A Theoretical Framework for Preventing Class Collapse in Supervised Contrastive Learning

Chungpa Lee
Yonsei University

Jeongheon Oh
Bank of Korea
Yonsei University[*]

Kibok Lee
Yonsei University

Jy-yong Sohn[†]
Yonsei University

Abstract

Supervised contrastive learning (SupCL) has emerged as a prominent approach in representation learning, leveraging both supervised and self-supervised losses. However, achieving an optimal balance between these losses is challenging; failing to do so can lead to class collapse, reducing discrimination among individual embeddings in the same class. In this paper, we present theoretically grounded guidelines for SupCL to prevent class collapse in learned representations. Specifically, we introduce the Simplex-to-Simplex Embedding Model (SSEM), a theoretical framework that models various embedding structures, including all embeddings that minimize the supervised contrastive loss. Through SSEM, we analyze how hyperparameters affect learned representations, offering practical guidelines for hyperparameter selection to mitigate the risk of class collapse. Our theoretical findings are supported by empirical results across synthetic and real-world datasets.

1 INTRODUCTION

Contrastive learning (CL) has recently demonstrated significant advancements in self-supervised representation learning, where data representations are learned by comparing views generated by augmentations of the training data (Chen et al., 2020; He et al., 2020; Radford et al., 2021). Specifically, CL maximizes the similarity between positive pairs, derived from different views of the same instance, while simultaneously minimizing the similarity between negative pairs, derived from views of distinct instances. While self-supervised CL does not leverage available supervision, such as class labels in classification tasks, incorporating such information can be advantageous for representation learning. To take advantage of this, Khosla et al. (2020) introduced supervised contrastive learning (SupCL), which extends self-supervised CL by treating views from different instances within the same class as positive pairs. Furthermore, Islam et al. (2021) empirically analyzed the effectiveness of SupCL through experiments on the transferability of representations learned both with and without supervision.

However, recent studies have pointed out that optimizing the supervised contrastive loss often results in *class collapse*, where the embedding vectors of all instances within the same class converge into the same point in the embedding space (Graf et al., 2021; Papyan et al., 2020). Class collapse significantly degrades the generalizability of learned representations by eliminating within-class variance that is potentially crucial for effective transfer learning (Islam et al., 2021; Chen et al., 2022). While several follow-up studies have explored the conditions under which class collapse occurs and proposed strategies to prevent it (Feng et al., 2022; Chen et al., 2022; Wang et al., 2023; Xue et al., 2023), their analyses are often limited to specific conditions, such as assumptions about data distribution. The precise mechanism underlying class collapse in SupCL is not yet fully understood.

To this end, we provide a theoretical analysis of the behavior of embeddings that minimize the supervised contrastive loss (SupCL loss), offering guidelines on how to avoid class collapse. Our analysis applies to a broad range of data configurations, including varying numbers of classes, instances, and augmentations, whenever the SupCL loss is given as a convex combination of the supervised and self-supervised contrastive losses. Our contributions are summarized as follows:

- In Sec. 4, we propose the Simplex-to-Simplex Embedding Model (SSEM), a theoretical framework for modeling diverse embedding structures. We prove that embeddings $U^\star$ that minimize the SupCL loss can only be found within SSEM. This result establishes SSEM as a fundamental tool for analyzing embeddings learned in SupCL.

- In Sec. 5, we provide guidelines for designing the SupCL loss to mitigate class collapse in learned embeddings. In particular, we derive a mathematical expression for the variance of embeddings $U^\star$, a key metric for assessing class collapse when the within-class variance is zero. Furthermore, we characterize the relationship between embedding variances and the hyperparameters.

- In Sec. 6, we present experiments on both synthetic and real-world datasets, demonstrating that our theoretical findings hold in practice. Specifically, the variance of the learned embeddings aligns with our theoretical predictions, allowing us to identify optimal hyperparameters in SupCL that balance within-class and between-class variance, leading to improved transfer learning performance.

2 RELATED WORK

SupCL Methods. SupCL leverages supervised information by jointly utilizing the supervised contrastive loss and the InfoNCE-based self-supervised loss (Khosla et al., 2020), demonstrating superior performance compared to existing CL methods that only use self-supervised loss (Islam et al., 2021; Gunel et al., 2021). To further enhance SupCL, Chen et al. (2022) replace the InfoNCE-based self-supervised loss with the class-conditional InfoNCE loss. Meanwhile, Feng et al. (2022) refine the supervised contrastive loss by selecting only the k nearest neighbors within each class as positive pairs, while Wang et al. (2023) introduce hierarchical supervision to balance instance-level and class-level information. Building on these approaches, Oh and Lee (2024) incorporate supervision into asymmetric non-CL methods (Grill et al., 2020; Chen and He, 2021) to further improve representation learning.

Despite these advancements, most existing methods require extensive hyperparameter tuning to achieve optimal performance. Rather than proposing a new SupCL method, this paper conducts a theoretical analysis on the role of hyperparameters in the SupCL loss function, ensuring that searching within a narrower region is sufficient and thus enabling more efficient hyperparameter optimization.

Analysis on SupCL. Although considerable theoretical research has been conducted on CL (Arora et al., 2019; Parulekar et al., 2023; Wen and Li, 2021; Yang et al., 2023), the understanding of SupCL remains underexplored. Notably, Xue et al. (2023) showed that the bias of gradient descent can cause subclass representation collapse and suppress harder class-relevant features. However, their analysis is specific to spectral contrastive loss (HaoChen et al., 2021), which is rarely used in practice. Moreover, most existing studies rely on strict assumptions about data distribution, limiting their applicability to real-world datasets. Instead of focusing on a specialized loss function or imposing such assumptions, we directly optimize the widely adopted SupCL loss (Khosla et al., 2020; Islam et al., 2021; Oh and Lee, 2024), ensuring broader applicability across diverse datasets.

Understanding SupCL Through Embedding Structures. Several previous works aim at understanding the optimal embedding structures minimizing the supervised loss and/or contrastive loss. In the supervised learning setup, one well-known phenomenon of optimal embeddings is neural collapse (Papyan et al., 2020), where the embeddings collapse to a simplex Equiangular Tight Frame (ETF). Similarly, in the case of CL, Lu and Steinerberger (2022) showed that the optimal embeddings minimizing the softmax-based contrastive loss construct the simplex ETF, and Lee et al. (2024) generalized the result to the cases of minimizing other CL losses including the sigmoid-based loss (Zhai et al., 2023). For SupCL, the optimal embeddings that minimize the supervised contrastive loss result in class collapse, where embeddings of the same class collapse to a single point of simplex ETF (Graf et al., 2021).

Notably, Chen et al. (2022) introduced the class-conditional InfoNCE loss (Oord et al., 2018) to spread out the embedding vectors within each class, and combined it with the supervised contrastive loss. They demonstrated that non-collapsed embeddings can achieve the lower value of the combined loss than collapsed ones in the specific case where the number of classes is two or three, without identifying the optimal embeddings. Although Chen et al. (2022) provided a meaningful direction of designing the SupCL loss in a way that the optimal embeddings do not suffer from class collapse, this work cannot be extended to general cases when the number of classes is more than three. In addition, none of existing works examined the behavior of the optimal embeddings in the SupCL setup, specifically regarding how to construct a loss function that consistently avoids class collapse.

Building on prior works, this paper specifies the optimal embeddings that minimize a convex combination of supervised and self-supervised losses, providing guidelines for selecting hyperparameters to avoid class collapse in SupCL.

3 PROBLEM FORMULATION

We consider the problem of training an encoder f that maps the feature $\boldsymbol{x}$ into the embedding $\boldsymbol{u} = f(\boldsymbol{x})$ by using SupCL. The training sample is categorized into m classes, and the sample size for each class is n. Every instance is augmented, *i.e.,* generating similar instances by data augmentation techniques; the number of augmentation for each instance is denoted by p. We use the notation $\boldsymbol{x}_{i,j,k}$ to represent the feature of k-th augmentation of j-th instance in i-th class for $i \in [m], j \in [n]$ and $k \in [p]$, where we define $[m] := \{1, 2, \cdots, m\}$ for positive integer m.

The output of the encoder f, also referred to *embedding*, is denoted as $\boldsymbol{u}_{i,j,k} = f(\boldsymbol{x}_{i,j,k}) \in \mathbb{R}^d$, where d is the embedding dimension. To streamline the notation, we define several sets of embedding vectors:

- *Same-instance embedding set*:
  This is the set of embeddings for the j-th instance in i-th class, denoted by $\boldsymbol{U}_{i,j} = \{\boldsymbol{u}_{i,j,k}\}_{k \in [p]}$ for all $i \in [m]$ and $j \in [n]$.

- *Same-class embedding set*:
  This set contains the embeddings of all instances in the i-th class, denoted by $\boldsymbol{U}_i = \cup_{j \in [n]} \boldsymbol{U}_{i,j}$ for all $i \in [m]$.

- *Entire embedding set*:
  This set includes the embeddings of all instances, represented as $\boldsymbol{U} = \cup_{i \in [m]} \boldsymbol{U}_i$.

Note that the number of embeddings in each set is $|\boldsymbol{U}| = mnp$ and $|\boldsymbol{U}_i| = np$ for all $i \in [m]$. Throughout the paper, we assume that the encoder is normalized, *i.e.,* $\|f(\boldsymbol{x})\|_2 = 1$ for all input $\boldsymbol{x}$, which is widely used in related works (Wang et al., 2017; Wu et al., 2018; Tian et al., 2020; Wang and Isola, 2020; Zimmermann et al., 2021; Sreenivasan et al., 2023; Lee et al., 2024).

The encoder is trained by optimizing the SupCL loss denoted as

$$\mathcal{L}(\boldsymbol{U}) := (1 - \alpha)\,\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U}) + \alpha\,\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U}), \quad (1)$$

where $\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U})$ is the supervised contrastive loss that considers the supervision (class information of each instance), $\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U})$ is the self-supervised contrastive loss that does not make use of the class information, and

$\alpha \in [0, 1]$ is the coefficient for combining two losses. Here, each loss term is defined as

$$\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U}) = -\frac{1}{mn(n-1)p^2} \quad (2)$$

$$\sum_{\substack{i \in [m] \\ j \neq j' \in [n]}} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_{i,j'}} \log \frac{\exp(\boldsymbol{u}^\top \boldsymbol{v}/\tau)}{\sum_{\boldsymbol{w} \in \boldsymbol{U}} \exp(\boldsymbol{u}^\top \boldsymbol{w}/\tau)}$$

and

$$\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U}) = -\frac{1}{mnp^2} \quad (3)$$

$$\sum_{\substack{i \in [m] \\ j \in [n]}} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_{i,j}} \log \frac{\exp(\boldsymbol{u}^\top \boldsymbol{v}/\tau)}{\sum_{\boldsymbol{w} \in \boldsymbol{U}} \exp(\boldsymbol{u}^\top \boldsymbol{w}/\tau)},$$

where $j \neq j' \in [n]$ is the simplified notation representing $j \in [n]$ and $j' \in [n] \setminus \{j\}$. Note that $\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U})$ in (2) is slightly different from what was proposed in the original supervised contrastive learning paper (Khosla et al., 2020) which does not include the condition $j \neq j'$ in the first summation of (2). We add this condition to make sure that the positive pairs ($\boldsymbol{u}$ and $\boldsymbol{v}$) counted in $\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U})$ and $\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U})$ do not overlap; the augmented entities $\boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{U}_{i,j}$ of the same instance are counted in $\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U})$, while the augmented entities ($\boldsymbol{u} \in \boldsymbol{U}_{i,j}$ and $\boldsymbol{v} \in \boldsymbol{U}_{i,j'}$) of different instances in the same class are counted in $\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U})$.

To make our key findings easier to understand, we first consider the simplest case where $p = 1$, *i.e.,* each instance has only one augmentation. In such a case, the index k for the augmentation (in the embedding vector $\boldsymbol{u}_{i,j,k}$) disappears, and the loss terms in (2) and (3) reduce to

$$\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U}) = -\frac{1}{mn(n-1)} \sum_{\substack{i \in [m] \\ j \neq j' \in [n]}} \log \frac{\exp(\boldsymbol{u}_{i,j}^\top \boldsymbol{u}_{i,j'}/\tau)}{\sum_{\boldsymbol{w} \in \boldsymbol{U}} \exp(\boldsymbol{u}_{i,j}^\top \boldsymbol{w}/\tau)}$$

$$(4)$$

and

$$\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U}) = -\frac{1}{mn} \sum_{\substack{i \in [m] \\ j \in [n]}} \log \frac{\exp(\boldsymbol{u}_{i,j}^\top \boldsymbol{u}_{i,j}/\tau)}{\sum_{\boldsymbol{w} \in \boldsymbol{U}} \exp(\boldsymbol{u}_{i,j}^\top \boldsymbol{w}/\tau)}. \quad (5)$$

Under the above setting, we focus on understanding the optimal embedding set

$$\boldsymbol{U}^\star := \arg\min_{\boldsymbol{U}} \mathcal{L}(\boldsymbol{U}) \quad (6)$$

that minimizes the SupCL loss in (1). In Sec. 4 and Sec. 5, we provide our theoretical results on the optimal embedding set $\boldsymbol{U}^\star$ when $p = 1$. These results are extended to general $p > 1$ case in Appendix A.

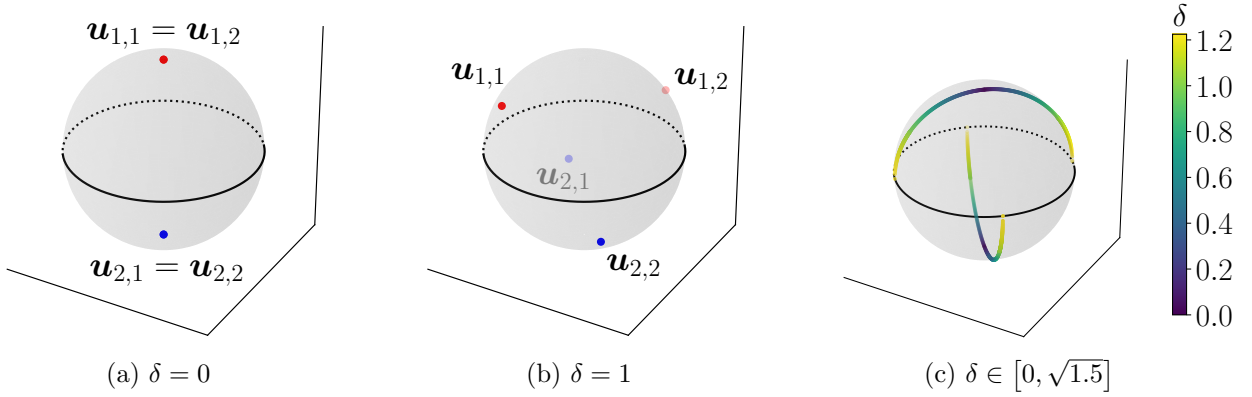|		
(a) $\delta = 0$	(b) $\delta = 1$	(c) $\delta \in \left[0, \sqrt{1.5}\right]$

Figure 1: Illustration of the proposed Simplex-to-Simplex Embedding Model (SSEM) in Def. 2, where both the number of classes (m) and the number of instances per class (n) are set to 2. The set of embedding vectors in SSEM is denoted by $\boldsymbol{U} = \{\boldsymbol{u}_{1,1}, \boldsymbol{u}_{1,2}, \boldsymbol{u}_{2,1}, \boldsymbol{u}_{2,2}\}$, where the superscript δ in (7) is omitted for simplicity. Each embedding's first subscript index indicates its class, with embeddings of class 1 drawn in red and those of class 2 in blue. The embeddings are visualized for different values of δ: (a) When $\delta = 0$, SSEM is equal to 1-simplex ETF, which is when class collapse happens. (b) When $\delta = 1$, SSEM is equal to 3-simplex ETF, where every embedding is equidistant. (c) When δ varies in the range of $\left[0, \sqrt{1.5}\right]$, we visualize the trajectory of $\boldsymbol{u}_{1,1}$ and $\boldsymbol{u}_{1,2}$ in the upper arc, and the trajectory of $\boldsymbol{u}_{2,1}$ and $\boldsymbol{u}_{2,2}$ in the lower arc, where the color in the trajectory transits from purple to yellow as δ increases.

4 OPTIMAL EMBEDDING

In this section, we first define Simplex-to-Simplex Embedding Model (SSEM), a framework of embedding sets that models different types of geometric embedding vectors. Then, we show that the optimal embedding set $\boldsymbol{U}^\star$ that minimizes the SupCL loss is only included in SSEM; note that this result is helpful for analyzing the properties of optimal embeddings in the following sections.

4.1 Simplex-to-Simplex Embedding Model

Before defining our proposed SSEM, we recall an embedding set called simplex equi-angular tight frame (ETF), where each vector is equally spaced from every other vector:

Definition 1 (Simplex ETF). *A set of n vectors $\boldsymbol{U}$ on the d-dimensional unit sphere is called $(n-1)$-simplex ETF, if*

$$\|\boldsymbol{u}\|_2^2 = 1 \text{ and } \boldsymbol{u}^\top \boldsymbol{v} = -\frac{1}{n-1}, \quad \forall \boldsymbol{u} \in \boldsymbol{U}, \boldsymbol{v} \in \boldsymbol{U} \setminus \{\boldsymbol{u}\}.$$

Note that $(n-1)$-simplex ETF exists when $d \geq n-1$.

