $d = \frac{N}{2}$.

Synthetic data without Encoder We train embedding vectors $\{\boldsymbol{u}_i, \boldsymbol{v}_i\}_{i=1}^N$, each of which is initialized by sampling from a standard normal distribution
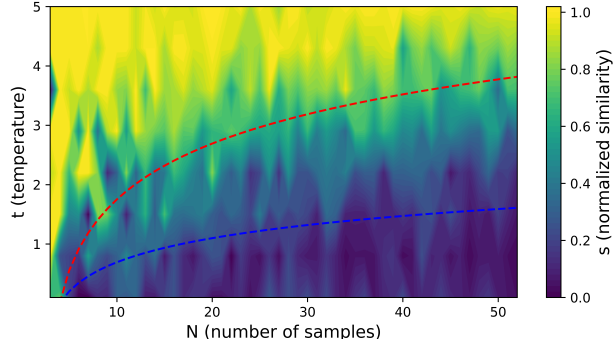

Figure 3: The normalized similarity $s = \frac{1}{2}(1 + \frac{1}{N} \sum_{i=1}^N \boldsymbol{u}_i^\top \boldsymbol{v}_i)$ of positive pairs measured for the embeddings trained by sigmoid loss $\mathcal{L}^{\text{sig}}$, for various N and t when $d = N$. Unlike the setting for Fig 2, we train a encoder (two-layer fully-connected ReLU network) which outputs embeddings, rather than directly optimizing the embedding vectors.

$\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, followed by projecting it to a unit sphere. Unlike real-world scenarios, we do not use encoder structures; instead, we directly optimize the embedding vectors $\{\boldsymbol{u}_i, \boldsymbol{v}_i\}_{i=1}^N$ by training them for $50,000$ steps with the sigmoid loss, using learning rate $\gamma = 0.5$. During the training process, the updated embeddings are projected back to the unit sphere, to make sure that $\boldsymbol{u}_i^\top \boldsymbol{u}_i = \boldsymbol{v}_i^\top \boldsymbol{u}_i = 1$ for all $i \in [N]$. Here, we test various temperature parameters $t = b$, ranging from 0.1 to 5.

Given the trained embeddings $\{\boldsymbol{u}_i, \boldsymbol{v}_i\}_{i=1}^N$, we measure the normalized similarity of positive pairs, denoted by $s = \frac{1}{2}(1 + \frac{1}{N} \sum_{i=1}^N \boldsymbol{u}_i^\top \boldsymbol{v}_i)$. Note that $s = 1$ holds when all positive pairs are aligned, while $s = 0$ holds when all positive pairs are antipodal.

Fig. 2 shows the normalized similarity s for various N and t. The above plot is the result for $d = N$, while the below plot is for $d = \lceil N/2 \rceil$. We have two dashed lines visualizing the theoretical threshold values obtained in Theorem 2. The blue line satisfies $t = \frac{N-1}{N} \log(N-3)$, the minimum value allowing the optimal embedding to be the simplex ETF having $s = 1$. The red line satisfies $t = \frac{1}{2} \log \frac{N-2}{2}$, the maximum value allowing the optimal embedding to be the antipodal solution having $s = 0$. Thus, our theoretical results claim that the set of (N, t) points above the blue line have $s = 1$, while the points below the red line have $s = 0$, which coincide with the empirical results in Fig. 2.

Synthetic data with Encoder Furthermore, we extend our experiments to more realistic setting where we have an encoder that generates the embedding vectors. In contrast to the previous synthetic experiments where the encoder output is directly optimized, we

employ a two-layer fully-connected ReLU network denoted as f. Especially, N pair of instances $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^N$ are generated from a standard normal distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{2d})$, which passes through the encoder f that transforms each $2d$-dimensional vector ($\boldsymbol{x}_i$ or $\boldsymbol{y}_i$) into a d-dimensional embedding vector ($f(\boldsymbol{x}_i)$ or $f(\boldsymbol{y}_i)$) on a unit sphere. Then, the encoder is trained to minimize the sigmoid loss $\mathcal{L}^{\mathrm{sig}}$. Fig. 3 shows the similarity of paired embeddings as a function of the temperature parameter t used in the sigmoid loss, when $d = N$. This result exhibits a tendency similar to Fig 2, implying that the empirical results in a realistic setting also align with the theoretical findings.