Recall that our goal is to find the optimal embedding set $\boldsymbol{U}^\star$ that minimizes the SupCL loss

$$\mathcal{L}(\boldsymbol{U}) = (1 - \alpha)\, \mathcal{L}_{\text{Sup}}(\boldsymbol{U}) + \alpha\, \mathcal{L}_{\text{Self}}(\boldsymbol{U})$$

in (1), which is a convex combination of $\mathcal{L}_{\text{Self}}(\boldsymbol{U})$ and $\mathcal{L}_{\text{Sup}}(\boldsymbol{U})$. According to recent works, the opti-

mal embedding set that minimizes $\mathcal{L}_{\text{Self}}(\boldsymbol{U})$ follows the $(mn - 1)$-simplex ETF (Lu and Steinerberger, 2022; Lee et al., 2024), while the optimal embedding set that minimizes $\mathcal{L}_{\text{Sup}}(\boldsymbol{U})$ follows the $(m - 1)$-simplex ETF (Graf et al., 2021). This means that we already know the solution $\boldsymbol{U}^\star$ when $\alpha = 0$ or $\alpha = 1$, but not for the case of $0 < \alpha < 1$. To find the solution $\boldsymbol{U}^\star$ for all α values, we propose SSEM, a framework that models the *transition* from $(nm - 1)$-simplex ETF to $(m-1)$-simplex ETF. To be specific, SSEM uses a single parameter δ that effectively controls the shift between two embedding sets, *i.e.*, $(nm-1)$-simplex ETF and $(m-1)$-simplex ETF. Note that SSEM stands for "Simplex-to-Simplex Embedding Model", due to its ability to explain the transition from one simplex ETF to another simplex ETF by changing δ. We formally define SSEM as below:

Definition 2 (Simplex-to-Simplex Embedding Model). *Let positive integers m, n, and a real value $\delta \in \left[0, \sqrt{\frac{mn-1}{m(n-1)}}\right]$ be given. We define Simplex-to-Simplex Embedding Model, denoted by (m, n, δ)-SSEM, as the set of mn vectors*

$$\boldsymbol{U}^\delta = \left\{\boldsymbol{u}_{i,j}^\delta\right\}_{i\in[m], j\in[n]} \tag{7}$$

satisfying the following:

For all $i \neq i' \in [m]$ and $j \neq j' \in [n]$,

$$\|\boldsymbol{u}^\delta\|_2^2 = 1 \qquad\qquad \forall \boldsymbol{u}^\delta \in \boldsymbol{U}_{i,j}^\delta, \tag{8}$$

$$(\boldsymbol{u}^\delta)^\top \boldsymbol{v}^\delta = 1 - \delta^2 \frac{mn}{mn-1} \quad \forall \boldsymbol{u}^\delta \in \boldsymbol{U}_{i,j}^\delta, \boldsymbol{v}^\delta \in \boldsymbol{U}_{i,j'}^\delta, \quad (9)$$

$$(\boldsymbol{u}^\delta)^\top \boldsymbol{v}^\delta = -\frac{1}{m-1} + \delta^2 \frac{m(n-1)}{(m-1)(mn-1)}$$
$$\forall \boldsymbol{u}^\delta \in \boldsymbol{U}_i^\delta, \boldsymbol{v}^\delta \in \boldsymbol{U}_{i'}^\delta, \quad (10)$$

where $\boldsymbol{U}_{i,j}^\delta := \{\boldsymbol{u}_{i,j}^\delta\}$ and $\boldsymbol{U}_i^\delta := \cup_{j \in [n]} \boldsymbol{U}_{i,j}^\delta$.

Let $\boldsymbol{u}^\delta$ and $\boldsymbol{v}^\delta$ be two distinct embedding vectors chosen from the set of mn vectors forming (m, n, δ)-SSEM in Def. 2. If $\boldsymbol{u}^\delta$ and $\boldsymbol{v}^\delta$ are from different instances (j, j') of the same class i, their cosine similarity decreases as the parameter δ increases, as shown in (9). On the other hand, if $\boldsymbol{u}^\delta$ and $\boldsymbol{v}^\delta$ are from instances of different classes i, i', their cosine similarity increases as δ gets larger, as stated in (10). Every embedding in (m, n, δ)-SSEM has unit norm, as described in (8). As an example, Fig. 1 visualizes the $(m, n, \delta)-$SSEM for $m = n = 2$ and various δ values, where the embedding dimension is set to $d = 3$.

The below proposition shows the existence of SSEM when the embedding dimension is sufficiently large, the proof of which is in Appendix A.2.

Proposition 1 (Existence of SSEM). *Suppose $mn \geq 2$ and $d \geq mn - 1$ hold. Let a set of mn vectors $\{\boldsymbol{w}_{i,j}\}_{i \in [m], j \in [n]}$ forms the $(mn - 1)$-simplex ETF in $\mathbb{R}^d$. For a given $\delta \in \left[0, \sqrt{\frac{mn-1}{m(n-1)}}\right]$, define the set of mn vectors $\boldsymbol{U}^\delta := \left\{\boldsymbol{u}_{i,j}^\delta\right\}_{i \in [m], j \in [n]}$ as*

$$\boldsymbol{u}_{i,j}^\delta := \delta \boldsymbol{w}_{i,j} + h(\delta) \sum_{j' \in [n]} \boldsymbol{w}_{i,j'} \in \mathbb{R}^d \quad \forall i \in [m], j \in [n],$$

where

$$h(\delta) := -\frac{\delta}{n} \pm \frac{1}{n} \sqrt{\frac{\delta^2 m(1-n) + (mn-1)}{m-1}}.$$

Then, the set of mn vectors $\boldsymbol{U}^\delta$ constructs (m, n, δ)-SSEM.

The dimensionality assumption $d \geq mn - 1$ for SSEM in Proposition 1 is necessary because the simplex ETF defined in Def. 1 exists only when the dimension is sufficiently large. Investigating the existence of SSEM for $d < mn - 1$ is closely related to the Thomson problem (Thomson, 1904), which has been extensively studied in prior work (Sustik et al., 2007; Fickus et al., 2012; Fickus and Mixon, 2015; Azarija and Marc, 2018). This challenge aligns with a common assumption in theoretical studies, where embeddings learned through supervised learning (Graf et al., 2021) and self-supervised learning (Lu and Steinerberger, 2022; Lee et al., 2024) are typically analyzed under the condition that the dimension is sufficiently large. While our mathematical results in the following sections hold

strictly for $d \geq mn - 1$, our experimental findings in Sec. 6 indicate that the properties of the optimal embedding discussed in Sec. 5 remain valid even when $d < mn - 1$.

4.2 Optimal Embeddings are in SSEM

Now we show that the optimal embedding set $\boldsymbol{U}^\star$ in (6) that minimizes the SupCL loss is only included in SSEM, the proof of which is in Appendix A.3.

Theorem 1 (Optimality of SSEM). *Suppose $mn \geq 2$ and $d \geq mn - 1$ hold. Then, all embedding sets $\boldsymbol{U}^\star$ that minimize the loss $\mathcal{L}(\boldsymbol{U})$ in (1) are included in the SSEM in Def. 2, i.e.,*

$$\forall \boldsymbol{U}^\star \in \arg\min_{\boldsymbol{U}} \mathcal{L}(\boldsymbol{U}), \exists! \delta \in [0, 1] \text{ such that } \boldsymbol{U}^\delta = \boldsymbol{U}^\star. \quad (11)$$

Theorem 1 implies that we can identify the optimal embedding set as follows: we first represent the SupCL loss $\mathcal{L}(\boldsymbol{U})$ in terms of δ (by substituting the inner products $\boldsymbol{u}^T \boldsymbol{v}$ and $\boldsymbol{u}^T \boldsymbol{w}$ in (2) and (3) with the right-hand sides of (8), (9), and (10)), and then find the optimal $\delta^\star \in [0, 1]$ that minimizes the loss.

Note that the range of the optimal $\delta^\star$ in Theorem 1 is $[0, 1]$, which is a subset of $\left[0, \sqrt{\frac{mn-1}{m(n-1)}}\right]$ in Def. 2. The following proposition and remark provide some implications obtained from the distinct ranges of δ.

Proposition 2. *Let $\boldsymbol{U}^\delta$ be a set of embedding vectors forming SSEM, as defined in Def. 2, where m and n can be arbitrarily chosen. For all $i \neq i' \in [m]$, $\boldsymbol{u}^\delta, \boldsymbol{v}^\delta \in \boldsymbol{U}_i^\delta$ and $\boldsymbol{w}^\delta \in \boldsymbol{U}_{i'}^\delta$,*

$$(\boldsymbol{u}^\delta)^\top \boldsymbol{v}^\delta \geq (\boldsymbol{u}^\delta)^\top \boldsymbol{w}^\delta$$

holds, if and only if $\delta \in [0, 1]$.

Proposition 2 implies that for the optimal embedding set that minimizes the SupCL loss, the embeddings of the instances from the same class are always closer to each other, compared with the embeddings of the instances from different classes, which is desired. The proof of this proposition is in Appendix A.2.

Remark 1. *Chen et al. (2022) propose a class-conditional version of the InfoNCE loss $\mathcal{L}_{\text{cNCE}}(\boldsymbol{U})$, where negative pairs are restricted within each class, and combine it with $\mathcal{L}_{\text{Sup}}(\boldsymbol{U})$ in (2), thus defining the loss as $(1 - \alpha) \mathcal{L}_{\text{Sup}}(\boldsymbol{U}) + \alpha \mathcal{L}_{\text{cNCE}}(\boldsymbol{U})$ for some $\alpha \in [0, 1]$. In Appendix A.5, we show that SSEM with $\delta = \sqrt{\frac{mn-1}{m(n-1)}}$ (> 1) is one of optimal embedding sets that minimize $\mathcal{L}_{\text{cNCE}}(\boldsymbol{U})$, implying that the embedding vectors minimizing $\mathcal{L}_{\text{cNCE}}(\boldsymbol{U})$ satisfy $(\boldsymbol{u}^\delta)^\top \boldsymbol{v}^\delta < (\boldsymbol{u}^\delta)^\top \boldsymbol{w}^\delta$ for all $i \neq i' \in [m]$, $\boldsymbol{u}^\delta, \boldsymbol{v}^\delta \in \boldsymbol{U}_i^\delta$*

and $\boldsymbol{w}^\delta \in \boldsymbol{U}_{i'}^\delta$ from Proposition 2. In other words, training with the loss proposed by Chen et al. (2022) may incur undesired result, where embeddings of instances in different classes are closer to each other, compared with embeddings of instances in the same class. Therefore, we recommend relying on the traditional self-supervised contrastive loss $\mathcal{L}_{\text{Self}}(\boldsymbol{U})$ as in (1) instead of $\mathcal{L}_{\text{cNCE}}(\boldsymbol{U})$, to avoid such issue.

5 PREVENTING EMBEDDINGS FROM CLASS COLLAPSE

In this section, we investigate the optimal embedding sets from the perspective of variance (Fisher, 1936; Rao, 1948) and identify the conditions on the training settings in order to prevent the *class collapse* where all embeddings of the instances in the same class collapses to a single vector.

5.1 Variance of Embeddings

For a given embedding set, we define two types of variances, within-class variance and between-class variance:

Definition 3 (Variance of Embeddings). *Let $\boldsymbol{U}$ be a set of embedding vectors for mn instances, and $\boldsymbol{U}_i$ be the subset of $\boldsymbol{U}$ corresponding to the embeddings for instances in i-th class, as in Sec. 3. For all $i \in [m]$, the i-th within-class variance of $\boldsymbol{U}$ is defined as*

$$\text{Var}[\boldsymbol{U}_i] := \frac{1}{n} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \|\boldsymbol{u} - \mathbb{E}[\boldsymbol{U}_i]\|_2^2, \qquad (12)$$

where $\mathbb{E}[\boldsymbol{U}_i] := \dfrac{1}{n} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \boldsymbol{u}$ is the expectation of the embedding vectors in $\boldsymbol{U}_i$. We also define the between-class variance of $\boldsymbol{U}$ as

$$\text{Var}^{\text{Btwn}}[\boldsymbol{U}] := \frac{1}{m} \sum_{i \in [m]} \|\mathbb{E}[\boldsymbol{U}_i] - \mathbb{E}[\boldsymbol{U}]\|_2^2. \qquad (13)$$

The variances in Def. 3 are important metrics that capture the behavior of embedding vectors. When the i-th within-class variance $\text{Var}[\boldsymbol{U}_i]$ in (12) is zero, we cannot distinguish the instances in the same class, which is known as the *class collapse*. When the between-class variance $\text{Var}^{\text{Btwn}}[\boldsymbol{U}]$ in (13) is zero, we cannot separate different classes. Therefore, we want both metrics to be large enough so that different embedding vectors are separable. However, Proposition 3 shows that the sum of these variances, which is also known as the total variance, is bounded, the proof of which is in Appendix A.1.

Proposition 3 (Bounded Variance). *Let a set of embedding vectors $\boldsymbol{U}$ in Sec. 3 lie on the d-dimensional*

unit sphere, i.e., $\|\boldsymbol{u}\|_2^2 = 1$ for all $\boldsymbol{u} \in \boldsymbol{U}$. Then, the sum of all within-class variances and the between-class variance is bounded as

$$\frac{1}{m} \sum_{i \in [m]} \text{Var}[\boldsymbol{U}_i] + \text{Var}^{\text{Btwn}}[\boldsymbol{U}] \le 1. \qquad (14)$$

Here, the maximum is achieved when the centroid of the embedding vectors is at the origin, i.e., $\mathbb{E}[\boldsymbol{U}] = \boldsymbol{0}$.

We found that the embeddings forming SSEM achieves the upper bound in (14), as formally stated below. The proof of this proposition is in Appendix A.1.

Proposition 4 (Variance of SSEM). *Let $\boldsymbol{U}^\delta$ be the embedding set forming (m, n, δ)-SSEM in Def. 2. For any $\delta \in \left[0, \sqrt{\frac{mn-1}{m(n-1)}}\right]$,*

$$\text{Var}\big[\boldsymbol{U}_i^\delta\big] = \delta^2 \frac{m(n-1)}{mn-1} \qquad \forall i \in [m],$$

$$\text{Var}^{\text{Btwn}}\big[\boldsymbol{U}^\delta\big] = 1 - \delta^2 \frac{m(n-1)}{mn-1}$$

hold. Therefore,

$$\frac{1}{m} \sum_{i \in [m]} \text{Var}\big[\boldsymbol{U}_i^\delta\big] + \text{Var}^{\text{Btwn}}\big[\boldsymbol{U}^\delta\big] = 1.$$

Remark 2. *The left-hand-side of (14) is the sum of the within-class variance and the between-class variance of embeddings $\boldsymbol{U}$. The within-class variance reflects the diversity of embeddings within each class, while the between-class variance corresponds to the separability between different classes.*

Recall that SupCL has two objectives: (i) increasing the separation between instances from different classes and (ii) encouraging diversity among embeddings of instances within the same class. In this context, maximizing both variances is desirable, meaning the total variance should reach its maximum value. This maximum is attained when equality holds in (14).

Notably, by combining Theorem 1 and Proposition 4, we conclude that the optimal embedding set $\boldsymbol{U}^\star$ in (6), which minimizes the SupCL loss, always achieves the maximum variance in (14).

5.2 Preventing Class Collapse

In this section, we discuss how to prevent the optimal embedding set (that minimizes the loss $\mathcal{L}(\boldsymbol{U})$ in (1)) from the class collapse, by using the theoretical results provided in Sec. 4 and Sec. 5.1. Recall that as shown in Proposition 4 and Remark 2, the optimal embedding set $\boldsymbol{U}^\star$ minimizing the SupCL loss has the maximum variance (which is desired), and the balance between

the within-class variance $\mathrm{Var}[\boldsymbol{U}_i^\delta]$ and the between-class variance $\mathrm{Var}^{\mathrm{Btwn}}[\boldsymbol{U}^\delta]$ is controlled by the parameter δ; one can easily check that the class-collapse occurs when $\delta = 0$. The below theorem provides the conditions on the loss-combining coefficient α and the temperature τ, in order to avoid the class collapse; see Appendix A.4 for the proof.

Theorem 2 (Preventing Class Collapse). *Let $\boldsymbol{U}^\star$ be the set of optimal embedding vectors that minimizes the loss $\mathcal{L}(\boldsymbol{U})$ in* (1). *Then, the class collapse does not happen, i.e., $\mathrm{Var}[\boldsymbol{U}_i^\star] > 0$ for all $i \in [m]$, if and only if the loss-combining coefficient α satisfies*

$$\alpha \in \left(\frac{mn - 1 + \exp\left(\frac{m}{m-1}/\tau\right)}{mn - n + n \cdot \exp\left(\frac{m}{m-1}/\tau\right)}, 1 \right] \quad (15)$$

for a given temperature $\tau > 0$. This necessary and sufficient condition for preventing class collapse can be re-written as

$$\tau \in \left(0, \frac{1}{\left(1 - \frac{1}{m}\right) \cdot \log\left(\frac{mn - 1 - \alpha(m-1)n}{\alpha n - 1}\right)} \right) \quad (16)$$

for a given $\alpha \in \left(\frac{1}{n}, 1\right]$.

The above theorem provides practical guidance for selecting appropriate hyperparameters (α and τ) to guarantee that the embeddings trained with the loss do not suffer from class collapse. According to (15), the minimum value of α that prevents class collapse increases monotonically with respect to $\tau > 0$. For example, when $\tau = 0.5$ and $m = 10$, with sufficiently large n, class collapse is prevented if $\alpha \in [0.549, 1]$. On the other hand, for $\tau = 0.9$ and $m = 10$ with sufficiently large n, class collapse is avoided when using $\alpha \in [0.804, 1]$. This relationship is also illustrated in the red line in the top plot of Fig. 2. Consequently, when a smaller τ is used, a wider range of α avoids class collapse. Furthermore, the minimum α required to prevent class collapse converges to $\frac{1}{n}$ as τ goes to zero, indicating that α must always be greater than $\frac{1}{n}$.

How about the condition on τ for a given $\alpha \in (\frac{1}{n}, 1]$? According to (16), the maximum temperature parameter τ (to avoid class collapse) converges to $\left((1 - \frac{1}{m}) \cdot \log(1 + \frac{1-\alpha}{\alpha}m)\right)^{-1}$, in the asymptotic regime of large n. In the standard setting of $\alpha = 0.5$, this condition reduces to $\tau \lesssim \frac{1}{\log m}$. This suggests that the convention $\tau = 0.1$ (*e.g.,* Khosla et al. (2020)) is a reasonable choice, unless the number of classes is extremely large, *i.e.,* $m \geq \exp(10)$.

All in all, the results in Theorem 2 imply that to guarantee the learned embeddings do not suffer from class collapse, it is necessary to satisfy $\alpha \geq \frac{1}{n}$ and $\tau \leq \frac{1}{\log m}$.
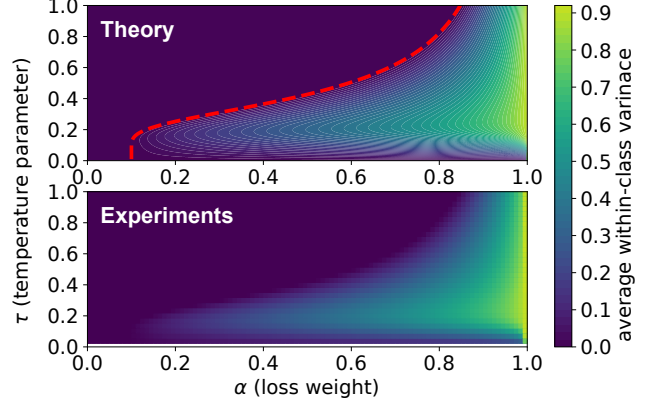

Figure 2: The within-class variance (averaged over different classes) of the learned embedding set ($\frac{1}{m}\sum_{i \in [m]} \mathrm{Var}[\boldsymbol{U}_i]$ in (12)), for various loss-combining coefficient α and temperature τ. (Top): Derived from theoretical results in Sec. 5, (Bottom): Computed from the experiments on synthetic datasets in Sec. 6.1. One can confirm that both results (shown at the top and the bottom figures) are well aligned. Here, the red dashed line at the top figure indicates the boundary of regions having zero within-class variance, *i.e.,* when class collapse happens.

6 EXPERIMENTS

In this section, we provide experimental results showing that our theoretical analysis on the within-class variance of learned embeddings (given in Sec. 4 and Sec. 5) provides practical guidelines on configurations of supervised contrastive learning to avoid class collapse. We run experiments on synthetic datasets and real image datasets, where the source code is given in https://github.com/leechungpa/ssem-supcl.

6.1 Experiments on Synthetic Data

Setup. Following previous works that reported experimental results on CL using synthetic datasets (Sreenivasan et al., 2023; Lee et al., 2024), we consider the scenario of *directly* optimizing the embedding set $\boldsymbol{U}$, instead of training an encoder that maps from each data to its embedding vector. Initially, all embedding vectors in $\boldsymbol{U}$ are randomly sampled from the 100-dimensional standard Gaussian distribution and then normalized to ensure unit norm. We then optimize these vectors to minimize the SupCL loss $\mathcal{L}(\boldsymbol{U})$ in (1) over $1,000$ epochs using the Adam optimizer, where the learning rate is set to 0.5. Note that at each training step, the updated embeddings are projected back onto the unit sphere to ensure that they maintain unit norm. Here, we have $|\boldsymbol{U}| = mnp = 200$ embeddings, which are categorized
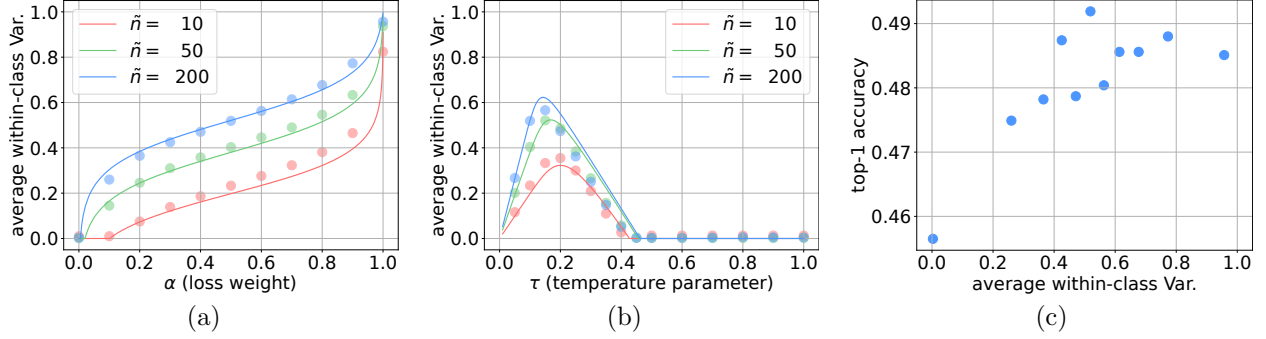
Figure 3: Average within-class variance of the learned embeddings obtained in theory (lines) and by experiments (dots), measured on CIFAR-10 dataset when ResNet-18 encoder is used. (a), (b): Dependency of the within-class variance on α and τ, for various per-class batch sizes $\tilde{n} = 10, 50, 200$. The values obtained in experiments match with those computed from our theoretical results. (c): Relationship between the within-class variance of the learned embeddings and the transfer learning performance (when transferred to CIFAR-100) of the CIFAR-10 trained ResNet-18 encoder. We set $\tau = 0.1$ and $\tilde{n} = 200$, and run experiments on various $\alpha = 0.0, 0.1, \cdots, 1.0$. Note that embeddings having a moderate amount of within-class variances achieves the highest performance.

as $m = 10$ classes, $n = 10$ instances per class, and $p = 2$ augmentations per instance.

Within-Class Variance of Learned Embeddings. Fig. 2 shows the average within-class variance $(\frac{1}{m} \sum_{i \in [m]} \text{Var}[\boldsymbol{U}_i]$ in (12)) of the learned embedding set that minimizes the SupCL loss in (1), for various α and τ. The top figure presents the average within-class variance derived from theoretical results (based on Theorem 1 and Proposition 4), whereas the bottom figure shows the average within-class variance computed from embeddings trained on our synthetic datasets. At each (α, τ) pair, the average within-class variance obtained in theory well matches with those computed in experiments. These results show that our analysis on the variance of learned embeddings in Sec. 5 is valid for synthetic datasets.

6.2 Experiments on Real Data

Setup. We run experiments on two image datasets: CIFAR-10 (Krizhevsky et al., 2009) and ImageNet-100 (Deng et al., 2009). We use ResNet architecture (He et al., 2016) for the encoder, and 2-layer MLP for the projector, where the output dimension of the projector is set to $d = 128$.

After training embeddings using the SupCL loss $\mathcal{L}(\boldsymbol{U})$ in (1), we measure the average within-class variance $(\frac{1}{m} \sum_{i \in [m]} \text{Var}[\boldsymbol{U}_i]$ in (12)) of the embeddings (normalized output of the projector) of the train data, without any augmentation. We then remove the projector head and evaluate the performance of pretrained encoder on various downstream classification tasks by using linear probing. Following recent works (Korn-

blith et al., 2019; Lee et al., 2021; Oh and Lee, 2024), we evaluate the top-1 accuracy or the mean per class accuracy, depending on the downstream tasks. Details of the datasets, as well as the training and evaluation processes, are provided in Appendix C.

Within-Class Variance of Learned Embeddings. Fig. 3 shows the average within-class variance measured for CIFAR-10 dataset. To be specific, Fig. 3a shows the dependency on the loss-combining coefficient α, when temperature is set to $\tau = 0.1$, while Fig. 3b shows the dependency on τ, when $\alpha = 0.5$.

We test on three different batch sizes: the per-class batch size (*i.e.,* the number of instances from the same class and contained in the same batch) is set to $\tilde{n} = 10, 50, 200$. Here, we use balanced batches, where the number of instances per class is equal in every batch. Since CIFAR-10 datasets are categorized into $m = 10$ classes, we have batch sizes of $m\tilde{n} = 100, 500, 2000$, respectively.

In Fig.3a and Fig.3b, we compare two types of within-class variances: one computed based on the theoretical results in Sec.5 (shown as solid lines) and the other obtained from experiments (shown as dots). Note that in order to reflect the batch effect, we use $\tilde{n}$ instead of n in the expressions given in Sec.5 when plotting the solid line representing the theoretical results.

One can observe that for each $\tilde{n}$, the solid line (derived from theory) aligns with the dots (obtained from experiments), confirming that our analysis on the variance of learned embeddings in Sec. 5 is valid for real datasets.

Table 1: Transfer learning performance evaluated on various downstream tasks. We first train a ResNet-50 encoder on ImageNet-100 by minimizing the SupCL loss $\mathcal{L}(\mathbf{U})$ in (1) with $\tau = 0.1$ for various α values. Then, we evaluate the pretrained encoder on multiple downstream classification tasks using linear probing, measuring the performance via top-1 accuracy or mean per-class accuracy. Here, for evaluating the second column of the table, we compute the within-class variance of each class using the embeddings of the original training data (without any augmentations), and take average of the within-class variances. The best result for each downstream dataset is highlighted in bold. Notably, the model with a moderate amount of average within-class variance shows the highest average performance.

α	Avg. within-class var.	Avg. accuracy	CIFAR10	CIFAR100	Caltech101	CUB200	Dog	DTD	Flowers102	Food101	MIT67	Pets	SUN397
0.0	0.014	68.00	88.34	68.25	**88.87**	36.05	62.04	65.64	88.46	57.98	63.13	79.05	50.19
0.2	0.021	68.16	88.70	68.96	87.35	35.45	62.10	65.11	88.46	59.26	64.40	80.09	49.89
0.4	0.078	68.42	88.91	69.08	87.60	37.35	62.39	66.33	88.19	59.78	63.28	79.56	50.20
0.5	0.133	**69.06**	**89.43**	69.45	88.35	**38.48**	**62.78**	66.33	**89.49**	60.36	63.43	**80.68**	50.88
0.6	0.192	68.72	89.34	**70.07**	87.50	35.99	61.56	66.33	87.82	61.03	**65.45**	79.75	**51.13**
0.8	0.325	68.36	88.13	68.42	86.06	36.46	60.73	**66.49**	89.29	**62.72**	64.55	78.48	50.60
1.0	0.830	63.06	84.15	63.17	78.64	30.40	46.19	65.00	85.71	62.02	62.91	67.60	47.89

Relationship between Within-Class Variance and Transfer Learning Performance. Now we provide experimental results showing that the within-class variance of the learned embeddings is a good indicator of the transfer learning performance. Fig. 3c shows the relationship between the within-class variance and the transfer learning performance, when the embeddings are used to classify CIFAR-100 datasets. Here, we set $\tau = 0.1$ and $\tilde{n} = 200$, and evaluate on various α values ranging from 0 to 1.

The result in Fig. 3c indicates that embeddings suffering from class collapse, which have zero within-class variance, exhibit the lowest top-1 accuracy on CIFAR-100 classification. In contrast, embeddings with a moderate amount of within-class variance (roughly between 0.4 and 0.8) achieve the highest top-1 accuracy on CIFAR-100 classification.

This trend, where the best performance occurs when embeddings maintain a moderate amount of within-class variance, is also observed in more complex architectures and datasets. In Table 1, we report the relationship between the transfer learning performance and the within-class variance for the embeddings learned on ImageNet-100 dataset. Specifically, we train a ResNet-50 encoder on ImageNet-100 using the SupCL loss $\mathcal{L}(\mathbf{U})$ in (1) with $\tau = 0.1$ for various α values, resulting in each trained encoder exhibiting a different within-class variance. We then evaluate its transfer learning performance across different downstream datasets, as well as the within-class variance of the trained model. In addition, we run experiments on other transfer learning tasks including object detection and few-shot learning tasks, results of which are provided in Appendix C.4. Across all these tasks, the findings consistently demonstrate that embeddings with a moderate amount of within-class variance yield the best performance.

7 CONCLUSION

This paper explores the behavior of embeddings trained with supervised contrastive learning, from the perspective of variance. First, we prove that Simplex-to-Simplex Embedding Model (SSEM), a class of embedding sets defined by us, contains the optimal embedding set that minimizes the supervised contrastive loss. Then, we provide theoretical analysis on the behaviors of optimal embeddings from the perspective of within-class and between-class variances and offer guidelines on training configurations to prevent the learned embeddings from class collapse, *i.e.,* when the within-class variance is zero. This theoretical result aligns well with our experimental findings on synthetic and real datasets. Specifically, the within-class variance of the learned embeddings (computed in our experiments) matches the mathematical expressions derived in theory. Moreover, our theory-driven guideline for preventing class collapse is empirically validated, showing consistency with conventional practices.

Acknowledgements

References

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. (2019). A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5628–5637. 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.

Bossard, L., Guillaumin, M., and Van Gool, L. (2014). Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*.

Chen, M., Fu, D. Y., Narayan, A., Zhang, M., Song, Z., Fatahalian, K., and Ré, C. (2022). Perfectly balanced: Improving transfer and robustness of supervised contrastive learning. In *International Conference on Machine Learning*, pages 3090–3122. PMLR.

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

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.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. (2014). Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338.

Fei-Fei, L., Fergus, R., and Perona, P. (2004). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE.

Feng, Y., Jiang, J., Tang, M., Jin, R., and Gao, Y. (2022). Rethinking supervised pre-training for better downstream transferring. In *International Conference on Learning Representations*.

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.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188.

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.

Gunel, B., Du, J., Conneau, A., and Stoyanov, V. (2021). Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. (2021). Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011.

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.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Islam, A., Chen, C.-F. R., Panda, R., Karlinsky, L., Radke, R., and Feris, R. (2021). A broad study on the transferability of visual representations with contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8845–8855.

Khosla, A., Jayadevaprakash, N., Yao, B., and Fei-Fei, L. (2011). Novel datasets for fine-grained image categorization. In *First Workshop on Fine Grained Visual Categorization, CVPR. Citeseer. Citeseer. Citeseer*, volume 5, page 2. Citeseer.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. (2020). Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Kornblith, S., Shlens, J., and Le, Q. V. (2019). Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

Lee, C., Chang, J., and Sohn, J.-y. (2024). Analysis of using sigmoid loss for contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1747–1755. PMLR.

Lee, H., Lee, K., Lee, K., Lee, H., and Shin, J. (2021). Improving transferability of representations via augmentation-aware self-supervision. *Advances in Neural Information Processing Systems*, 34:17710–17722.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer.

Liu, D. C. and Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528.

Loshchilov, I. and Hutter, F. (2017). SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

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

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. (2013). Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Nilsback, M.-E. and Zisserman, A. (2008). Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, pages 722–729. IEEE.

Oh, J. and Lee, K. (2024). On the effectiveness of supervision in asymmetric non-contrastive learning. In *International Conference on Machine Learning*.

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

Oreshkin, B., Rodríguez López, P., and Lacoste, A. (2018). Tadam: Task dependent adaptive metric for improved few-shot learning. *Advances in neural information processing systems*, 31.

Papyan, V., Han, X., and Donoho, D. L. (2020). Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40):24652–24663.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. (2012). Cats and dogs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE.

Parulekar, A., Collins, L., Shanmugam, K., Mokhtari, A., and Shakkottai, S. (2023). Infonce loss provably learns cluster-preserving representations. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1914–1961. PMLR.

Quattoni, A. and Torralba, A. (2009). Recognizing indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 413–420. IEEE.

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.

Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

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*.

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.

Thomson, J. J. (1904). On the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 7(39):237–265.

Tian, Y., Krishnan, D., and Isola, P. (2020). Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer.

Wang, C., Zheng, W., Zhu, Z., Zhou, J., and Lu, J. (2023). Opera: omni-supervised representation learning with hierarchical supervisions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5559–5570.

Wang, F., Xiang, X., Cheng, J., and Yuille, A. L. (2017). Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049.

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.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., and Perona, P. (2011). Caltech-ucsd birds 200. Technical Report CNS-TR-2011-001, Technical Report CNS-TR-2010-001, California Institute of Technology.

Wen, Z. and Li, Y. (2021). Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3485–3492. IEEE.

Xue, Y., Joshi, S., Gan, E., Chen, P.-Y., and Mirzasoleiman, B. (2023). Which features are learnt by contrastive learning? on the role of simplicity bias in class collapse and feature suppression. In *International Conference on Machine Learning*, pages 38938–38970. PMLR.

Yang, R., Li, X., Jiang, B., and Li, S. (2023). Understanding representation learnability of nonlinear self-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10807–10815.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. (2023). Sigmoid loss for language image pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. (2021). Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] See Sec 3.

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes] Following prior works, we use a normalized encoder f, *i.e.,* $\|f(\boldsymbol{x})\|_2 = 1$ for all input $\boldsymbol{x}$, as described in Sec 3. To streamline the notation, we present our theoretical results in the main paper for the case of a single augmentation ($p = 1$). These results are extended to the general case ($p > 1$) in Appendix A.

   (b) Complete proofs of all theoretical results. [Yes] See Appendix A.

   (c) Clear explanations of any assumptions. [Yes] See Sec 3.

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] See Appendix C.

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] See Appendix C.

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] See Appendix C.

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] See Appendix C.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes] See Appendix C.

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A   PROOFS

A.1   Proposition About Embedding Vectors on the Unit Sphere

The definitions of within-class and between-class variances are reformulated as follows:

Definition A.1 (Restatement of Def. 3). *Let $\boldsymbol{U}$ be a set of embedding vectors, and $\boldsymbol{U}_i$ be the subset of $\boldsymbol{U}$ corresponding to the embeddings for instances in i-th class, as in Sec. 3. For all $i \in [m]$, the i-th within-class variance of $\boldsymbol{U}$ is defined as*

$$\mathrm{Var}[\boldsymbol{U}_i] := \frac{1}{|\boldsymbol{U}_i|} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \|\boldsymbol{u} - \mathbb{E}[\boldsymbol{U}_i]\|_2^2,$$

where $\mathbb{E}[\boldsymbol{U}_i] := \dfrac{1}{|\boldsymbol{U}_i|} \displaystyle\sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \boldsymbol{u}$ is the expectation of the embedding vectors in $\boldsymbol{U}_i$. We also define the between-class variance of $\boldsymbol{U}$ as

$$\mathrm{Var}^{\mathrm{Btwn}}[\boldsymbol{U}] := \sum_{i \in [m]} \frac{|\boldsymbol{U}_i|}{|\boldsymbol{U}|} \|\mathbb{E}[\boldsymbol{U}_i] - \mathbb{E}[\boldsymbol{U}]\|_2^2.$$

Proposition A.1. *Consider a set of n vectors $\boldsymbol{U}$ on the sphere of radius $r > 0$, i.e., $\|\boldsymbol{u}\|_2 = r$ for all $\boldsymbol{u} \in \boldsymbol{U}$. Then, the variance of set $\boldsymbol{U}$ can be rewritten as*

$$\mathrm{Var}[\boldsymbol{U}] = r^2 - \|\mathbb{E}[\boldsymbol{U}]\|_2^2 = \frac{|\boldsymbol{U}| - 1}{|\boldsymbol{U}|} \cdot r^2 - \frac{1}{|\boldsymbol{U}|^2} \sum_{\boldsymbol{u} \in \boldsymbol{U}} \sum_{\boldsymbol{v} \in \boldsymbol{U} \setminus \{\boldsymbol{u}\}} \boldsymbol{u}^\top \boldsymbol{v}.$$

Proof. According to Def. A.1, the variance of vectors on the unit sphere is determined as follows:

$$\begin{aligned}
\mathrm{Var}[\boldsymbol{U}] &= \frac{1}{|\boldsymbol{U}|} \sum_{\boldsymbol{u} \in \boldsymbol{U}} \|\boldsymbol{u} - \mathbb{E}[\boldsymbol{U}]\|_2^2 \\
&= \frac{1}{|\boldsymbol{U}|} \sum_{\boldsymbol{u} \in \boldsymbol{U}} \|\boldsymbol{u}\|_2^2 + \|\mathbb{E}[\boldsymbol{U}]\|_2^2 - \frac{2}{|\boldsymbol{U}|} \sum_{\boldsymbol{u} \in \boldsymbol{U}} \boldsymbol{u}^\top \mathbb{E}[\boldsymbol{U}] \\
&= \frac{1}{|\boldsymbol{U}|} \sum_{\boldsymbol{u} \in \boldsymbol{U}} \|\boldsymbol{u}\|_2^2 - \|\mathbb{E}[\boldsymbol{U}]\|_2^2 \\
&= r^2 - \|\mathbb{E}[\boldsymbol{U}]\|_2^2 \\
&= r^2 - \frac{1}{|\boldsymbol{U}|^2} \left\| \sum_{\boldsymbol{u} \in \boldsymbol{U}} \boldsymbol{u} \right\|_2^2 \\
&= r^2 - \frac{1}{|\boldsymbol{U}|^2} \sum_{\boldsymbol{u} \in \boldsymbol{U}} \sum_{\boldsymbol{v} \in \boldsymbol{U}} \boldsymbol{u}^\top \boldsymbol{v} \\
&= \frac{|\boldsymbol{U}| - 1}{|\boldsymbol{U}|} \cdot r^2 - \frac{1}{|\boldsymbol{U}|^2} \sum_{\boldsymbol{u} \in \boldsymbol{U}} \sum_{\boldsymbol{v} \in \boldsymbol{U} \setminus \{\boldsymbol{u}\}} \boldsymbol{u}^\top \boldsymbol{v}.
\end{aligned}$$