Remark 5. *Our theoretical results are based on the assumption of $d \geq N$, while our empirical results have similar behaviors for both $d \geq N$ and $d < N$ cases. This empirical finding can be interpreted as follows. Recall that when d is sufficiently large (i.e., $d \geq N$), the learned embeddings satisfy both uniformity and alignment (Wang and Isola, 2020), while the uniformity is not attainable for smaller d (especially, $d < N$). Our empirical results, which measure the alignment of positive pairs, imply that smaller d does not harm the alignment property, although the uniformity is broken.*

6   Conclusion

This paper investigated the embedding structure learned by contrastive learning with the sigmoid loss, motivated by the recent success of SigLIP, a variant of CLIP that uses sigmoid loss instead of InfoNCE loss. During the route towards answering the question of finding the optimal embedding for sigmoid loss, we first defined the *double-Constant Embedding Model* (CCEM) framework that models various embedding structures. Based on our observation that the optimal embeddings are in the CCEM, we used the CCEM to specify the optimal embedding structure learned by the sigmoid loss. When the temperature parameter is sufficiently large, the optimal embedding forms a simplex ETF structure, while the optimal embedding reduces to the antipodal structure when the temperature parameter is set to be a small number, which coincides with the empirical observations made by the authors of SigLIP. We also conducted experiments on synthetic datasets, the results of which coincide with our theoretical results on the behavior of optimal embeddings.

Acknowledgement

References

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. (2019). A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*.

Awasthi, P., Dikkala, N., and Kamath, P. (2022). Do more negative samples necessarily hurt in contrastive learning? In *International Conference on Machine Learning*, pages 1101–1116. PMLR.

Azarija, J. and Marc, T. (2018). There is no (75, 32, 10, 16) strongly regular graph. *Linear Algebra and its Applications*, 557:62–83.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. (2020b). Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255.

Chen, T., Luo, C., and Li, L. (2021). Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845.

Chen, X. and He, K. (2021). Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758.

El-Gebeily, M. and Fiagbedzi, Y. (2004). On certain properties of the regular n-simplex. *International Journal of Mathematical Education in Science and Technology*, 35(4):617–629.

Elizalde, B., Deshmukh, S., Al Ismail, M., and Wang, H. (2023). Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Fickus, M., Jasper, J., Mixon, D. G., Peterson, J. D., and Watson, C. E. (2018). Equiangular tight frames with centroidal symmetry. *Applied and Computational Harmonic Analysis*, 44(2):476–496.

Fickus, M. and Mixon, D. G. (2015). Tables of the existence of equiangular tight frames. *arXiv preprint arXiv:1504.00253*.

Fickus, M., Mixon, D. G., and Tremain, J. C. (2012). Steiner equiangular tight frames. *Linear algebra and its applications*, 436(5):1014–1027.

Gale, D. (1953). On inscribing n-dimensional sets in a regular n-simplex. *Proceedings of the American Mathematical Society*, 4(2):222–225.

Giorgi, J., Nitski, O., Wang, B., and Bader, G. (2020). Declutr: Deep contrastive learning for unsupervised textual representations. *arXiv preprint arXiv:2006.03659*.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. (2023). Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Goel, S., Bansal, H., Bhatia, S., Rossi, R., Vinay, V., and Grover, A. (2022). Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719.

Goyal, P., Mahajan, D., Gupta, A., and Misra, I. (2019). Scaling and benchmarking self-supervised visual representation learning. In *Proceedings of the ieee/cvf International Conference on computer vision*, pages 6391–6400.

Graf, F., Hofer, C., Niethammer, M., and Kwitt, R. (2021). Dissecting supervised contrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Gutmann, M. U. and Hyvärinen, A. (2012). Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of machine learning research*, 13(2).

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Huang, W., Yi, M., Zhao, X., and Jiang, Z. (2023). Towards the generalization of contrastive self-supervised learning.

Hyvarinen, A., Sasaki, H., and Turner, R. (2019). Nonlinear ica using auxiliary variables and generalized

contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. (2020). A survey on contrastive self-supervised learning. *Technologies*, 9(1):2.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Kukleva, A., Böhle, M., Schiele, B., Kuehne, H., and Rupprecht, C. (2023). Temperature schedules for self-supervised contrastive methods on long-tail data. *arXiv preprint arXiv:2303.13664*.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. (2021). Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. (2021). Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876.

Lu, J. and Steinerberger, S. (2022). Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241.

Ma, S., Zeng, Z., McDuff, D., and Song, Y. (2020). Active contrastive learning of audio-visual video representations. *arXiv preprint arXiv:2009.09805*.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. (2022). Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer.

Oh Song, H., Xiang, Y., Jegelka, S., and Savarese, S. (2016). Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012.

O'Neill, B. (2021). The double-constant matrix, centering matrix and equicorrelation matrix: Theory and applications. *arXiv preprint arXiv:2109.05814*.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Saunshi, N., Ash, J., Goel, S., Misra, D., Zhang, C., Arora, S., Kakade, S., and Krishnamurthy, A. (2022). Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pages 19250–19286. PMLR.

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., and Khandeparkar, H. (2019). A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Sohn, K. (2016). Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Sreenivasan, K., Lee, K., Lee, J.-G., Lee, A., Cho, J., Sohn, J.-y., Papailiopoulos, D., and Lee, K. (2023). Mini-batch optimization of contrastive loss. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Strohmer, T. and Heath Jr, R. W. (2003). Grassmannian frames with applications to coding and communication. *Applied and computational harmonic analysis*, 14(3):257–275.

Sustik, M. A., Tropp, J. A., Dhillon, I. S., and Heath Jr, R. W. (2007). On the existence of equiangular tight frames. *Linear Algebra and its applications*, 426(2-3):619–635.

Tian, Y., Sun, C., Poole, B., Krishnan, D., Schmid, C., and Isola, P. (2020). What makes for good views for contrastive learning? *Advances in neural information processing systems*, 33:6827–6839.

Wang, F. and Liu, H. (2021). Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504.

Wang, T. and Isola, P. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Wu, Z., Wang, S., Gu, J., Khabsa, M., Sun, F., and Ma, H. (2020). Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., and Xu, C. (2021). Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid loss for language image pre-training. *arXiv preprint arXiv:2303.15343*.

Zhang, C., Zhang, K., Pham, T. X., Niu, A., Qiao, Z., Yoo, C. D., and Kweon, I. S. (2022a). Dual temperature helps contrastive learning without many negative samples: Towards understanding and simplifying moco. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14441–14450.

Zhang, C., Zhang, K., Zhang, C., Pham, T. X., Yoo, C. D., and Kweon, I. S. (2022b). How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning. *arXiv preprint arXiv:2203.16262*.

Zhang, O., Wu, M., Bayrooti, J., and Goodman, N. (2021). Temperature as uncertainty in contrastive learning. *arXiv preprint arXiv:2110.04403*.

A   Comparison of Various Contrastive Loss Functions

As metioned in Sec. 1, there have been various approaches of defining contrastive loss function in the past year. In this section, we compare and highlight the distinctive points of the three loss functions relevant to our paper; NCE loss $\mathcal{L}^{\mathrm{NCE}}$ (Gutmann and Hyvärinen, 2010, 2012), InfoNCE loss $\mathcal{L}^{\mathrm{InfoNCE}}$ (Oord et al., 2018; Radford et al., 2021), and Sigmoid loss $\mathcal{L}^{\mathrm{sig}}$ (Zhai et al., 2023).

To simplify the notation, we omit the instance pair index i, the temperature parameter, and the bias term. For a given embedding $\boldsymbol{u}$, we consider the the positive embedding $\boldsymbol{v}_+$ and the negative embeddings $\{\boldsymbol{v}_-^{(j)}\}_{j=1}^N$ with respect to $\boldsymbol{u}$. If there is only one negative embedding, we remove the index (j) and denote it as $\boldsymbol{v}_-$. Then, the three loss functions are defined as follows.

$$
\begin{aligned}
\mathcal{L}^{\mathrm{NCE}}(\boldsymbol{u}, \boldsymbol{V}) &= -\log\left(\frac{\exp(\boldsymbol{u}^\top \boldsymbol{v}_+)}{\exp(\boldsymbol{u}^\top \boldsymbol{v}_+) + \exp(\boldsymbol{u}^\top \boldsymbol{v}_-)}\right) \\
&= -\log\left(\frac{1}{1 + \exp\left(-\boldsymbol{u}^\top \boldsymbol{v}_+ + \boldsymbol{u}^\top \boldsymbol{v}_-\right)}\right) \\
\mathcal{L}^{\mathrm{InfoNCE}}(\boldsymbol{u}, \boldsymbol{V}) &= -\log\left(\frac{\exp(\boldsymbol{u}^\top \boldsymbol{v}_+)}{\exp(\boldsymbol{u}^\top \boldsymbol{v}_+) + \sum_j \exp\left(\boldsymbol{u}^\top \boldsymbol{v}_-^{(j)}\right)}\right) \\
&= -\log\left(\frac{1}{1 + \sum_j \exp\left(-\boldsymbol{u}^\top \boldsymbol{v}_+ + \boldsymbol{u}^\top \boldsymbol{v}_-^{(j)}\right)}\right) \\
\mathcal{L}^{\mathrm{sig}}(\boldsymbol{u}, \boldsymbol{V}) &= -\log\left(\frac{1}{1 + \exp\left(-\boldsymbol{u}^\top \boldsymbol{v}_+\right)}\right) - \sum_j \log\left(\frac{1}{1 + \exp\left(\boldsymbol{u}^\top \boldsymbol{v}_-^{(j)}\right)}\right)
\end{aligned}
$$