$\square$

Proposition A.2 (Restatement of Proposition 3). *For $i \in [m]$, let $\boldsymbol{U}_i$ be the set of vectors on the sphere with radius $r > 0$, i.e., $\|\boldsymbol{u}\|_2 = r$ for all $i \in [m]$ and $\boldsymbol{u} \in \boldsymbol{U}_i$. Define the entire set as $\boldsymbol{U} := \cup_{i \in [m]} \boldsymbol{U}_i$. Then, the sum of variances is bounded as*

$$\sum_{i \in [m]} \frac{|\boldsymbol{U}_i|}{|\boldsymbol{U}|} \mathrm{Var}[\boldsymbol{U}_i] + \mathrm{Var}^{\mathrm{Btwn}}[\boldsymbol{U}] \leq r^2.$$

The maximum is achieved when the centroid of the vectors is at the origin; that is,

$$\mathbb{E}[\boldsymbol{U}] = \boldsymbol{0}.$$

Proof. For simplicity of notation, let $n_i = |U_i|$ and $N = \sum_{i \in [m]} n_i$. Note that the sum of following inner products is zero.

$$\sum_{i \in [m]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \left(\boldsymbol{u} - \mathbb{E}[\boldsymbol{U}_i]\right)^\top \left(\mathbb{E}[\boldsymbol{U}_i] - \mathbb{E}[\boldsymbol{U}]\right) = \sum_{i \in [m]} \left(\sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \boldsymbol{u} - n_i \cdot \mathbb{E}[\boldsymbol{U}_i]\right)^\top \left(\mathbb{E}[\boldsymbol{U}_i] - \mathbb{E}[\boldsymbol{U}]\right)$$

$$= \sum_{i \in [m]} \boldsymbol{0}^\top \left(\mathbb{E}[\boldsymbol{U}_i] - \mathbb{E}[\boldsymbol{U}]\right)$$

$$= 0.$$

Then, we can decompose the summation of norms as follows.

$$\sum_{\boldsymbol{u} \in \boldsymbol{U}} \|\boldsymbol{u} - \mathbb{E}[\boldsymbol{U}]\|_2^2 = \sum_{i \in [m]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \|\boldsymbol{u} - \mathbb{E}[\boldsymbol{U}]\|_2^2$$

$$= \sum_{i \in [m]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \|\boldsymbol{u} - \mathbb{E}[\boldsymbol{U}_i] + \mathbb{E}[\boldsymbol{U}_i] - \mathbb{E}[\boldsymbol{U}]\|_2^2$$

$$= \sum_{i \in [m]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \|\boldsymbol{u} - \mathbb{E}[\boldsymbol{U}_i]\|_2^2 + \sum_{i \in [m]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \|\mathbb{E}[\boldsymbol{U}_i] - \mathbb{E}[\boldsymbol{U}]\|_2^2 + \sum_{i \in [m]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} 2\left(\boldsymbol{u} - \mathbb{E}[\boldsymbol{U}_i]\right)^\top \left(\mathbb{E}[\boldsymbol{U}_i] - \mathbb{E}[\boldsymbol{U}]\right)$$

$$= \sum_{i \in [m]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \|\boldsymbol{u} - \mathbb{E}[\boldsymbol{U}_i]\|_2^2 + \sum_{i \in [m]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \|\mathbb{E}[\boldsymbol{U}_i] - \mathbb{E}[\boldsymbol{U}]\|_2^2$$

$$= \sum_{i \in [m]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \|\boldsymbol{u} - \mathbb{E}[\boldsymbol{U}_i]\|_2^2 + \sum_{i \in [m]} n_i \|\mathbb{E}[\boldsymbol{U}_i] - \mathbb{E}[\boldsymbol{U}]\|_2^2.$$

As a result, the variance of the entire set can be rewritten as follows.

$$\text{Var}(\boldsymbol{W}) = \frac{1}{N} \sum_{\boldsymbol{u} \in \boldsymbol{U}} \|\boldsymbol{u} - \mathbb{E}[\boldsymbol{U}]\|_2^2$$

$$= \frac{1}{N} \sum_{i \in [m]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \|\boldsymbol{u} - \mathbb{E}[\boldsymbol{U}_i]\|_2^2 + \frac{1}{N} \sum_{i \in [m]} n_i \|\mathbb{E}[\boldsymbol{U}_i] - \mathbb{E}[\boldsymbol{U}]\|_2^2$$

$$= \sum_{i \in [m]} \frac{n_i}{N} \text{Var}[\boldsymbol{U}_i] + \sum_{i \in [m]} \frac{n_i}{N} \|\mathbb{E}[\boldsymbol{U}_i] - \mathbb{E}[\boldsymbol{U}]\|_2^2$$

$$= \sum_{i \in [m]} \frac{n_i}{N} \text{Var}[\boldsymbol{U}_i] + \text{Var}^{\text{Btwn}}[\boldsymbol{U}].$$

To establish the upper bound, from Proposition A.1,

$$\text{Var}[\boldsymbol{U}] = r^2 - \frac{1}{n^2} \left\|\sum_{\boldsymbol{u} \in \boldsymbol{U}} \boldsymbol{u}\right\|_2^2$$

$$\leq r^2, \tag{A.1}$$

which implies

$$\sum_{i \in [m]} \frac{n_i}{N} \text{Var}[\boldsymbol{U}_i] + \text{Var}^{\text{Btwn}}[\boldsymbol{U}] \leq r^2.$$

The equality condition in (A.1) is

$$\mathbb{E}[\boldsymbol{U}] = \frac{1}{N} \sum_{\boldsymbol{u} \in \boldsymbol{U}} \boldsymbol{u} = \boldsymbol{0}.$$

$\square$

A.2   Properties of SSEM

We redefine the the Simplex-to-Simplex Embedding Model (SSEM) for a general $p \in \mathbb{N}$ as follows:

Definition A.2 (Restatement of Def. 2). *Let positive integers m,n,p, and a real non-negative number $\delta \in \left[0, \sqrt{\frac{mn-1}{m(n-1)}}\right]$ be given. We define Simplex-to-Simplex Embedding Model, denoted by (m, n, p, δ)-SSEM, as the set of mnp vectors*

$$\boldsymbol{U}^\delta = \left\{\boldsymbol{u}_{i,j,k}^\delta\right\}_{i\in[m],j\in[n],k\in[p]}$$

satisfying the following:

For all $i \neq i' \in [m]$ and $j \neq j' \in [n]$,

$$(\boldsymbol{u}^\delta)^\top \boldsymbol{v}^\delta = 1 \qquad\qquad\qquad \forall \boldsymbol{u}^\delta, \boldsymbol{v}^\delta \in \boldsymbol{U}_{i,j}^\delta, \qquad\qquad \text{(A.2)}$$

$$(\boldsymbol{u}^\delta)^\top \boldsymbol{v}^\delta = 1 - \delta^2 \frac{mn}{mn-1} \qquad\qquad \forall \boldsymbol{u}^\delta \in \boldsymbol{U}_{i,j}^\delta, \boldsymbol{v}^\delta \in \boldsymbol{U}_{i,j'}^\delta, \qquad\qquad \text{(A.3)}$$

$$(\boldsymbol{u}^\delta)^\top \boldsymbol{v}^\delta = -\frac{1}{m-1} + \delta^2 \frac{m(n-1)}{(m-1)(mn-1)} \qquad \forall \boldsymbol{u}^\delta \in \boldsymbol{U}_i^\delta, \boldsymbol{v}^\delta \in \boldsymbol{U}_{i'}^\delta, \qquad \text{(A.4)}$$

where $\boldsymbol{U}_{i,j}^\delta := \{\boldsymbol{u}_{i,j,k}^\delta\}_{k\in[p]}$ and $\boldsymbol{U}_i^\delta := \cup_{j\in[n]}\boldsymbol{U}_{i,j}^\delta$.

Proposition A.3. *Let $\boldsymbol{U}^\delta$ be a set of embedding vectors forming SSEM, as defined in Def. A.2, where m, n and p can be arbitrarily chosen. For all $i \neq i' \in [m]$, $\boldsymbol{u}^\delta, \boldsymbol{v}^\delta \in \boldsymbol{U}_i^\delta$ and $\boldsymbol{w}^\delta \in \boldsymbol{U}_{i'}^\delta$,*

$$(\boldsymbol{u}^\delta)^\top \boldsymbol{v}^\delta \geq (\boldsymbol{u}^\delta)^\top \boldsymbol{w}^\delta \qquad\qquad\qquad\qquad \text{(A.5)}$$

holds, if and only if $\delta \in [0, 1]$.

Proof. For all $i \neq i' \in [m]$, $\boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{U}_i^\delta$ and $\boldsymbol{w} \in \boldsymbol{U}_{i'}^\delta$, the following holds from the definition of SSEM:

$$(\boldsymbol{u}^\delta)^\top \boldsymbol{v}^\delta = \begin{cases} 1 & \text{if there exist } j \in [m] \text{ such that } \boldsymbol{u}^\delta, \boldsymbol{v}^\delta \in \boldsymbol{U}_{i,j}^\delta \\ 1 - \delta^2 \frac{mn}{mn-1} & \text{otherwise} \end{cases}$$

$$\geq \min\left(1, 1 - \delta^2 \frac{mn}{mn-1}\right)$$

$$= 1 - \delta^2 \frac{mn}{mn-1},$$

$$(\boldsymbol{u}^\delta)^\top \boldsymbol{w}^\delta = -\frac{1}{m-1} + \delta^2 \frac{m(n-1)}{(m-1)(mn-1)}.$$

To determine the necessary and sufficient condition for (A.5),

$$(\boldsymbol{u}^\delta)^\top \boldsymbol{v}^\delta - (\boldsymbol{u}^\delta)^\top \boldsymbol{w}^\delta \geq 1 - \delta^2 \frac{mn}{mn-1} + \frac{1}{m-1} - \delta^2 \frac{m(n-1)}{(m-1)(mn-1)}$$

$$= \frac{m}{m-1} - \delta^2 \frac{m}{m-1}$$

$$= \frac{m}{m-1}(1 - \delta^2),$$

which implies that (A.5) holds if and only if $\delta \in [0, 1]$, as δ is defined to be within the range $\left[0, \sqrt{\frac{mn-1}{m(n-1)}}\right]$. $\qquad \square$

Proposition A.4. *Let $\boldsymbol{U}^\delta$ be the embedding set forming (m, n, p, δ)-SSEM in Def. A.2. For any $\delta \in \left[0, \sqrt{\frac{mn-1}{m(n-1)}}\right]$,*

$$\mathrm{Var}\left[\boldsymbol{U}_i^\delta\right] = \delta^2 \frac{m(n-1)}{mn-1} \qquad \forall i \in [m], \qquad and \qquad \mathrm{Var}^{\mathrm{Btwn}}\left[\boldsymbol{U}^\delta\right] = 1 - \delta^2 \frac{m(n-1)}{mn-1}$$

Therefore,

$$\frac{1}{m} \sum_{i \in [m]} \mathrm{Var}\left[\boldsymbol{U}_i^\delta\right] + \mathrm{Var}^{\mathrm{Btwn}}\left[\boldsymbol{U}^\delta\right] = 1.$$

Proof. Note that the followings hold from the definition of SSEM in Def. A.2:

For all $i \in [m]$,

$$\left\|\sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}_i^\delta} \boldsymbol{u}^\delta\right\|_2^2 = \left(\sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}_i^\delta} \boldsymbol{u}^\delta\right)^\top \left(\sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}_i^\delta} \boldsymbol{u}^\delta\right)$$

$$= np^2 + n(n-1)p^2\left(1 - \delta^2 \frac{mn}{mn-1}\right)$$

$$= n^2 p^2 - \delta^2 \frac{mn^2(n-1)p^2}{mn-1},$$

$$\left(\sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}_i^\delta} \boldsymbol{u}^\delta\right)^\top \left(\sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}^\delta} \boldsymbol{u}^\delta\right) = np^2 + n(n-1)p^2\left(1 - \delta^2 \frac{mn}{mn-1}\right)$$

$$+ (m-1)n^2 p^2\left(-\frac{1}{m-1} + \delta^2 \frac{m(n-1)}{(m-1)(mn-1)}\right)$$

$$= -\delta^2 \frac{mn^2(n-1)p^2}{mn-1} + \delta^2 \frac{mn^2(n-1)p^2}{mn-1}$$

$$= 0,$$

$$\left(\sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}^\delta} \boldsymbol{u}^\delta\right)^\top \left(\sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}^\delta} \boldsymbol{u}^\delta\right) = \sum_{i \in [m]} \left(\sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}_i^\delta} \boldsymbol{u}^\delta\right)^\top \left(\sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}^\delta} \boldsymbol{u}^\delta\right)$$

$$= 0,$$

which implies

$$\mathbb{E}\left[\boldsymbol{U}^\delta\right] = \frac{1}{mnp} \sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}^\delta} \boldsymbol{u}^\delta = \boldsymbol{0}$$

for any $\delta \in \left[0, \sqrt{\frac{mn-1}{m(n-1)}}\right]$. Then, the between-class variance of SSEM is determined as below.

$$\mathrm{Var}^{\mathrm{Btwn}}\left[\boldsymbol{U}^\delta\right] = \sum_{i \in [m]} \frac{np}{mnp} \left\|\mathbb{E}\left[\boldsymbol{U}_i^\delta\right] - \mathbb{E}\left[\boldsymbol{U}^\delta\right]\right\|_2^2$$

$$= \frac{1}{m} \sum_{i \in [m]} \left\|\frac{1}{np} \sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}_i^\delta} \boldsymbol{u}^\delta\right\|_2^2$$

$$= \sum_{i \in [m]} \frac{1}{m}\left(\frac{1}{n^2 p^2}\left(n^2 p^2 - \delta^2 \frac{mn^2(n-1)p^2}{mn-1}\right) + 0\right)$$

$$= \sum_{i \in [m]} \frac{1}{m}\left(1 - \delta^2 \frac{m(n-1)}{mn-1}\right)$$

$$= 1 - \delta^2 \frac{m(n-1)}{mn-1}.$$

Moreover, the i-th within-class variance for all $i \in [m]$ is determined as follows, based on Proposition A.1,

$$\mathrm{Var}\big[\boldsymbol{U}_i^\delta\big] = 1 - \frac{1}{n^2 p^2}\left\|\sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}_i^\delta} \boldsymbol{u}^\delta\right\|_2^2 = 1 - \frac{1}{n^2 p^2}\left(n^2 p^2 - \delta^2 \frac{mn^2(n-1)p^2}{mn-1}\right) = \delta^2 \frac{m(n-1)}{mn-1}.$$

Therefore, the following holds for any $\delta \in \left[0, \sqrt{\frac{mn-1}{m(n-1)}}\right]$:

$$\mathrm{Var}^{\mathrm{Btwn}}\big[\boldsymbol{U}^\delta\big] + \sum_{i \in [m]} \frac{1}{m}\mathrm{Var}\big[\boldsymbol{U}_i^\delta\big] = 1 - \delta^2 \frac{m(n-1)}{mn-1} + \delta^2 \frac{m(n-1)}{mn-1} = 1.$$

$\square$

Theorem A.1 (Existence of SSEM). *Suppose $mn \geq 2$ and $d \geq mn - 1$ hold. Let a set of mn vectors $\{\boldsymbol{w}_{i,j}\}_{i \in [m], j \in [n]}$ forms the $(mn-1)$-simplex ETF in $\mathbb{R}^d$. For a given $\delta \in \left[0, \sqrt{\frac{mn-1}{m(n-1)}}\right]$, define the set of mnp vectors $\boldsymbol{U}^\delta := \{\boldsymbol{u}_{i,j,k}^\delta\}_{i \in [m], j \in [n], k \in [p]}$ as*

$$\boldsymbol{u}_{i,j,k}^\delta := \delta\boldsymbol{w}_{i,j} + h(\delta)\sum_{j \in [n]} \boldsymbol{w}_{i,j} \in \mathbb{R}^d \qquad \forall i \in [m], j \in [n], k \in [p], \tag{A.6}$$

where

$$h(\delta) := -\frac{\delta}{n} \pm \frac{1}{n}\sqrt{\frac{\delta^2 m(1-n) + (mn-1)}{m-1}}.$$

Then, the set of mnp vectors $\boldsymbol{U}^\delta$ constructs SSEM.

Proof. From $d \geq mn - 1$, the $(mn-1)$-simplex ETF of Def. 1 exists in $\mathbb{R}^d$, implying that the set of mnp vectors $\boldsymbol{U}^\delta$ exist. What remains to be proved is that $\boldsymbol{U}^\delta$ follows SSEM of Def. A.2.