There are two points to note. First, $\mathcal{L}^{\mathrm{NCE}}$ has only one negative pair, whereas $\mathcal{L}^{\mathrm{InfoNCE}}$, an extension of $\mathcal{L}^{\mathrm{NCE}}$, includes multiple negative pairs, similar to $\mathcal{L}^{\mathrm{sig}}$. Second, the NCE-based loss functions, $\mathcal{L}^{\mathrm{NCE}}$ and $\mathcal{L}^{\mathrm{InfoNCE}}$, are derived from the softmax function, while $\mathcal{L}^{\mathrm{sig}}$ is formulated based on the sigmoid function. As a result, $\mathcal{L}^{\mathrm{sig}}$ can separate the term for the positive pair and negative pairs through summation, whereas the NCE-based losses cannot.

B   Proofs

Note that the $(N-1)$-regular-simplex is a set of N vectors which makes the geometry of simplex with equal edges. Detailed concepts and properties are on El-Gebeily and Fiagbedzi (2004); Gale (1953); Fickus et al. (2018); Sustik et al. (2007).

B.1   Proof of Lemma 1

Lemma. *Suppose $d \geq N$. Let $\{\boldsymbol{u}_i, \boldsymbol{v}_i\}_{i=1}^N$ be $2N$ distinct vectors in $\mathbb{R}^d$, satisfying $\boldsymbol{u}_i^\top \boldsymbol{u}_i = \boldsymbol{v}_i^\top \boldsymbol{v}_i = 1$ for all $i \in [N]$. Then, for every $c \in [-1, 1]$, there exists unique $\delta \geq 0$ such that*

$$
(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta)) = \underset{\left\{(\boldsymbol{U}, \boldsymbol{V}): \frac{1}{N}\sum_{i=1}^N \boldsymbol{u}_i^\top \boldsymbol{v}_i = c\right\}}{\arg\min} \sum_{i=1}^N \sum_{j \in [N] \setminus \{i\}} \boldsymbol{u}_i^\top \boldsymbol{v}_j. \tag{11}
$$

Proof. By using Jensen's inequality,

$$
\sum_{i=1}^{N} \|\boldsymbol{u}_i - \boldsymbol{v}_i\|_2^2 \geq N \left\| \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{u}_i - \boldsymbol{v}_i) \right\|_2^2
$$

$$
= N \left\| \frac{1}{N} \sum_{i=1}^{N} (\boldsymbol{u}_i + \boldsymbol{v}_i) \right\|_2^2 - 4N \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i \right)^\top \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{v}_i \right)
$$

$$
\geq -4N \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i \right)^\top \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{v}_i \right)
$$

where the equality conditions are

$$
\begin{cases} \boldsymbol{u}_i - \boldsymbol{v}_i = \boldsymbol{c}' \text{ for all } i \in [N], \text{ for some constant vector } \boldsymbol{c}' \\ \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i + \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{v}_i = 0 \end{cases} . \tag{12}
$$

This implies that

$$
\left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i \right)^\top \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{v}_i \right) \geq -\frac{1}{4N} \sum_{i=1}^{N} \|\boldsymbol{u}_i - \boldsymbol{v}_i\|_2^2. \tag{13}
$$

Note that the inner product of centroids is

$$
\left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i \right)^\top \left(\frac{1}{N} \sum_{j=1}^{N} \boldsymbol{v}_j \right) = \frac{1}{N^2} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \in [N] \setminus \{i\}} \boldsymbol{u}_i^\top \boldsymbol{v}_j.
$$

Since $\|\boldsymbol{u}_i - \boldsymbol{v}_i\|_2^2 = 2 - 2\boldsymbol{u}_i^\top \boldsymbol{v}_i$ for all $i \in [N]$, we have

$$
\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \in [N] \setminus \{i\}} \boldsymbol{u}_i^\top \boldsymbol{v}_j = \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i \right)^\top \left(\frac{1}{N} \sum_{j=1}^{N} \boldsymbol{v}_j \right) - \frac{1}{N^2} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i
$$

$$
= \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i \right)^\top \left(\frac{1}{N} \sum_{j=1}^{N} \boldsymbol{v}_j \right) + \frac{1}{2N^2} \sum_{i=1}^{N} \|\boldsymbol{u}_i - \boldsymbol{v}_i\|_2^2 - \frac{1}{N}
$$

$$
\geq -\frac{1}{4N} \sum_{i=1}^{N} \|\boldsymbol{u}_i - \boldsymbol{v}_i\|_2^2 + \frac{1}{2N^2} \sum_{i=1}^{N} \|\boldsymbol{u}_i - \boldsymbol{v}_i\|_2^2 - \frac{1}{N} \tag{14}
$$

$$
= \frac{N-2}{2N^2} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i - \frac{1}{2} \tag{15}
$$

where the inequality is from Eq. 13.

Recall that our target problem in Eq. 11 is about finding the optimal embedding matrices $(\boldsymbol{U}, \boldsymbol{V})$ that minimizes $\sum_{i=1}^{N} \sum_{j \in [N] \setminus \{i\}} \boldsymbol{u}_i^\top \boldsymbol{v}_j$, under the constraint of $\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i = c$. One can confirm that the necessary and sufficient condition for achieving this optimal $(\boldsymbol{U}, \boldsymbol{V})$ is nothing but the equality condition for Eq. 14, which is in Eq. 12.

Note that the construction of the CCEM in Def. 3 fulfills these equality conditions of Eq. 12, which also attains the minimum of Eq. 15, since

$$
\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \in [N] \setminus \{i\}} \left(\boldsymbol{u}_i^\delta \right)^\top \boldsymbol{v}_j^\delta = -\frac{\frac{1}{N} + \frac{N-1}{N} \delta^2}{1 + \delta^2}
$$

$$
= \frac{N-2}{2N^2} \sum_{i=1}^{N} \left(\boldsymbol{u}_i^\delta \right)^\top \boldsymbol{v}_i^\delta - \frac{1}{2}.
$$

holds from Prop. 1.

Thus, we can conclude that there exists unique δ such that $(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta))$ is the optimal embedding. Actually, we can specify such δ: from Eq. 4, $\delta = \sqrt{\frac{1-c}{1+c}}$ holds when $\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i = c$. The uniqueness of δ arises from the fact that $\sqrt{\frac{1-c}{1+c}}$ is a strictly decreasing function with respect to $c \in [-1, 1]$.

$\square$

B.2 Proof of Corollary 3

Corollary. *Consider InfoNCE loss* $\mathcal{L}^{\mathrm{InfoNCE}}(\boldsymbol{U}, \boldsymbol{V})$ *defined in Eq. 1. Then,*

$$\min_{\boldsymbol{U}, \boldsymbol{V}} \mathcal{L}^{\mathrm{InfoNCE}}(\boldsymbol{U}, \boldsymbol{V}) = \min_{\delta \geq 0} \mathcal{L}^{\mathrm{InfoNCE}}(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta)).$$

Proof. Define the function h of $c \in [-1, 1]$ as

$$h(c) := \min_{\left\{(\boldsymbol{U}, \boldsymbol{V}): \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i = c\right\}} \sum_{i=1}^{N} \sum_{j \in [N] \backslash \{i\}} \boldsymbol{u}_i^\top \boldsymbol{v}_j.$$

To simplify the notation, define the convex function

$$l(x) = \log \left(1 + (N-1) \exp \left(\frac{1}{tN(N-1)} x\right)\right)$$

for any $N \geq 3$ and $t > 0$.

By using Jensen's inequality on the loss $\mathcal{L}^{\mathrm{InfoNCE}}$ in Eq. 1,

$$
\begin{aligned}
\mathcal{L}^{\mathrm{InfoNCE}} &= \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + \sum_{j \in [N] \backslash \{i\}} \exp \left(\boldsymbol{u}_i^\top (\boldsymbol{v}_j - \boldsymbol{v}_i)/t\right)\right) + \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + \sum_{j \in [N] \backslash \{i\}} \exp \left((\boldsymbol{u}_j - \boldsymbol{u}_i)^\top \boldsymbol{v}_i/t\right)\right) \\
&\geq \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + (N-1) \exp \left(\frac{1}{N-1} \sum_{j \in [N] \backslash \{i\}} \boldsymbol{u}_i^\top (\boldsymbol{v}_j - \boldsymbol{v}_i)/t\right)\right) \\
&\quad + \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + (N-1) \exp \left(\frac{1}{N-1} \sum_{j \in [N] \backslash \{i\}} (\boldsymbol{u}_j - \boldsymbol{u}_i)^\top \boldsymbol{v}_i/t\right)\right) \\
&\geq \log \left(1 + (N-1) \exp \left(\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \in [N] \backslash \{i\}} \boldsymbol{u}_i^\top (\boldsymbol{v}_j - \boldsymbol{v}_i)/t\right)\right) \\
&\quad + \log \left(1 + (N-1) \exp \left(\frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \in [N] \backslash \{i\}} (\boldsymbol{u}_j - \boldsymbol{u}_i)^\top \boldsymbol{v}_i/t\right)\right) \\
&= 2l \left(\sum_{i=1}^{N} \sum_{j \in [N] \backslash \{i\}} \boldsymbol{u}_i^\top \boldsymbol{v}_j - (N-1) \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i\right) \\
&\geq 2l \left(h \left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i\right) - (N-1) \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i\right)
\end{aligned}
$$

where the equality conditions of Jensen's inequalities are $\boldsymbol{u}_i^\top (\boldsymbol{v}_j - \boldsymbol{v}_i) = c_1$ and $(\boldsymbol{u}_j - \boldsymbol{u}_i)^\top \boldsymbol{v}_i = c_2$ for all $i, j \in [N]$ with $i \neq j$. The last inequality holds since l is an increasing function.