For simplicity, the vector $\boldsymbol{v}_i$ is defined as

$$\boldsymbol{v}_i := \sum_{j \in [n]} \boldsymbol{w}_{i,j} \in \mathbb{R}^d$$

for all $i \in [m]$. This simplifies $\boldsymbol{u}_{i,j,k}^\delta$ as

$$\boldsymbol{u}_{i,j,k}^\delta = \delta\boldsymbol{w}_{i,j} + h(\delta)\boldsymbol{v}_i \qquad \forall i \in [m], j \in [n], k \in [p].$$

From the definition of the $(mn-1)$-simplex ETF, the followings hold for all $i \neq i' \in [m]$ and $j \in [n]$:

$$\boldsymbol{v}_i^\top \boldsymbol{v}_i = n - n(n-1)\frac{1}{mn-1} = \frac{(m-1)n^2}{mn-1} \quad (> 0), \tag{A.7}$$

$$\boldsymbol{v}_i^\top \boldsymbol{v}_{i'} = -n^2\frac{1}{mn-1} = -\frac{n^2}{mn-1} \quad (< 0), \tag{A.8}$$

$$\boldsymbol{w}_{i,j}^\top \boldsymbol{v}_i = 1 - (n-1)\frac{1}{mn-1} = \frac{(m-1)n}{mn-1} \quad (> 0), \tag{A.9}$$

$$\boldsymbol{w}_{i,j}^\top \boldsymbol{v}_{i'} = -n\frac{1}{mn-1} = -\frac{n}{mn-1} \quad (< 0). \tag{A.10}$$

Using the above results in (A.7)-(A.10), we can show that $\boldsymbol{U}^\delta$ defined in (A.6) satisfies the conditions of SSEM, which are (A.2), (A.3), and (A.4) in Def. A.2, as below:

[**Condition of** (A.2)] For all $i \in [m], j \in [n], k, k' \in [p]$,

$$(\boldsymbol{u}_{i,j,k}^{\delta})^{\top} \boldsymbol{u}_{i,j,k'}^{\delta} = \delta^2 \boldsymbol{w}_{i,j}^{\top} \boldsymbol{w}_{i,j} + h(\delta)^2 \boldsymbol{v}_i^{\top} \boldsymbol{v}_i + 2\delta h(\delta) \boldsymbol{w}_{i,j}^{\top} \boldsymbol{v}_i \tag{A.11}$$

$$= \delta^2 + h(\delta)^2 \frac{(m-1)n^2}{mn-1} + 2\delta h(\delta) \frac{(m-1)n}{mn-1} \tag{A.12}$$

$$= \delta^2 + \frac{(m-1)n}{mn-1} \left(h(\delta)^2 \cdot n + 2\delta h(\delta) \right)$$

$$= \delta^2 + \frac{(m-1)n}{mn-1} \cdot \frac{mn-1}{(m-1)n} \cdot (1-\delta^2) \tag{A.13}$$

$$= 1,$$

where the second and third terms in (A.12) are results from using (A.7) and (A.9), respectively. Moreover, the second term in (A.13) comes from

$$h(\delta)^2 \cdot n = \left(-\frac{\delta}{n} \pm \frac{1}{n} \sqrt{\frac{\delta^2 m(1-n) + (mn-1)}{m-1}} \right)^2 \cdot n$$

$$= \left(\frac{\delta^2}{n^2} + \frac{\delta^2 m(1-n) + (mn-1)}{(m-1)n^2} \mp \frac{2\delta}{n^2} \sqrt{\frac{\delta^2 m(1-n) + (mn-1)}{m-1}} \right) \cdot n$$

$$= \left(\frac{\delta^2(2m-mn-1)}{(m-1)n^2} + \frac{mn-1}{(m-1)n^2} \mp \frac{2\delta}{n^2} \sqrt{\frac{\delta^2 m(1-n) + (mn-1)}{m-1}} \right) \cdot n$$

$$= \frac{\delta^2(2m-mn-1)}{(m-1)n} \mp \frac{2\delta}{n} \sqrt{\frac{\delta^2 m(1-n) + (mn-1)}{m-1}} + \frac{mn-1}{(m-1)n}$$

$$= \frac{\delta^2(2m-mn-1)}{(m-1)n} - \frac{2\delta^2}{n} - 2\delta h(\delta) + \frac{mn-1}{(m-1)n}$$

$$= -2\delta h(\delta) - \frac{\delta(mn-1)}{(m-1)n} + \frac{mn-1}{(m-1)n}$$

$$= -2\delta h(\delta) + \frac{mn-1}{(m-1)n} \cdot (1-\delta^2).$$

[**Condition of** (A.3)] For all $i \in [m], j \neq j' \in [n], k, k' \in [p]$,

$$(\boldsymbol{u}_{i,j,k}^{\delta})^{\top} \boldsymbol{u}_{i,j',k'}^{\delta} = \delta^2 \boldsymbol{w}_{i,j}^{\top} \boldsymbol{w}_{i,j'} + h(\delta)^2 \boldsymbol{v}_i^{\top} \boldsymbol{v}_i + \delta h(\delta) \boldsymbol{w}_{i,j}^{\top} \boldsymbol{v}_i + \delta h(\delta) \boldsymbol{w}_{i,j'}^{\top} \boldsymbol{v}_i$$

$$= \delta^2 \boldsymbol{w}_{i,j}^{\top} \boldsymbol{w}_{i,j'} + h(\delta)^2 \boldsymbol{v}_i^{\top} \boldsymbol{v}_i + 2\delta h(\delta) \boldsymbol{w}_{i,j}^{\top} \boldsymbol{v}_i$$

$$= \delta^2 \boldsymbol{w}_{i,j}^{\top} \boldsymbol{w}_{i,j'} + (\boldsymbol{u}_{i,j,k}^{\delta})^{\top} \boldsymbol{u}_{i,j,k'}^{\delta} - \delta^2 \boldsymbol{w}_{i,j}^{\top} \boldsymbol{w}_{i,j} \tag{A.14}$$

$$= -\delta^2 \frac{1}{mn-1} + 1 - \delta^2 \tag{A.15}$$

$$= 1 - \delta^2 \frac{mn}{mn-1},$$

where (A.14) follows from (A.11), and the first and third terms in (A.15) come form the definition of the $(mn-1)$-simplex ETF in Def. 1.

[**Condition of** (A.4)] For all $i \neq i' \in [m], j, j' \in [n], k, k' \in [p]$,

$$
\begin{aligned}
(\boldsymbol{u}_{i,j,k}^{\delta})^{\top} \boldsymbol{u}_{i',j',k'}^{\delta} &= \delta^2 \boldsymbol{w}_{i,j}^{\top} \boldsymbol{w}_{i',j'} + h(\delta)^2 \boldsymbol{v}_i^{\top} \boldsymbol{v}_{i'} + \delta h(\delta) \boldsymbol{w}_{i,j}^{\top} \boldsymbol{v}_{i'} + \delta h(\delta) \boldsymbol{w}_{i',j'}^{\top} \boldsymbol{v}_i \\
&= \delta^2 \boldsymbol{w}_{i,j}^{\top} \boldsymbol{w}_{i',j'} + h(\delta)^2 \boldsymbol{v}_i^{\top} \boldsymbol{v}_{i'} + 2\delta h(\delta) \boldsymbol{w}_{i,j}^{\top} \boldsymbol{v}_{i'} \\
&= -\delta^2 \frac{1}{mn-1} - h(\delta)^2 \frac{n^2}{mn-1} - 2\delta h(\delta) \frac{n}{mn-1} &&\text{(A.16)} \\
&= -\delta^2 \frac{1}{mn-1} - \frac{1}{m-1} \cdot \left(h(\delta)^2 \frac{(m-1)n^2}{mn-1} - 2\delta h(\delta) \frac{(m-1)n}{mn-1} \right) \\
&= -\delta^2 \frac{1}{mn-1} - \frac{1}{m-1} \cdot (1 - \delta^2) &&\text{(A.17)} \\
&= -\frac{1}{m-1} + \delta^2 \frac{m(n-1)}{(m-1)(mn-1)},
\end{aligned}
$$

where the first term in (A.16) comes from the definition of the the the $(mn-1)$-simplex ETF in Def. 1, and the second and third terms in (A.16) comes from (A.8) and (A.10), respectively. Moreover, (A.17) comes from reformatting (A.12) which is equal to 1. $\qquad\square$

A.3 The Optimality of SSEM

First, we prove one proposition and two lemmas that are necessary for proving the theorem, which shows the optimality of SSEM.

Proposition A.5. *Let $\boldsymbol{a} \in \mathbb{R}^d$ be a given vector where all elements are non-negative. Then, the following holds for any set of n vectors $\{\boldsymbol{w}_i \in \mathbb{R}^d\}_{i \in [n]}$:*

$$\frac{1}{n} \sum_{i \in [n]} \log \left(\boldsymbol{a}^\top \exp(\boldsymbol{w}_i) \right) \geq \log \left(\boldsymbol{a}^\top \exp \left(\frac{1}{n} \sum_{i \in [n]} \boldsymbol{w}_i \right) \right), \tag{A.18}$$

where exp *is an element-wise exponential function.*

When the given vector $\boldsymbol{a}$ has all positive elements, the equality condition of (A.18) is

$$\boldsymbol{w}_i = \boldsymbol{c} \qquad \forall i \in [n],$$

for some $\boldsymbol{c} \in \mathbb{R}^d$.

Proof. Note that the both terms inside the logarithm in (A.18) equal zero when $\boldsymbol{a} = \boldsymbol{0} \in \mathbb{R}^d$, *i.e.,*

$$\boldsymbol{a}^\top \exp(\boldsymbol{w}_i) = 0 = \boldsymbol{a}^\top \exp \left(\frac{1}{n} \sum_{i \in [n]} \boldsymbol{w}_i \right).$$

Therefore, it suffices to consider the case where $\boldsymbol{a}$ contains only positive elements.

Let the continuous function $f(\boldsymbol{w}) = \log \left(\boldsymbol{a}^\top \exp(\boldsymbol{w}) \right)$ be defined for the given vector $\boldsymbol{a}$. Then, for any $\boldsymbol{w}, \boldsymbol{v} \in \mathbb{R}^d$, the followings hold:

$$\begin{aligned} \frac{1}{2} f(\boldsymbol{w}) + \left(1 - \frac{1}{2} \right) f(\boldsymbol{v}) &= \frac{1}{2} \left(\log \left(\boldsymbol{a}^\top \exp(\boldsymbol{w}) \right) + \log \left(\boldsymbol{a}^\top \exp(\boldsymbol{v}) \right) \right) \\ &= \log \sqrt{\boldsymbol{a}^\top \exp(\boldsymbol{w}) \cdot \boldsymbol{a}^\top \exp(\boldsymbol{v})} \\ &\geq \log \left(\boldsymbol{a}^\top \exp \left(\frac{1}{2} \boldsymbol{w} + \frac{1}{2} \boldsymbol{v} \right) \right) \\ &= f \left(\frac{1}{2} \boldsymbol{w} + \left(1 - \frac{1}{2} \right) \boldsymbol{v} \right), \end{aligned} \tag{A.19}$$

where the equality condition of (A.19) is $\exp \left(\frac{1}{2} \boldsymbol{w} \right) = \exp \left(\frac{1}{2} \boldsymbol{v} \right)$ from the Cauchy–Schwarz inequality. This directly implies that f is a convex function due to the continuity of f.

As a result, (A.18) holds from the Jensen's inequality as below:

$$\frac{1}{n} \sum_{i \in [n]} \log \left(\boldsymbol{a}^\top \exp(\boldsymbol{w}_i) \right) = \frac{1}{n} \sum_{i \in [n]} f(\boldsymbol{w}_i) \geq f \left(\frac{1}{n} \sum_{i \in [n]} \boldsymbol{w}_i \right) = \log \left(\boldsymbol{a}^\top \exp \left(\frac{1}{n} \sum_{i \in [n]} \boldsymbol{w}_i \right) \right),$$

where the equality condition of (A.19) is simplified as

$$\boldsymbol{w}_i = \boldsymbol{c} \qquad \forall i \in [n],$$

for some $\boldsymbol{c} \in \mathbb{R}^d$. $\qquad \square$

Lemma A.1. *Let $\boldsymbol{U} := \{\boldsymbol{u}_{i,j,k}\}_{i \in [m], j \in [n], k \in [p]}$ be a set of mnp vectors in $\mathbb{R}^d$, satisfying $\|\boldsymbol{u}\|_2^2 = 1$ for all $\boldsymbol{u} \in \boldsymbol{U}$. Additionally, define the sets $\boldsymbol{U}_{i,j} := \{\boldsymbol{u}_{i,j,k}\}_{k \in [p]}$ and $\boldsymbol{U}_i := \cup_{j \in [n]} \boldsymbol{U}_{i,j}$ for all $i \in [m]$ and $j \in [n]$. Then, for every constant $c \in [-mn, mn(n-1)]$, there exists a unique $\delta^\star(c) \in \left[0, \sqrt{\frac{mn-1}{m(n-1)}} \right]$ such that*

$$\boldsymbol{U}^{\delta^\star(c)} \in \underset{\boldsymbol{U}}{\arg\min} \left\{ \sum_{i \in [m]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \sum_{\boldsymbol{v} \in \boldsymbol{U} \setminus \boldsymbol{U}_i} \boldsymbol{u}^\top \boldsymbol{v} \,\middle|\, \sum_{i \in [m], j \neq j' \in [n]} \mathbb{E}[\boldsymbol{U}_{i,j}]^\top \mathbb{E}[\boldsymbol{U}_{i,j'}] = c \right\}.$$

Specifically, $\delta^\star(c) = \sqrt{\frac{mn-1}{mn} - \frac{mn-1}{m^2 n^2 (n-1)} c}$.

Proof. First of all, the possible values of c are bounded within the interval $[-mn, mn(n-1)]$: the maximum value of c can be achieved when $\boldsymbol{u} = \boldsymbol{v}$ for all $\boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{U}$, while the minimum is determined as below.

$$c = \sum_{i \in [m], j \neq j' \in [n]} \mathbb{E}[\boldsymbol{U}_{i,j}]^\top \mathbb{E}[\boldsymbol{U}_{i,j'}] \tag{A.20}$$

$$= \frac{1}{p^2} \sum_{i \in [m], j \neq j' \in [n]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_{i,j'}} \boldsymbol{u}^\top \boldsymbol{v}$$

$$= \frac{1}{p^2} \sum_{i \in [m]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \sum_{\boldsymbol{v} \in \boldsymbol{U}_i} \boldsymbol{u}^\top \boldsymbol{v} - \frac{1}{p^2} \sum_{i \in [m], j \in [n]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_{i,j}} \boldsymbol{u}^\top \boldsymbol{v}$$

$$= \frac{1}{p^2} \sum_{i \in [m]} \left\| \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \boldsymbol{u} \right\|_2^2 - \frac{1}{p^2} \sum_{i \in [m], j \in [n]} \left\| \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \boldsymbol{u} \right\|_2^2$$

$$\geq 0 - mn$$

Now consider the minimization problem. Given that $\|\boldsymbol{u}\|_2^2 = 1$ for all $\boldsymbol{u} \in \boldsymbol{U}$, the following holds:

$$\left\| \sum_{\boldsymbol{u} \in \boldsymbol{U}} \boldsymbol{u} \right\|_2^2 = \sum_{\boldsymbol{u} \in \boldsymbol{U}} \sum_{\boldsymbol{v} \in \boldsymbol{U}} \boldsymbol{u}^\top \boldsymbol{v}$$

$$= \sum_{i \in [m], j \in [n]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_{i,j}} \boldsymbol{u}^\top \boldsymbol{v} + \sum_{i \in [m], j \neq j' \in [n]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_{i,j'}} \boldsymbol{u}^\top \boldsymbol{v} + \sum_{i \in [m]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \sum_{\boldsymbol{v} \in \boldsymbol{U} \setminus \boldsymbol{U}_i} \boldsymbol{u}^\top \boldsymbol{v},$$

which implies

$$\sum_{i \in [m]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_i} \sum_{\boldsymbol{v} \in \boldsymbol{U} \setminus \boldsymbol{U}_i} \boldsymbol{u}^\top \boldsymbol{v} = \left\| \sum_{\boldsymbol{u} \in \boldsymbol{U}} \boldsymbol{u} \right\|_2^2 - \sum_{i \in [m], j \in [n]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_{i,j}} \boldsymbol{u}^\top \boldsymbol{v} - \sum_{i \in [m], j \neq j' \in [n]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_{i,j'}} \boldsymbol{u}^\top \boldsymbol{v}$$

$$\geq 0 - \sum_{i \in [m], j \in [n]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_{i,j}} 1 - \sum_{i \in [m], j \neq j' \in [n]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_{i,j'}} \boldsymbol{u}^\top \boldsymbol{v} \tag{A.21}$$

$$= -mnp^2 - p^2 c. \tag{A.22}$$

Note that the equality conditions of (A.21) are

$$\sum_{\boldsymbol{u} \in \boldsymbol{U}} \boldsymbol{u} = 0, \tag{A.23}$$

$$\boldsymbol{u}^\top \boldsymbol{v} = 1 \qquad\qquad \forall i \in [m], j \in [n], \boldsymbol{u} \in \boldsymbol{U}_{i,j}, \boldsymbol{v} \in \boldsymbol{U}_{i,j}, \tag{A.24}$$

implying that the centroid of the embedding vectors is at the origin and that every embedding vector of each instance is the same, regardless of augmentations.

For any $\delta \in \left[0, \sqrt{\frac{mn-1}{m(n-1)}}\right]$, embedding vectors $\boldsymbol{U}^\delta$ of SSEM in Def. A.2 fulfill the equality conditions in (A.23) and (A.24), as follows:

$$\left\| \sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}^\delta} \boldsymbol{u}^\delta \right\|_2^2 = \sum_{i \in [m], j \in [n]} \sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}_{i,j}^\delta} \sum_{\boldsymbol{v}^\delta \in \boldsymbol{U}_{i,j}^\delta} \left(\boldsymbol{u}^\delta\right)^\top \boldsymbol{v}^\delta + \sum_{i \in [m], j \neq j' \in [n]} \sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}_{i,j}^\delta} \sum_{\boldsymbol{v}^\delta \in \boldsymbol{U}_{i,j'}^\delta} \left(\boldsymbol{u}^\delta\right)^\top \boldsymbol{v}^\delta$$

$$+ \sum_{i \in [m]} \sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}_i^\delta} \sum_{\boldsymbol{v}^\delta \in \boldsymbol{U}^\delta \setminus \boldsymbol{U}_i^\delta} \left(\boldsymbol{u}^\delta\right)^\top \boldsymbol{v}^\delta$$

$$= mnp^2 + mn(n-1)p^2 \left(1 - \delta^2 \frac{mn}{mn-1}\right) + m(m-1)n^2 p^2 \left(-\frac{1}{m-1} + \delta^2 \frac{m(n-1)}{(m-1)(mn-1)}\right)$$

$$= 0,$$

implying (A.23) holds, and (A.24) holds from (A.2) in Def A.2.

Therefore, we can conclude the existence of a unique $\delta \in \left[0, \sqrt{\frac{mn-1}{m(n-1)}}\right]$ such that $\boldsymbol{U}^\delta$ of SSEM represents the optimal embedding from (A.22), specified as:

$$-mnp^2 - p^2 = \sum_{i \in [m]} \sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}_i^\delta} \sum_{\boldsymbol{v}^\delta \in \boldsymbol{U}^\delta \setminus \boldsymbol{U}_i^\delta} \left(\boldsymbol{u}^\delta\right)^\top \boldsymbol{v}^\delta$$

$$= m(m-1)n^2p^2\left(-\frac{1}{m-1} + \delta^2 \frac{m(n-1)}{(m-1)(mn-1)}\right)$$

$$= -mn^2p^2 + \delta^2 \frac{m^2n^2(n-1)p^2}{mn-1},$$

which is equal to

$$\delta = \sqrt{\frac{mn-1}{m^2n^2(n-1)p^2}\left(mn(n-1)p^2 - p^2\right)}$$

$$= \sqrt{\frac{mn-1}{mn} - \frac{mn-1}{m^2n^2(n-1)}c}. \tag{A.25}$$

The uniqueness of δ comes from the fact that (A.25) is a strictly decreasing function of $c \in [-mn, mn(n-1)]$. $\qquad\square$

Lemma A.2. *Let $\boldsymbol{U} := \{\boldsymbol{u}_{i,j,k}\}_{i \in [m], j \in [n], k \in [p]}$ be a set of mnp vectors in $\mathbb{R}^d$, satisfying $\|\boldsymbol{u}\|_2^2 = 1$ for all $\boldsymbol{u} \in \boldsymbol{U}$. Additionally, define the sets $\boldsymbol{U}_{i,j} := \{\boldsymbol{u}_{i,j,k}\}_{k \in [p]}$ and $\boldsymbol{U}_i := \cup_{j \in [n]} \boldsymbol{U}_{i,j}$ for all $i \in [m]$ and $j \in [n]$. Then, for every constant $c \in [0, 2mn^2]$, there exists a unique $\delta^\star(c) \in \left[0, \sqrt{\frac{mn-1}{mn(n-1)}}\right]$ such that*

$$\boldsymbol{U}^{\delta^\star(c)} \in \arg\max_{\boldsymbol{U}} \left\{ \sum_{i \in [m], j \neq j' \in [n]} \mathbb{E}[\boldsymbol{U}_{i,j}]^\top \mathbb{E}[\boldsymbol{U}_{i,j'}] \,\middle|\, \sum_{i \in [m], j \neq j' \in [n]} \left\|\mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}]\right\|_2^2 = c \right\}. \tag{A.26}$$

Specifically, $\delta^\star(c) = \sqrt{\frac{mn-1}{2m^2n^2(n-1)} \cdot c}$.

Proof. First of all, the possible values of c are bounded within the interval $[0, 2mn^2]$, because $\|\boldsymbol{u}\|_2^2 = 1$ for all $\boldsymbol{u} \in \boldsymbol{U}$. Especially, the possible maximum value of c, which is $2mn^2$, can be attained as below.

$$c = \sum_{i \in [m], j \neq j' \in [n]} \left\|\mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}]\right\|_2^2$$

$$= \sum_{i \in [m], j \neq j' \in [n]} \left(\left\|\mathbb{E}[\boldsymbol{U}_{i,j}]\right\|_2^2 + \left\|\mathbb{E}[\boldsymbol{U}_{i,j'}]\right\|_2^2 - 2\mathbb{E}[\boldsymbol{U}_{i,j}]^\top \mathbb{E}[\boldsymbol{U}_{i,j'}]\right)$$

$$= 2(n-1) \sum_{i \in [m], j \in [n]} \left\|\mathbb{E}[\boldsymbol{U}_{i,j}]\right\|_2^2 - 2 \sum_{i \in [m], j \neq j' \in [n]} \mathbb{E}[\boldsymbol{U}_{i,j}]^\top \mathbb{E}[\boldsymbol{U}_{i,j'}] \tag{A.27}$$

$$= 2(n-1) \sum_{i \in [m], j \in [n]} \left\|\mathbb{E}[\boldsymbol{U}_{i,j}]\right\|_2^2 - 2\left(\sum_{i \in [m]} \left\|\sum_{j \in [n]} \mathbb{E}[\boldsymbol{U}_{i,j}]\right\|_2^2 - \sum_{i \in [m], j \in [n]} \left\|\mathbb{E}[\boldsymbol{U}_{i,j}]\right\|_2^2\right)$$

$$= 2n \sum_{i \in [m], j \in [n]} \left\|\mathbb{E}[\boldsymbol{U}_{i,j}]\right\|_2^2 - 2 \sum_{i \in [m]} \left\|\sum_{j \in [n]} \mathbb{E}[\boldsymbol{U}_{i,j}]\right\|_2^2$$

$$\leq 2n \sum_{i \in [m], j \in [n]} 1 - 0$$

$$= 2mn^2,$$

where $\left\|\mathbb{E}[\boldsymbol{U}_{i,j}]\right\|_2^2 = \left\|\frac{1}{p}\sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \boldsymbol{u}\right\|_2^2 \leq \frac{1}{p}\sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \|\boldsymbol{u}\|_2^2 = 1$ from Jensen's inequality.

Now, consider the main maximization problem as given in (A.26). From (A.27),

$$
\begin{aligned}
c &= 2(n-1) \sum_{i \in [m], j \in [n]} \left\| \mathbb{E}[\boldsymbol{U}_{i,j}] \right\|_2^2 - 2 \sum_{i \in [m], j \neq j' \in [n]} \mathbb{E}[\boldsymbol{U}_{i,j}]^\top \mathbb{E}[\boldsymbol{U}_{i,j'}] \\
&\leq 2(n-1) \sum_{i \in [m], j \in [n]} 1 - 2 \sum_{i \in [m], j \neq j' \in [n]} \mathbb{E}[\boldsymbol{U}_{i,j}]^\top \mathbb{E}[\boldsymbol{U}_{i,j'}] \qquad\qquad\text{(A.28)} \\
&= 2mn(n-1) - 2 \sum_{i \in [m], j \neq j' \in [n]} \mathbb{E}[\boldsymbol{U}_{i,j}]^\top \mathbb{E}[\boldsymbol{U}_{i,j'}]
\end{aligned}
$$

which implies

$$
\sum_{i \in [m], j \neq j' \in [n]} \mathbb{E}[\boldsymbol{U}_{i,j}]^\top \mathbb{E}[\boldsymbol{U}_{i,j'}] \leq mn(n-1) - \frac{c}{2}. \tag{A.29}
$$

The equality condition of (A.28) is

$$
\left\| \mathbb{E}[\boldsymbol{U}_{i,j}] \right\|_2^2 = 1 \qquad \forall i \in [m], j \in [n],
$$

implying that every embedding vector of each instance is the same, regardless of augmentation. For any $\delta \in \left[0, \sqrt{\frac{mn-1}{m(n-1)}}\right]$, the set of embedding vectors $\boldsymbol{U}^\delta$ of SSEM fulfills the equality condition in (A.28) from (A.2) in Def A.2 as follows:

$$
\left\| \mathbb{E}\left[\boldsymbol{U}_{i,j}^\delta\right] \right\|_2^2 = \left\| \frac{1}{p} \sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}_{i,j}^\delta} \boldsymbol{u}^\delta \right\|_2^2 = \frac{1}{p^2} \sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}_{i,j}^\delta} \sum_{\boldsymbol{v}^\delta \in \boldsymbol{U}_{i,j}^\delta} \left(\boldsymbol{u}^\delta\right)^\top \boldsymbol{v}^\delta = 1 \qquad\qquad \forall i \in [m], j \in [n].
$$

Therefore, we can conclude the existence of a unique $\delta \in \left[0, \sqrt{\frac{mn-1}{m(n-1)}}\right]$ such that $\boldsymbol{U}^\delta$ of SSEM represents the optimal embedding from (A.29), specified as:

$$
\begin{aligned}
mn(n-1)p^2 - \frac{cp^2}{2} &= p^2 \sum_{i \in [m], j \neq j' \in [n]} \mathbb{E}[\boldsymbol{U}_{i,j}]^\top \mathbb{E}[\boldsymbol{U}_{i,j'}] \\
&= \sum_{i \in [m], j \neq j' \in [n]} \sum_{\boldsymbol{u}^\delta \in \boldsymbol{U}_{i,j}^\delta} \sum_{\boldsymbol{v}^\delta \in \boldsymbol{U}_{i,j'}^\delta} \left(\boldsymbol{u}^\delta\right)^\top \boldsymbol{v}^\delta \\
&= mn(n-1)p^2 \left(1 - \delta^2 \frac{mn}{mn-1}\right) \\
&= mn(n-1)p^2 - \delta^2 \frac{m^2 n^2 (n-1)p^2}{mn-1},
\end{aligned}
$$

which is equal to

$$
\delta = \sqrt{\frac{mn-1}{2m^2 n^2 (n-1)} \cdot c} \quad . \tag{A.30}
$$

The uniqueness of δ comes from the fact that (A.30) is a strictly increasing function of $c \in [0, 2mn^2]$. $\qquad\square$

Theorem A.2 (Optimality of SSEM). *Suppose $mn \geq 2$ and $d \geq mn-1$ hold. Then, all embedding sets $\boldsymbol{U}^\star$ that minimize the loss $\mathcal{L}(\boldsymbol{U})$ in (1) are included in the SSEM in Def. 2, i.e.,*

$$
\forall \boldsymbol{U}^\star \in \arg\min_{\boldsymbol{U}} \mathcal{L}(\boldsymbol{U}), \quad \exists! \delta \in [0,1] \text{ such that } \boldsymbol{U}^\delta = \boldsymbol{U}^\star.
$$

Specifically,

$$
\delta^\star = \begin{cases} 0, & \text{if } h\big(0; m, n, \tau, \alpha\big) \geq 0, \\ \delta \in (0,1] \text{ such that } h\left(\delta^2 \frac{mn}{mn-1}; m, n, \tau, \alpha\right) = 0, & \text{otherwise,} \end{cases} \tag{A.31}
$$

where

$$h(x; m, n, \tau, \alpha) = (1 - \alpha) - \alpha(n - 1) \cdot \exp(-x/\tau)$$
$$+ (mn - 1 - \alpha(m - 1)n) \cdot \exp\left(\left(-\frac{m}{m-1} + x\frac{n-1}{(m-1)n}\right)/\tau\right).$$

Proof. We want to find the optimal embeddings that minimize a SupCL loss $\mathcal{L}(\boldsymbol{U})$ as defined in (1), which is a convex combination of $\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U})$ in (4) and $\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U})$ in (5). These can be rewritten as follows:

$$\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U}) = -\frac{1}{mn(n-1)p^2} \sum_{i \in [m], j \neq j' \in [n]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_{i,j'}} \log \frac{\exp(\boldsymbol{u}^\top \boldsymbol{v}/\tau)}{\sum_{\boldsymbol{w} \in \boldsymbol{U}} \exp(\boldsymbol{u}^\top \boldsymbol{w}/\tau)}$$

$$= \frac{1}{mn(n-1)p^2} \sum_{i \in [m], j \neq j' \in [n]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_{i,j'}} \log \left(\sum_{\boldsymbol{w} \in \boldsymbol{U}} \exp(\boldsymbol{u}^\top (\boldsymbol{w} - \boldsymbol{v})/\tau) \right)$$

$$= \frac{1}{mn(n-1)p^2} \sum_{i \in [m], j \in [n]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_i \setminus \boldsymbol{U}_{i,j}} \log \left(\sum_{\boldsymbol{w} \in \boldsymbol{U}} \exp(\boldsymbol{u}^\top (\boldsymbol{w} - \boldsymbol{v})/\tau) \right), \qquad (A.32)$$

$$\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U}) = -\frac{1}{mnp^2} \sum_{i \in [m], j \in [n]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_{i,j}} \log \frac{\exp(\boldsymbol{u}^\top \boldsymbol{v}/\tau)}{\sum_{\boldsymbol{w} \in \boldsymbol{U}} \exp(\boldsymbol{u}^\top \boldsymbol{w}/\tau)}$$

$$= \frac{1}{mnp^2} \sum_{i \in [m], j \in [n]} \sum_{\boldsymbol{u} \in \boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v} \in \boldsymbol{U}_{i,j}} \log \left(\sum_{\boldsymbol{w} \in \boldsymbol{U}} \exp(\boldsymbol{u}^\top (\boldsymbol{w} - \boldsymbol{v})/\tau) \right). \qquad (A.33)$$

We first consider minimizing $\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U})$ in (A.32), and then minimizing $\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U})$ in (A.33) in a similar manner.

Note that the set of all embeddings $\boldsymbol{U}$ can be partitioned to three disjoint sets as

$$\boldsymbol{U} = (\boldsymbol{U} \setminus \boldsymbol{U}_i) \dot{\cup} (\boldsymbol{U}_i \setminus \boldsymbol{U}_{i,j}) \dot{\cup} \boldsymbol{U}_{i,j} \qquad \forall i \in [m], j \in [n],$$

where $\dot{\cup}$ denotes the disjoint union. Then, the term inside the logarithm in (A.32) can be decomposed to three terms. Specifically, for all $i \in [m], j \neq j' \in [n]$, $\boldsymbol{u} \in \boldsymbol{U}_{i,j}$, and $\boldsymbol{v} \in \boldsymbol{U}_{i,j'}$,

$$\sum_{\boldsymbol{w} \in \boldsymbol{U}} \exp(\boldsymbol{u}^\top (\boldsymbol{w} - \boldsymbol{v})/\tau) = \sum_{\boldsymbol{w} \in \boldsymbol{U} \setminus \boldsymbol{U}_i} \exp(\boldsymbol{u}^\top (\boldsymbol{w} - \boldsymbol{v})/\tau) + \sum_{\boldsymbol{w} \in \boldsymbol{U}_i \setminus \boldsymbol{U}_{i,j}} \exp(\boldsymbol{u}^\top (\boldsymbol{w} - \boldsymbol{v})/\tau) + \sum_{\boldsymbol{w} \in \boldsymbol{U}_{i,j}} \exp(\boldsymbol{u}^\top (\boldsymbol{w} - \boldsymbol{v})/\tau)$$

$$\geq (m-1)np \cdot \exp\left(\frac{1}{(m-1)np} \sum_{\boldsymbol{w} \in \boldsymbol{U} \setminus \boldsymbol{U}_i} \boldsymbol{u}^\top \boldsymbol{w}/\tau - \boldsymbol{u}^\top \boldsymbol{v}/\tau\right)$$

$$+ (n-1)p \cdot \exp\left(\frac{1}{(n-1)p} \sum_{\boldsymbol{w} \in \boldsymbol{U}_i \setminus \boldsymbol{U}_{i,j}} \boldsymbol{u}^\top \boldsymbol{w}/\tau - \boldsymbol{u}^\top \boldsymbol{v}/\tau\right)$$

$$+ p \cdot \exp\left(\frac{1}{p} \sum_{\boldsymbol{w} \in \boldsymbol{U}_{i,j}} \boldsymbol{u}^\top \boldsymbol{w}/\tau - \boldsymbol{u}^\top \boldsymbol{v}/\tau\right), \qquad (A.34)$$

where the inequality in (A.34) comes from using Jensen's inequality three times. The equality in (A.34) is achieved if there exist some constants $c_1, c_2, c_3 \in \mathbb{R}$ such that the following conditions hold for all $i \in [m], j \neq j' \in [n]$, $\boldsymbol{u} \in \boldsymbol{U}_{i,j}$, and $\boldsymbol{v} \in \boldsymbol{U}_{i,j'}$:

$$\boldsymbol{u}^\top (\boldsymbol{w} - \boldsymbol{v}) = c_1 \qquad \forall \boldsymbol{w} \in \boldsymbol{U} \setminus \boldsymbol{U}_i, \qquad (A.35)$$
$$\boldsymbol{u}^\top (\boldsymbol{w} - \boldsymbol{v}) = c_2 \qquad \forall \boldsymbol{w} \in \boldsymbol{U}_i \setminus \boldsymbol{U}_{i,j}, \qquad (A.36)$$
$$\boldsymbol{u}^\top (\boldsymbol{w} - \boldsymbol{v}) = c_3 \qquad \forall \boldsymbol{w} \in \boldsymbol{U}_{i,j}. \qquad (A.37)$$

By using the above result in (A.34) to $\mathcal{L}_{\text{Sup}}(\boldsymbol{U})$ in (A.32),

$$
\mathcal{L}_{\text{Sup}}(\boldsymbol{U}) \geq \frac{1}{mn(n-1)p^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_i\backslash\boldsymbol{U}_{i,j}} \log\left((m-1)np\cdot\exp\left(\frac{1}{(m-1)np}\sum_{\boldsymbol{w}\in\boldsymbol{U}\backslash\boldsymbol{U}_i}\boldsymbol{u}^\top\boldsymbol{w}/\tau - \boldsymbol{u}^\top\boldsymbol{v}/\tau\right)\right.
$$
$$
+ (n-1)p\cdot\exp\left(\frac{1}{(n-1)p}\sum_{\boldsymbol{w}\in\boldsymbol{U}_i\backslash\boldsymbol{U}_{i,j}}\boldsymbol{u}^\top\boldsymbol{w}/\tau - \boldsymbol{u}^\top\boldsymbol{v}/\tau\right)
$$
$$
\left. + p\cdot\exp\left(\frac{1}{p}\sum_{\boldsymbol{w}\in\boldsymbol{U}_{i,j}}\boldsymbol{u}^\top\boldsymbol{w}/\tau - \boldsymbol{u}^\top\boldsymbol{v}/\tau\right)\right)
$$

$$
\geq \frac{1}{m}\sum_{i\in[m]}\log\left((m-1)np\cdot\exp\left(\frac{1}{n(n-1)p^2}\sum_{j\in[n]}\sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}}\sum_{\boldsymbol{v}\in\boldsymbol{U}_i\backslash\boldsymbol{U}_{i,j}}\frac{1}{(m-1)np}\sum_{\boldsymbol{w}\in\boldsymbol{U}\backslash\boldsymbol{U}_i}\boldsymbol{u}^\top\boldsymbol{w}/\tau\right.\right.
$$
$$
\left. - \frac{1}{n(n-1)p^2}\sum_{j\in[n]}\sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}}\sum_{\boldsymbol{v}\in\boldsymbol{U}_i\backslash\boldsymbol{U}_{i,j}}\boldsymbol{u}^\top\boldsymbol{v}/\tau\right)
$$
$$
+ (n-1)p\cdot\exp\left(\frac{1}{n(n-1)p^2}\sum_{j\in[n]}\sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}}\sum_{\boldsymbol{v}\in\boldsymbol{U}_i\backslash\boldsymbol{U}_{i,j}}\frac{1}{(n-1)p}\sum_{\boldsymbol{w}\in\boldsymbol{U}_i\backslash\boldsymbol{U}_{i,j}}\boldsymbol{u}^\top\boldsymbol{w}/\tau\right.
$$
$$
\left. - \frac{1}{n(n-1)p^2}\sum_{j\in[n]}\sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}}\sum_{\boldsymbol{v}\in\boldsymbol{U}_i\backslash\boldsymbol{U}_{i,j}}\boldsymbol{u}^\top\boldsymbol{v}/\tau\right)
$$
$$
+ p\cdot\exp\left(\frac{1}{n(n-1)p^2}\sum_{j\in[n]}\sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}}\sum_{\boldsymbol{v}\in\boldsymbol{U}_i\backslash\boldsymbol{U}_{i,j}}\frac{1}{p}\sum_{\boldsymbol{w}\in\boldsymbol{U}_{i,j}}\boldsymbol{u}^\top\boldsymbol{w}/\tau\right.
$$
$$
\left.\left. - \frac{1}{n(n-1)p^2}\sum_{j\in[n]}\sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}}\sum_{\boldsymbol{v}\in\boldsymbol{U}_i\backslash\boldsymbol{U}_{i,j}}\boldsymbol{u}^\top\boldsymbol{v}/\tau\right)\right) \tag{A.38}
$$

$$
= \frac{1}{m}\sum_{i\in[m]}\log\left((m-1)np\cdot\exp\left(\frac{1}{(m-1)n^2p^2}\sum_{\boldsymbol{u}\in\boldsymbol{U}_i}\sum_{\boldsymbol{w}\in\boldsymbol{U}\backslash\boldsymbol{U}_i}\boldsymbol{u}^\top\boldsymbol{w}/\tau\right.\right.
$$
$$
\left. - \frac{1}{n(n-1)p^2}\sum_{j\in[n]}\sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}}\sum_{\boldsymbol{v}\in\boldsymbol{U}_i\backslash\boldsymbol{U}_{i,j}}\boldsymbol{u}^\top\boldsymbol{v}/\tau\right)
$$
$$
+ (n-1)p\cdot\exp\left(\frac{1}{n(n-1)p^2}\sum_{j\in[n]}\sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}}\sum_{\boldsymbol{w}\in\boldsymbol{U}_i\backslash\boldsymbol{U}_{i,j}}\boldsymbol{u}^\top\boldsymbol{w}/\tau\right.
$$
$$
\left. - \frac{1}{n(n-1)p^2}\sum_{j\in[n]}\sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}}\sum_{\boldsymbol{v}\in\boldsymbol{U}_i\backslash\boldsymbol{U}_{i,j}}\boldsymbol{u}^\top\boldsymbol{v}/\tau\right)
$$
$$
\left.\left. + p\cdot\exp\left(\frac{1}{np^2}\sum_{j\in[n]}\sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}}\sum_{\boldsymbol{w}\in\boldsymbol{U}_{i,j}}\boldsymbol{u}^\top\boldsymbol{w}/\tau - \frac{1}{n(n-1)p^2}\sum_{j\in[n]}\sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}}\sum_{\boldsymbol{v}\in\boldsymbol{U}_i\backslash\boldsymbol{U}_{i,j}}\boldsymbol{u}^\top\boldsymbol{v}/\tau\right)\right)\right) \tag{A.39}
$$

where the inequality in (A.38) holds from Proposition A.5. Note that the value inside the second exponential in (A.39) is zero.