For every $c \in [-1, 1]$, by Lemma 1, there is a unique $\delta(c) \geq 0$ such that

$$(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta)) := \underset{\left\{(\boldsymbol{U}, \boldsymbol{V}): \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i = c\right\}}{\arg \min} \sum_{i=1}^{N} \sum_{j \in [N] \backslash \{i\}} \boldsymbol{u}_i^\top \boldsymbol{v}_j$$

which implies

$$\sum_{i=1}^{N} \sum_{j \in [N] \backslash \{i\}} \left(\boldsymbol{u}_i^\delta\right)^\top \boldsymbol{v}_j^\delta = h\left(\frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{u}_i^\delta\right)^\top \boldsymbol{v}_i^\delta\right)$$

Moreover, every CCEM fulfills the equality conditions of Jensen's inequalities, due to its double-constant properties in Eq. 4. Thus,

$$\mathcal{L}^{\text{InfoNCE}}(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta)) = 2l\left(h\left(\frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{u}_i^\delta\right)^\top \boldsymbol{v}_i^\delta\right) - (N-1) \sum_{i=1}^{N} \left(\boldsymbol{u}_i^\delta\right)^\top \boldsymbol{v}_i^\delta\right)$$

Since δ is function of $\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i$, we have

$$\mathcal{L}^{\text{InfoNCE}} \geq 2l\left(h\left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i\right) - (N-1) \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i\right)$$

$$\geq 2 \min_{\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i} l\left(h\left(\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i\right) - (N-1) \sum_{i=1}^{N} \boldsymbol{u}_i^\top \boldsymbol{v}_i\right)$$

$$= 2 \min_{\delta \geq 0} l\left(h\left(\frac{1}{N} \sum_{i=1}^{N} \left(\boldsymbol{u}_i^\delta\right)^\top \boldsymbol{v}_i^\delta\right) - (N-1) \sum_{i=1}^{N} \left(\boldsymbol{u}_i^\delta\right)^\top \boldsymbol{v}_i^\delta\right)$$

$$= \min_{\delta \geq 0} \mathcal{L}^{\text{InfoNCE}}(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta)).$$

Thus, the optimal embeddings $(\boldsymbol{U}(\delta^\star), \boldsymbol{V}(\delta^\star))$ found from CCEM minimize the contrastive loss over all possible embeddings. $\qquad \square$

B.3 Proof of Corollary 1, using the proposed CCEM

Here we show that Corollary 1 proven by Lu and Steinerberger (2022), can be also proved using the CCEM (Double-constant Embedding Model) framework proposed by us in Def. 3.

Corollary (Equivalent to Theorem 1 of Lu and Steinerberger (2022)). *Let $\boldsymbol{U}, \boldsymbol{V}$ be the optimal embedding minimizing the InfoNCE loss $\mathcal{L}^{\text{InfoNCE}}$. Then $\boldsymbol{U} = \boldsymbol{V}$ holds and $\boldsymbol{U}$ forms a simplex ETF, regardless of the temperature parameter $t > 0$ and the number of paired instances $N \geq 3$.*

Proof. It is sufficient to use the CCEM in Def. 3 to find the minimum by Corollary 3. Then, by introducing the CCEM to the InfoNCE loss $\mathcal{L}^{\text{InfoNCE}}$ of Eq. 1,

$$\mathcal{L}^{\text{InfoNCE}}(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta)) = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\left(\boldsymbol{u}_i^\delta\right)^\top \boldsymbol{v}_i^\delta / t\right)}{\sum_{j=1}^{N} \exp\left(\left(\boldsymbol{u}_i^\delta\right)^\top \boldsymbol{v}_j^\delta / t\right)} - \frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp\left(\left(\boldsymbol{u}_i^\delta\right)^\top \boldsymbol{v}_i^\delta / t\right)}{\sum_{j=1}^{N} \exp\left(\left(\boldsymbol{u}_j^\delta\right)^\top \boldsymbol{v}_i^\delta / t\right)}$$

$$= 2 \log\left(1 + (N-1) \exp\left(-\frac{\frac{N}{N-1}}{1 + \delta^2} / t\right)\right)$$

which is strictly monotonic increasing function with respect to δ for all $t > 0$ and $N \geq 3$.