Moreover, applying Proposition A.5 one more to (A.39) results in

$$
\begin{aligned}
\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U}) \geq \log \bigg(& (m-1)np \cdot \exp \bigg(\frac{1}{m(m-1)n^2p^2} \sum_{i\in[m]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_i} \sum_{\boldsymbol{w}\in\boldsymbol{U}\setminus\boldsymbol{U}_i} \boldsymbol{u}^\top\boldsymbol{w}/\tau \\
& - \frac{1}{mn(n-1)p^2} \sum_{i\in[m]} \sum_{j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_i\setminus\boldsymbol{U}_{i,j}} \boldsymbol{u}^\top\boldsymbol{v}/\tau \bigg) \\
& + (n-1)p \\
& + p \cdot \exp \bigg(\frac{1}{mnp^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{w}\in\boldsymbol{U}_{i,j}} \boldsymbol{u}^\top\boldsymbol{w}/\tau \\
& - \frac{1}{mn(n-1)p^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_i\setminus\boldsymbol{U}_{i,j}} \boldsymbol{u}^\top\boldsymbol{v}/\tau \bigg) \bigg).
\end{aligned}
\tag{A.40}
$$

The equality conditions of (A.38) and (A.40), which use Proposition A.5, are also achieved if the conditions in (A.35), (A.36), and (A.37) are satisfied.

Note that the value inside the first exponential in (A.40) follows the below inequality:

$$
\begin{aligned}
& \frac{1}{m(m-1)n^2p^2} \sum_{i\in[m]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_i} \sum_{\boldsymbol{w}\in\boldsymbol{U}\setminus\boldsymbol{U}_i} \boldsymbol{u}^\top\boldsymbol{w}/\tau - \frac{1}{mn(n-1)p^2} \sum_{i\in[m]} \sum_{j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_i\setminus\boldsymbol{U}_{i,j}} \boldsymbol{u}^\top\boldsymbol{v}/\tau \\
& \geq -\frac{mnp^2}{m(m-1)n^2p^2}/\tau - \frac{1}{m(m-1)n^2p^2} \sum_{i\in[m],j\neq j'\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_{i,j'}} \boldsymbol{u}^\top\boldsymbol{v}/\tau \\
& \quad - \frac{1}{mn(n-1)p^2} \sum_{i\in[m]} \sum_{j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_i\setminus\boldsymbol{U}_{i,j}} \boldsymbol{u}^\top\boldsymbol{v}/\tau \\
& = -\frac{1}{(m-1)n}/\tau - \frac{mn-1}{m(m-1)n^2(n-1)p^2} \sum_{i\in[m]} \sum_{j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_i\setminus\boldsymbol{U}_{i,j}} \boldsymbol{u}^\top\boldsymbol{v}/\tau \\
& = -\frac{1}{(m-1)n}/\tau - \frac{mn-1}{m(m-1)n^2(n-1)} \sum_{i\in[m],j\neq j'\in[n]} \mathbb{E}[\boldsymbol{U}_{i,j}]^\top \mathbb{E}[\boldsymbol{U}_{i,j'}]/\tau \\
& \geq -\frac{1}{(m-1)n}/\tau - \frac{mn-1}{m(m-1)n^2(n-1)} \cdot mn(n-1)/\tau \\
& \quad + \frac{mn-1}{2m(m-1)n^2(n-1)} \sum_{i\in[m],j\neq j'\in[n]} \big\| \mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}] \big\|_2^2/\tau \\
& = -\frac{m}{m-1}/\tau - \frac{mn-1}{2m(m-1)n^2(n-1)} \sum_{i\in[m],j\neq j'\in[n]} \big\| \mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}] \big\|_2^2/\tau,
\end{aligned}
$$

(A.41)

(A.42)

where the inequality in (A.41) comes from (A.21) in Lemma A.1 with the equality condition of

$$
\bigg\| \sum_{\boldsymbol{u}\in\boldsymbol{U}} \boldsymbol{u} \bigg\|_2^2 = 0,
\tag{A.43}
$$

and the inequality in (A.42) comes from (A.29) in Lemma A.2 with the equality condition of

$$
\big\| \mathbb{E}[\boldsymbol{U}_{i,j}] \big\|_2^2 = 1 \qquad \forall i\in[m], j\in[n].
\tag{A.44}
$$

Moreover, the value inside the second exponential in (A.40) follows

$$\frac{1}{mnp^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{w}\in\boldsymbol{U}_{i,j}} \boldsymbol{u}^\top\boldsymbol{w}/\tau - \frac{1}{mn(n-1)p^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_i\setminus\boldsymbol{U}_{i,j}} \boldsymbol{u}^\top\boldsymbol{v}/\tau$$

$$= \frac{1}{mn} \sum_{i\in[m],j\in[n]} \|\mathbb{E}[\boldsymbol{U}_{i,j}]\|_2^2/\tau - \frac{1}{mn(n-1)} \sum_{i\in[m],j\neq j'\in[n]} \mathbb{E}[\boldsymbol{U}_{i,j}]^\top\mathbb{E}[\boldsymbol{U}_{i,j'}]\tau$$

$$= \frac{1}{2mn} \sum_{i\in[m],j\in[n]} \|\mathbb{E}[\boldsymbol{U}_{i,j}]\|_2^2/\tau + \frac{1}{2mn} \sum_{i\in[m],j'\in[n]} \|\mathbb{E}[\boldsymbol{U}_{i,j'}]\|_2^2/\tau - \frac{1}{2mn(n-1)} \sum_{i\in[m],j\neq j'\in[n]} 2\mathbb{E}[\boldsymbol{U}_{i,j}]^\top\mathbb{E}[\boldsymbol{U}_{i,j'}]\tau$$

$$= \frac{1}{2mn(n-1)} \sum_{i\in[m],j\neq j'\in[n]} \left\|\mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}]\right\|_2^2/\tau.$$

Therefore, applying the above results to (A.40) yields

$$\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U}) \geq \log\left(-\frac{m}{m-1}/\tau - \frac{mn-1}{2m(m-1)n^2(n-1)} \sum_{i\in[m],j\neq j'\in[n]} \left\|\mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}]\right\|_2^2/\tau\right.$$

$$\left. + (n-1)p + p\cdot\exp\left(\frac{1}{2mn(n-1)} \sum_{i\in[m],j\neq j'\in[n]} \left\|\mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}]\right\|_2^2/\tau\right)\right). \tag{A.45}$$

On the other hand, minimizing $\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U})$ in (A.33) in a manner similar to the approach described above yields the following result:

$$\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U}) = \frac{1}{mnp^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_{i,j}} \log\left(\sum_{\boldsymbol{w}\in\boldsymbol{U}} \exp(\boldsymbol{u}^\top(\boldsymbol{w}-\boldsymbol{v})/\tau)\right)$$

$$\geq \frac{1}{mnp^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_{i,j}} \log\left((m-1)np\cdot\exp\left(\frac{1}{(m-1)np} \sum_{\boldsymbol{w}\in\boldsymbol{U}\setminus\boldsymbol{U}_i} \boldsymbol{u}^\top\boldsymbol{w}/\tau - \boldsymbol{u}^\top\boldsymbol{v}/\tau\right)\right.$$

$$+ (n-1)p\cdot\exp\left(\frac{1}{(n-1)p} \sum_{\boldsymbol{w}\in\boldsymbol{U}_i\setminus\boldsymbol{U}_{i,j}} \boldsymbol{u}^\top\boldsymbol{w}/\tau - \boldsymbol{u}^\top\boldsymbol{v}/\tau\right)$$

$$\left. + p\cdot\exp\left(\frac{1}{p} \sum_{\boldsymbol{w}\in\boldsymbol{U}_{i,j}} \boldsymbol{u}^\top\boldsymbol{w}/\tau - \boldsymbol{u}^\top\boldsymbol{v}/\tau\right)\right) \tag{A.46}$$

where the inequality in (A.46) comes from using Jensen's inequality three times. The equality in (A.46) is achieved if there exist some constants $c_4, c_5, c_6 \in \mathbb{R}$ such that the following conditions hold for all $i \in [m], j \in [n]$, $\boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{U}_{i,j}$:

$$\boldsymbol{u}^\top(\boldsymbol{w}-\boldsymbol{v}) = c_4 \qquad \forall\boldsymbol{w}\in\boldsymbol{U}\setminus\boldsymbol{U}_i, \tag{A.47}$$

$$\boldsymbol{u}^\top(\boldsymbol{w}-\boldsymbol{v}) = c_5 \qquad \forall\boldsymbol{w}\in\boldsymbol{U}_i\setminus\boldsymbol{U}_{i,j}, \tag{A.48}$$

$$\boldsymbol{u}^\top(\boldsymbol{w}-\boldsymbol{v}) = c_6 \qquad \forall\boldsymbol{w}\in\boldsymbol{U}_{i,j}. \tag{A.49}$$

From the (A.46), the following holds.

$$
\begin{aligned}
\mathcal{L}_{\text{Self}}(\boldsymbol{U}) \geq \log \Bigg(& (m-1)np \cdot \exp \Bigg(\frac{1}{m(m-1)n^2p^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{w}\in\boldsymbol{U}\setminus\boldsymbol{U}_i} \boldsymbol{u}^\top\boldsymbol{w}/\tau \\
& - \frac{1}{mnp^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_{i,j}} \boldsymbol{u}^\top\boldsymbol{v}/\tau \Bigg) \\
& + (n-1)p \cdot \exp \Bigg(\frac{1}{mn(n-1)p^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{w}\in\boldsymbol{U}_i\setminus\boldsymbol{U}_{i,j}} \boldsymbol{u}^\top\boldsymbol{w}/\tau \\
& - \frac{1}{mnp^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_{i,j}} \boldsymbol{u}^\top\boldsymbol{v}/\tau \Bigg) + p \Bigg) \quad (\text{A.50}) \\
= \log \Bigg(& (m-1)np \cdot \exp \Bigg(\frac{1}{m(m-1)n^2p^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{w}\in\boldsymbol{U}\setminus\boldsymbol{U}_i} \boldsymbol{u}^\top\boldsymbol{w}/\tau \\
& - \frac{1}{mnp^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_{i,j}} \boldsymbol{u}^\top\boldsymbol{v}/\tau \Bigg) \\
& + (n-1)p \cdot \exp \Bigg(-\frac{1}{2mn(n-1)} \sum_{i\in[m],j\neq j'\in[n]} \big\|\mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}]\big\|_2^2/\tau \Bigg) + p \Bigg) \\
\geq \log \Bigg(& (m-1)np \cdot \exp \Bigg(\frac{1}{m(m-1)n^2p^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{w}\in\boldsymbol{U}\setminus\boldsymbol{U}_i} \boldsymbol{u}^\top\boldsymbol{w}/\tau - 1/\tau \Bigg) \\
& + (n-1)p \cdot \exp \Bigg(-\frac{1}{2mn(n-1)} \sum_{i\in[m],j\neq j'\in[n]} \big\|\mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}]\big\|_2^2/\tau \Bigg) + p \Bigg) \quad (\text{A.51})
\end{aligned}
$$

where the inequality in (A.50) holds from Proposition A.5, and the equality holds if (A.48), (A.48), and (A.49) are satisfied. Moreover, the equality condition for (A.51) is equivalent to (A.44).

By using (A.21) in Lemma A.1 and (A.29) in Lemma A.2,

$$
\begin{aligned}
\mathcal{L}_{\text{Self}}(\boldsymbol{U}) \geq \log \Bigg(& (m-1)np \cdot \exp \Bigg(-\frac{(m-1)n+1}{(m-1)n}/\tau - \frac{1}{m(m-1)n^2p^2} \sum_{i\in[m],j\neq j'\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{w}\in\boldsymbol{U}_{i,j'}} \boldsymbol{u}^\top\boldsymbol{w}/\tau \Bigg) \\
& + (n-1)p \cdot \exp \Bigg(-\frac{1}{2mn(n-1)} \sum_{i\in[m],j\neq j'\in[n]} \big\|\mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}]\big\|_2^2/\tau \Bigg) + p \Bigg) \quad (\text{A.52}) \\
\geq \log \Bigg(& (m-1)np \cdot \exp \Bigg(-\frac{m}{m-1}/\tau + \frac{1}{2m(m-1)n^2} \sum_{i\in[m],j\neq j'\in[n]} \big\|\mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}]\big\|_2^2/\tau \Bigg) \\
& + (n-1)p \cdot \exp \Bigg(-\frac{1}{2mn(n-1)} \sum_{i\in[m],j\neq j'\in[n]} \big\|\mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}]\big\|_2^2/\tau \Bigg) + p \Bigg), \quad (\text{A.53})
\end{aligned}
$$

where the equality conditions of (A.52) and (A.53) are fulfilled if (A.43) and (A.44) are satisfied.

Finally, by combining the results of minimizing each loss in (A.45) and (A.53), we obtain:

$$
\mathcal{L}(\boldsymbol{U}) = (1 - \alpha)\, \mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U}) + \alpha\, \mathcal{L}_{\mathrm{Self}}(\boldsymbol{U})
$$

$$
\geq (1 - \alpha) \log \left(- \frac{m}{m-1}/\tau - \frac{mn-1}{2m(m-1)n^2(n-1)} \sum_{i \in [m], j \neq j' \in [n]} \left\| \mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}] \right\|_2^2/\tau \right.
$$

$$
\left. + (n-1)p + p \cdot \exp\left(\frac{1}{2mn(n-1)} \sum_{i \in [m], j \neq j' \in [n]} \left\| \mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}] \right\|_2^2/\tau \right) \right)
$$

$$
+ \alpha \log \left((m-1)np \cdot \exp\left(- \frac{m}{m-1}/\tau + \frac{1}{2m(m-1)n^2} \sum_{i \in [m], j \neq j' \in [n]} \left\| \mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}] \right\|_2^2/\tau \right) \right.
$$

$$
\left. + (n-1)p \cdot \exp\left(- \frac{1}{2mn(n-1)} \sum_{i \in [m], j \neq j' \in [n]} \left\| \mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}] \right\|_2^2/\tau \right) + p \right) \tag{A.54}
$$

$$
:= l \left(\sum_{i \in [m], j \neq j' \in [n]} \left\| \mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}] \right\|_2^2 \right), \tag{A.55}
$$

where the function l is defined for simple notation, as (A.54) depends on the term $\sum_{i \in [m], j \neq j' \in [n]} \left\| \mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}] \right\|_2^2$.

Note that the SSEM $\boldsymbol{U}^\delta$ in Def. A.2 satisfies all of equality conditions in (A.35), (A.36), (A.37), (A.43), (A.44), (A.47), (A.48), and (A.49). Moreover, the SSEM $\boldsymbol{U}^\delta$ in Def. A.2 can attain all possible values of $\sum_{i \in [m], j \neq j' \in [n]} \left\| \mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}] \right\|_2^2$. As a result, the equality in (A.54) holds when $\boldsymbol{U}$ is substituted by $\boldsymbol{U}^\delta$. That is,

$$
\mathcal{L}(\boldsymbol{U}) \geq l \left(\sum_{i \in [m], j \neq j' \in [n]} \left\| \mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}] \right\|_2^2 \right)
$$

$$
\geq \min_{\boldsymbol{U}} l \left(\sum_{i \in [m], j \neq j' \in [n]} \left\| \mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}] \right\|_2^2 \right)
$$

$$
= \min_{\boldsymbol{U}^\delta} l \left(\sum_{i \in [m], j \neq j' \in [n]} \left\| \mathbb{E}[\boldsymbol{U}_{i,j}] - \mathbb{E}[\boldsymbol{U}_{i,j'}] \right\|_2^2 \right)
$$

$$
= \min_{\boldsymbol{U}^\delta} \mathcal{L}(\boldsymbol{U}^\delta)
$$

$$
= \min_{\delta} \mathcal{L}(\boldsymbol{U}^\delta).
$$

Now, we only need to determine $\delta^\star$ of SSEM in Def. A.2 such that minimize $\mathcal{L}(\boldsymbol{U}^{\delta^\star})$ in (1). Note that the denominator in each logarithm in $\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U}^\delta)$ and $\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U}^\delta)$ has the same value for a given δ, as follows. For any $\boldsymbol{u}^\delta \in \boldsymbol{U}^\delta$, by using (A.2)-(A.4) in Def. A.2,

$$
\sum_{\boldsymbol{w}^\delta \in \boldsymbol{U}^\delta} \exp\left((\boldsymbol{u}^\delta)^\top \boldsymbol{w}^\delta/\tau \right) = p \cdot \exp(1/\tau) + (n-1)p \cdot \exp\left(\left(1 - \delta^2 \frac{mn}{mn-1}\right)/\tau \right)
$$

$$
+ (m-1)np \cdot \exp\left(\left(-\frac{1}{m-1} + \delta^2 \frac{m(n-1)}{(m-1)(mn-1)} \right)/\tau \right)
$$

$$
:= g(\delta; m, n, p, \tau), \tag{A.56}
$$

where $g(\delta; m, n, p, \tau)$ in (A.56) is defined for the simple notation.