Therefore, $\mathcal{L}^{\text{InfoNCE}}$ minimize at $\delta = 0$ for all $t > 0$ and $N \geq 3$. $\qquad \square$

B.4 Proof of Theorem 2

Theorem (Optimal Embedding for sigmoid loss). *Consider the case when $t = b$, i.e., the temperature and the bias are identical. When $N = 3$, we have $\delta^\star = 0$, i.e., the optimal embedding for sigmoid loss forms a simplex ETF, regardless of t value. On the other hand, when $N \geq 4$, the optimal $\delta^\star$ is a monotonic decreasing function of t, satisfying*

$$\delta^\star = \begin{cases} 0, & \text{if and only if } t > \frac{N-1}{N} \log(N-3) \\ \infty, & \text{if and only if } t < \frac{1}{2} \log \frac{N-2}{2}. \end{cases}$$

Proof. It is sufficient to use the CCEM in Def. 3 to find the minimum by Corollary 2. Then, by introducing the CCEM of Def. 3 to the sigmoid loss $\mathcal{L}^{\text{sig}}$ of Eq. 2,

$$\mathcal{L}^{\text{sig}}(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta)) = \frac{1}{N} \sum_{i=1}^{N} \log\left(1 + \exp\left(-t\left(\boldsymbol{u}_i^\delta\right)^\top \boldsymbol{v}_i^\delta + b\right)\right) + \frac{1}{N} \sum_{i=1}^{N} \sum_{j \in [N] \backslash \{i\}} \log\left(1 + \exp\left(t\left(\boldsymbol{u}_i^\delta\right)^\top \boldsymbol{v}_j^\delta - b\right)\right)$$

$$= \log\left(1 + \exp\left(-t\frac{1-\delta^2}{1+\delta^2} + b\right)\right) + (N-1)\log\left(1 + \exp\left(-t\frac{\frac{1}{N-1} + \delta^2}{1+\delta^2} - b\right)\right)$$

$$= \log\left(1 + \exp\left(-t\frac{-2\delta^2}{1+\delta^2}\right)\right) + (N-1)\log\left(1 + \exp\left(-t\frac{\frac{N}{N-1} + 2\delta^2}{1+\delta^2}\right)\right).$$

Then,

$$\frac{\partial}{\partial\delta}\mathcal{L}^{\text{sig}} = \frac{\exp\left(-t\frac{-2\delta^2}{1+\delta^2}\right)}{1 + \exp\left(-t\frac{-2\delta^2}{1+\delta^2}\right)}\left(-t\frac{-4\delta}{(1+\delta^2)^2}\right) + (N-1)\frac{\exp\left(-t\frac{\frac{N}{N-1}+2\delta^2}{1+\delta^2}\right)}{1 + \exp\left(-t\frac{\frac{N}{N-1}+2\delta^2}{1+\delta^2}\right)}\left(-t\frac{\frac{2(N-2)}{N-1}\delta}{(1+\delta^2)^2}\right)$$

$$= \frac{2t\delta}{(1+\delta^2)^2}\left(\frac{2}{1 + \exp\left(t\frac{-2\delta^2}{1+\delta^2}\right)} - \frac{(N-2)}{1 + \exp\left(t\frac{\frac{N}{N-1}+2\delta^2}{1+\delta^2}\right)}\right)$$

$$= \frac{2t\delta}{(1+\delta^2)^2}\frac{g(N, t, \delta)}{\left(1 + \exp\left(t\frac{-2\delta^2}{1+\delta^2}\right)\right)\left(1 + \exp\left(t\frac{\frac{N}{N-1}+2\delta^2}{1+\delta^2}\right)\right)}$$

where we define $g(N, t, \delta) = (4 - N) + 2\exp\left(t\frac{\frac{N}{N-1}+2\delta^2}{1+\delta^2}\right) - (N-2)\exp\left(t\frac{-2\delta^2}{1+\delta^2}\right)$.

Note that $g(N, t, \delta)$ is strictly monotone increasing with respect to $\delta > 0$ for all $t > 0$ and $N \geq 3$, since

$$\frac{\partial}{\partial\delta}g(N, t, \delta) = \frac{4t(N-2)\delta}{(1+\delta^2)^2}\left(\frac{1}{N-1}\exp\left(t\frac{\frac{N}{N-1}+2\delta^2}{1+\delta^2}\right) + \exp\left(t\frac{-2\delta^2}{1+\delta^2}\right)\right)$$

$$> 0.$$

Also, note that $\frac{\partial}{\partial\delta}\mathcal{L}^{\text{sig}}\big|_{\delta=0} = 0$, $\frac{\partial}{\partial\delta}\mathcal{L}^{\text{sig}}\big|_{\delta\to\infty} = 0$, and $\frac{\partial^2}{\partial\delta^2}\mathcal{L}^{\text{sig}}\big|_{\delta=0} = 2t\frac{(3-N)+\exp\left(t\frac{N}{N-1}\right)}{1+\exp\left(t\frac{N}{N-1}\right)}$.

To find out the condition t of N that minimize $\mathcal{L}^{\text{sig}}$, we split the cases for $N = 3$ and $N \geq 4$.

[$N = 3$ case] First, consider $N = 3$. Then, for all $t > 0$,

$$\frac{\partial^2}{\partial\delta^2}\mathcal{L}^{\text{sig}}\bigg|_{\delta=0, N=3} = 2t\frac{\exp\left(t\frac{3}{2}\right)}{1 + \exp\left(t\frac{3}{2}\right)}$$

$$> 0$$

Furthermore,

$$g(3, t, \delta) = 1 + 2 \exp\left(t \frac{\frac{3}{2} + 2\delta^2}{1 + \delta^2}\right) - \exp\left(t \frac{-2\delta^2}{1 + \delta^2}\right)$$

$$= 1 + 2 \exp\left(t\left(2 - \frac{\frac{1}{2}}{1 + \delta^2}\right)\right) - \exp\left(t\left(-2 + \frac{2}{1 + \delta^2}\right)\right)$$

$$\geq 2 \exp\left(\frac{3}{2}t\right)$$

$$> 0$$

where the inequality holds by plugging $\delta = 0$, because it is monotone increasing with respect to δ^2. Since g is non-negative for all $t, \delta > 0$, it is also hold that $\frac{\partial}{\partial \delta} \mathcal{L}^{\mathrm{sig}}\big|_{\delta=0, N=3} > 0$ for all $\delta > 0$. Therefore, $0 = \arg\min_\delta \mathcal{L}^{\mathrm{sig}}(\boldsymbol{U}(\delta), \boldsymbol{V}(\delta))$ for $N = 3$.

[$N \geq 4$ case] Second, consider the case of $N \geq 4$.

For any $t > \frac{N-1}{N} \log(N-3)$, $\frac{\partial^2}{\partial \delta^2} \mathcal{L}^{\mathrm{sig}}\big|_{\delta=0} > 0$ holds and

$$g(N, t, \delta) \geq g(N, t, 0)$$

$$= 2(3 - N) + 2 \exp\left(t \frac{N}{N-1}\right)$$

$$> 0$$

where the first inequality comes from the fact that g is strictly monotone increasing with respect to $\delta \geq 0$. Since g is non-negative for above case, it is also hold that $\frac{\partial}{\partial \delta} \mathcal{L}^{\mathrm{sig}} > 0$ for all $\delta > 0$. As a result, $\mathcal{L}^{\mathrm{sig}}$ minimize at $\delta = 0$.

On the other hand, for $0 < t < \frac{1}{2} \log \frac{N-2}{2}$,

$$g(N, t, \delta) \leq g(N, t, \infty)$$

$$= (4 - N) + 2 \exp(2t) - (N - 2) \exp(-2t)$$

$$< 0$$

holds. Therefore, $\frac{\partial}{\partial \delta} \mathcal{L}^{\mathrm{sig}} < 0$ for all $\delta > 0$, which implies $\mathcal{L}^{\mathrm{sig}}$ minimize at $\delta = \infty$.

Finally, for $\frac{1}{2} \log \frac{N-2}{2} < t < \frac{N-1}{N} \log(N-3)$, by applying the inequality opposite to the above, we can easily get $g(N, t, 0) < 0$ and $g(N, t, \infty) > 0$. Since g is strictly increasing continuous function with respect to δ, there is unique $\delta^\star \in (0, \infty)$ such that $g(N, t, \delta^\star) = 0$ by the intermediate value theorem. Therefore, there is unique $\delta^\star \in (0, \infty)$ such that minimize $\mathcal{L}^{\mathrm{sig}}$.

Furthermore, since $\frac{\partial}{\partial \delta} g(N, t, \delta) > 0$ and $\frac{\partial}{\partial t} g(N, t, \delta) > 0$ holds for all $t, \delta > 0$ and $N \geq 3$, by taking partial derivative of $g(N, t, \delta^\star) = 0$ for fixed N,

$$0 = \left(\frac{\partial}{\partial t} g(N, t, \delta^\star)\right) + \left(\frac{\partial \delta^\star}{\partial t}\right) \left(\frac{\partial}{\partial \delta^\star} g(N, t, \delta^\star)\right)$$

implies $\frac{\partial \delta^\star}{\partial t} < 0$ for all $t, \delta > 0$ and $N \geq 3$. Therefore, the optimal $\delta^\star$ is a monotonic decreasing function of t for fixed N. $\square$