Then, the losses $\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U}^\delta)$ and $\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U}^\delta)$ are simplified as follows.

$$
\begin{aligned}
\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U}^\delta) &= -\frac{1}{mn(n-1)p^2} \sum_{i\in[m],j\neq j'\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_{i,j'}} \log \frac{\exp(\boldsymbol{u}^\top\boldsymbol{v}/\tau)}{\sum_{\boldsymbol{w}\in\boldsymbol{U}} \exp(\boldsymbol{u}^\top\boldsymbol{w}/\tau)} \\
&= -\frac{1}{mn(n-1)p^2} \sum_{i\in[m],j\neq j'\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_{i,j'}} \log \frac{\exp\left(\left(1-\delta^2\frac{mn}{mn-1}\right)/\tau\right)}{g(\delta;m,n,p,\tau)} \\
&= \log \frac{g(\delta;m,n,p,\tau)}{\exp\left(\left(1-\delta^2\frac{mn}{mn-1}\right)/\tau\right)}, \\
\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U}^\delta) &= -\frac{1}{mnp^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_{i,j}} \log \frac{\exp(\boldsymbol{u}^\top\boldsymbol{v}/\tau)}{\sum_{\boldsymbol{w}\in\boldsymbol{U}} \exp(\boldsymbol{u}^\top\boldsymbol{w}/\tau)} \\
&= -\frac{1}{mnp^2} \sum_{i\in[m],j\in[n]} \sum_{\boldsymbol{u}\in\boldsymbol{U}_{i,j}} \sum_{\boldsymbol{v}\in\boldsymbol{U}_{i,j}} \log \frac{\exp(1/\tau)}{g(\delta;m,n,p,\tau)} \\
&= \log \frac{g(\delta;m,n,p,\tau)}{\exp(1/\tau)}.
\end{aligned}
$$

As a result, $\mathcal{L}(\boldsymbol{U}^\delta)$ is rewritten as

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{U}^\delta) &= (1-\alpha)\,\mathcal{L}_{\mathrm{Sup}}(\boldsymbol{U}^\delta) + \alpha\,\mathcal{L}_{\mathrm{Self}}(\boldsymbol{U}^\delta) \\
&= (1-\alpha)\cdot\log \frac{g(\delta;m,n,p,\tau)}{\exp\left(\left(1-\delta^2\frac{mn}{mn-1}\right)/\tau\right)} + \alpha\cdot\log \frac{g(\delta;m,n,p,\tau)}{\exp(1/\tau)} \\
&= \log g(\delta;m,n,p,\tau) - (1-\alpha)\left(1-\delta^2\frac{mn}{mn-1}\right)/\tau - \alpha/\tau \\
&= \log\left(p\cdot\exp(1/\tau) + (n-1)p\cdot\exp\left(\left(1-\delta^2\frac{mn}{mn-1}\right)/\tau\right)\right. \\
&\qquad \left. + (m-1)np\cdot\exp\left(\left(-\frac{1}{m-1}+\delta^2\frac{m(n-1)}{(m-1)(mn-1)}\right)/\tau\right)\right) \\
&\quad - (1-\alpha)\cdot\left(1-\delta^2\frac{mn}{mn-1}\right)/\tau - \alpha\cdot(1/\tau) \\
&= \log\left(1+(n-1)\cdot\exp\left(-\tilde{\delta}/\tau\right)+(m-1)n\cdot\exp\left(\left(-\frac{m}{m-1}+\tilde{\delta}\frac{n-1}{(m-1)n}\right)/\tau\right)\right) \\
&\quad + \log\left(p\cdot\exp(1/\tau)\right)-(1-\alpha)(1-\tilde{\delta})/\tau-\alpha/\tau, \qquad\qquad\qquad\qquad\qquad\text{(A.57)} \\
&= \log\left(1+(n-1)\cdot\exp\left(-\tilde{\delta}/\tau\right)+(m-1)n\cdot\exp\left(\left(-\frac{m}{m-1}+\tilde{\delta}\frac{n-1}{(m-1)n}\right)/\tau\right)\right)+\log p+(1-\alpha)\tilde{\delta}/\tau,
\end{aligned}
$$

where $\tilde{\delta} := \delta^2\frac{mn}{mn-1} \in \left[0,\frac{n}{n-1}\right]$ in (A.57) is the monotonic increasing transformation of $\delta \in \left[0,\sqrt{\frac{mn-1}{m(n-1)}}\right]$.

To find $\tilde{\delta}^\star$ that minimize $\mathcal{L}(\boldsymbol{U}^{\tilde{\delta}})$,

$$
\begin{aligned}
\frac{\partial}{\partial\tilde{\delta}}\mathcal{L}(\boldsymbol{U}^{\tilde{\delta}}) &= \frac{-\frac{n-1}{\tau}\cdot\exp\left(-\tilde{\delta}/\tau\right)+\frac{n-1}{\tau}\cdot\exp\left(\left(-\frac{m}{m-1}+\tilde{\delta}\frac{n-1}{(m-1)n}\right)/\tau\right)}{1+(n-1)\cdot\exp\left(-\tilde{\delta}/\tau\right)+(m-1)n\cdot\exp\left(\left(-\frac{m}{m-1}+\tilde{\delta}\frac{n-1}{(m-1)n}\right)/\tau\right)}+(1-\alpha)/\tau \\
&= \frac{1}{\tau}\cdot\frac{-(n-1)\cdot\exp\left(-\tilde{\delta}/\tau\right)+(n-1)\cdot\exp\left(\left(-\frac{m}{m-1}+\tilde{\delta}\frac{n-1}{(m-1)n}\right)/\tau\right)}{1+(n-1)\cdot\exp\left(-\tilde{\delta}/\tau\right)+(m-1)n\cdot\exp\left(\left(-\frac{m}{m-1}+\tilde{\delta}\frac{n-1}{(m-1)n}\right)/\tau\right)}+\frac{1}{\tau}\cdot(1-\alpha) \\
&= \frac{1}{\tau}\cdot\frac{h(\tilde{\delta};m,n,\tau,\alpha)}{1+(n-1)\cdot\exp(-\tilde{\delta}/\tau)+(m-1)n\cdot\exp\left(\left(-\frac{m}{m-1}+\tilde{\delta}\frac{n-1}{(m-1)n}\right)/\tau\right)}, \qquad\qquad\text{(A.58)}
\end{aligned}
$$

where we define the function h in (A.58) as

$$h(\tilde{\delta}; m, n, \tau, \alpha) = (1 - \alpha) - \alpha(n - 1) \cdot \exp(-\tilde{\delta}/\tau)$$
$$+ (mn - 1 - \alpha(m - 1)n) \cdot \exp\left(\left(-\frac{m}{m-1} + \tilde{\delta}\frac{n-1}{(m-1)n}\right)/\tau\right).$$

Note that the denominator in (A.58) is always positive. Moreover, the function $h(\tilde{\delta}; m, n, \tau, \alpha)$ in (A.58) is monotonically increasing with respect to $\tilde{\delta} \in \left[0, \frac{n}{n-1}\right]$, and the following value is non-negative.

$$h\left(\frac{mn}{mn-1}; m, n, \tau, \alpha\right) = (1 - \alpha) - \alpha(n - 1) \cdot \exp\left(-\frac{mn}{mn-1}/\tau\right)$$
$$+ (mn - 1 - \alpha(m - 1)n) \cdot \exp\left(\left(-\frac{m}{m-1} + \frac{m(n-1)}{(m-1)(mn-1)}\right)/\tau\right)$$
$$= (1 - \alpha) - \alpha(n - 1) \cdot \exp\left(-\frac{mn}{mn-1}/\tau\right) + (mn - 1 - \alpha(m - 1)n) \cdot \exp\left(-\frac{mn}{mn-1}/\tau\right)$$
$$= (1 - \alpha) + (mn - 1)(1 - \alpha) \cdot \exp\left(-\frac{mn}{mn-1}/\tau\right)$$
$$\geq 0.$$

Therefore, $\tilde{\delta}^{\star}$, which minimizes $\mathcal{L}(\boldsymbol{U}^{\tilde{\delta}})$, can be determined as follows:

$$\tilde{\delta}^{\star} = \begin{cases} 0, & \text{if } h(0; m, n, \tau, \alpha) \geq 0, \\ \tilde{\delta} \in \left(0, \frac{mn}{mn-1}\right) \text{ such that } h(\tilde{\delta}; m, n, \tau, \alpha) = 0, & \text{otherwise.} \end{cases}$$

This can be rewritten as:

$$\delta^{\star} = \begin{cases} 0, & \text{if } h(0; m, n, \tau, \alpha) \geq 0, \\ \delta \in (0, 1], \text{ such that } h\left(\delta^2\frac{mn}{mn-1}; m, n, \tau, \alpha\right) = 0, & \text{otherwise.} \end{cases}$$

As a result, all embedding sets $\boldsymbol{U}^{\star}$ that minimize the loss $\mathcal{L}(\boldsymbol{U})$ in (1) are included in the SSEM as follows:

$$\forall \boldsymbol{U}^{\star} \in \arg\min_{\boldsymbol{U}} \mathcal{L}(\boldsymbol{U}), \quad \exists! \delta \in [0, 1] \text{ such that } \boldsymbol{U}^{\delta} = \boldsymbol{U}^{\star}.$$

where the uniqueness of δ arises from the monotonicity of h.

$\square$

A.4 Preventing Class Collapse

Theorem A.3. *Let $\boldsymbol{U}^\star$ be the set of optimal embedding vectors that minimizes the loss $\mathcal{L}(\boldsymbol{U})$ in (1). Then, the class collapse does not happen, i.e., $\mathrm{Var}[\boldsymbol{U}_i^\star] > 0$ for all $i \in [m]$, if and only if the loss-combining coefficient α satisfies*

$$\alpha \in \left(\frac{mn - 1 + \exp\left(\frac{m}{m-1}/\tau\right)}{mn - n + n \cdot \exp\left(\frac{m}{m-1}/\tau\right)}, 1 \right]$$

for a given temperature $\tau > 0$. This necessary and sufficient condition for preventing class-collapse can be re-written as

$$\tau \in \left(0, \frac{1}{\left(1 - \frac{1}{m}\right) \cdot \log\left(\frac{mn-1-\alpha(m-1)n}{\alpha n - 1}\right)} \right)$$

for a given $\alpha \in \left(\frac{1}{n}, 1 \right]$.

Proof. To find the condition for preventing class-collapse. Proposition A.4 implies that $\delta^\star$ of SSEM must be positive. Therefore, from (A.31) in Theorem A.2, the necessity and sufficient condition of preventing class-collapse is $h\big(0; m, n, \tau\big) < 0$ where

$$h\big(x; m, n, \tau, \alpha\big) = (1 - \alpha) - \alpha(n-1) \cdot \exp(-x/\tau)$$
$$+ (mn - 1 - \alpha(m-1)n) \cdot \exp\left(\left(-\frac{m}{m-1} + x\frac{n-1}{(m-1)n} \right) /\tau \right).$$

This condition can be rewritten as follows:

$$0 > h\big(0; m, n, \tau, \alpha\big)$$
$$= (1 - \alpha) - \alpha(n-1) \cdot \exp(0) + (mn - 1 - \alpha(m-1)n) \cdot \exp\left(-\frac{m}{m-1}/\tau \right)$$
$$= 1 - \alpha n + (mn - 1 - \alpha(m-1)n) \cdot \exp\left(-\frac{m}{m-1}/\tau \right) \qquad \text{(A.59)}$$
$$= -\alpha n \left(1 + (m-1) \cdot \exp\left(-\frac{m}{m-1}/\tau \right) \right) + 1 + (mn - 1) \cdot \exp\left(-\frac{m}{m-1}/\tau \right),$$

which is equal to

$$\alpha > \frac{1 + (mn-1) \cdot \exp\left(-\frac{m}{m-1}/\tau \right)}{n\left(1 + (m-1) \cdot \exp\left(-\frac{m}{m-1}/\tau \right) \right)} = \frac{mn - 1 + \exp\left(\frac{m}{m-1}/\tau\right)}{mn - n + n \cdot \exp\left(\frac{m}{m-1}/\tau\right)}.$$

Or equivalently, from (A.59),

$$\alpha n - 1 > (mn - 1 - \alpha(m-1)n) \cdot \exp\left(-\frac{m}{m-1}/\tau \right). \qquad \text{(A.60)}$$

Note that

$$\frac{mn-1}{mn-n} \geq \frac{mn-1}{mn-1} = 1 \geq \alpha,$$

which implies $mn - 1 - \alpha(m-1)n \geq 0$. Moreover,

$$\frac{\alpha n - 1}{mn - 1 - \alpha(m-1)n} = \frac{\alpha n - 1}{\alpha n - 1 + mn(1 - \alpha)} \leq \frac{\alpha n - 1}{\alpha n - 1} = 1.$$

Then the following conditions are equivalent to (A.60):

$$\frac{\alpha n - 1}{mn - 1 - \alpha(m-1)n} > \exp\left(-\frac{m}{m-1}/\tau \right),$$

$$\log\left(\frac{\alpha n - 1}{mn - 1 - \alpha(m-1)n}\right) > -\frac{m}{m-1}/\tau,$$

$$\tau < \frac{m}{(m-1)\cdot\log\left(\frac{mn-1-\alpha(m-1)n}{\alpha n-1}\right)} = \left(0, \frac{1}{(1-\frac{1}{m})\cdot\log\left(\frac{mn-1-\alpha(m-1)n}{\alpha n-1}\right)}\right).$$

Note that the minimum range of α is $\frac{1}{n}$, which is comes from

$$\frac{1}{n} = \lim_{\tau\to 0}\frac{mn - 1 + \exp\left(\frac{m}{m-1}/\tau\right)}{mn - n + n\cdot\exp\left(\frac{m}{m-1}/\tau\right)}.$$

□

A.5 The Optimality of Class-Conditional InfoNCE Loss

Proposition A.6. *Let the class-conditional InfoNCE loss $\mathcal{L}_{\mathrm{cNCE}}(U)$ be defined as*

$$\mathcal{L}_{\mathrm{cNCE}}(U) = -\frac{1}{mnp^2}\sum_{i\in[m],j\in[n]}\sum_{u\in U_{i,j}}\sum_{v\in U_{i,j}}\log\frac{\exp(u^\top v/\tau)}{\sum_{w\in U_i}\exp(u^\top w/\tau)}.$$

Suppose the embedding dimension satisfies $d \geq mn-1$ for given $m, n \in \mathbb{N}$ with $mn \geq 2$. Then, for any embedding set U,

$$\mathcal{L}_{\mathrm{cNCE}}(U) \geq \mathcal{L}_{\mathrm{cNCE}}(U^\delta)$$

holds for $\delta = \sqrt{\frac{mn-1}{m(n-1)}}$.

Proof. We minimize $\mathcal{L}_{\mathrm{cNCE}}(U)$ using a similar approach as in the proofs of Theorem A.2.

$$\mathcal{L}_{\mathrm{cNCE}}(U) = \frac{1}{mnp^2}\sum_{i\in[m],j\in[n]}\sum_{u\in U_{i,j}}\sum_{v\in U_{i,j}}\log\left(\sum_{w\in U_i}\exp(u^\top(w-v)/\tau)\right)$$

$$\geq \frac{1}{mnp^2}\sum_{i\in[m],j\in[n]}\sum_{u\in U_{i,j}}\sum_{v\in U_{i,j}}\log\left((n-1)p\cdot\exp\left(\frac{1}{(n-1)p}\sum_{w\in U_i\setminus U_{i,j}}u^\top w/\tau - u^\top v/\tau\right)\right.$$

$$\left. + p\cdot\exp\left(\frac{1}{p}\sum_{w\in U_{i,j}}u^\top w/\tau - u^\top v/\tau\right)\right) \tag{A.61}$$

where the inequality in (A.61) comes from using Jensen's inequality two times. The equality in (A.61) is achieved if there exist some constants $c_1, c_2 \in \mathbb{R}$ such that the following conditions hold for all $i \in [m], j \in [n], u, v \in U_{i,j}$:

$$u^\top(w - v) = c_1 \qquad \forall w \in U_i \setminus U_{i,j}, \tag{A.62}$$

$$u^\top(w - v) = c_2 \qquad \forall w \in U_{i,j}. \tag{A.63}$$

Then, the following holds,

$$\mathcal{L}_{\mathrm{cNCE}}(U) \geq \log\left((n-1)p\exp\left(\frac{1}{mn(n-1)p^2}\sum_{i\in[m],j\in[n]}\sum_{u\in U_{i,j}}\sum_{w\in U_i\setminus U_{i,j}}u^\top w/\tau\right.\right.$$

$$\left.\left. -\frac{1}{mnp^2}\sum_{i\in[m],j\in[n]}\sum_{u\in U_{i,j}}\sum_{v\in U_{i,j}}u^\top v/\tau\right) + p\right) \tag{A.64}$$

$$= \log\left((n-1)p\exp\left(-\frac{1}{2mn(n-1)}\sum_{i\in[m],j\neq j'\in[n]}\left\|\mathbb{E}[U_{i,j}] - \mathbb{E}[U_{i,j'}]\right\|_2^2/\tau\right) + p\right)$$

where the inequality in (A.64) holds from Proposition A.5, and the equality holds if (A.62) and (A.63) are satisfied.

Note that the SSEM $\boldsymbol{U}^\delta$ in Def. A.2 satisfies all of equality conditions in (A.62) and (A.63). Moreover, the SSEM $\boldsymbol{U}^\delta$ in Def. A.2 can attain all possible values of $\sum_{i\in[m],j\neq j'\in[n]}\left\|\mathbb{E}[\boldsymbol{U}_{i,j}]-\mathbb{E}[\boldsymbol{U}_{i,j'}]\right\|_2^2$. As a result, we only need to find $\delta\in\left[0,\sqrt{\frac{mn-1}{m(n-1)}}\right]$ that minimizes the loss, as below.

$$\mathcal{L}_{\mathrm{cNCE}}(\boldsymbol{U})\geq\min_\delta\mathcal{L}_{\mathrm{cNCE}}(\boldsymbol{U}^\delta).$$

Since the SSEM $\boldsymbol{U}^\delta$ satisfy the equality conditions of (A.61) and (A.64), $\mathcal{L}_{\mathrm{cNCE}}(\boldsymbol{U}^\delta)$ is rewritten as follows:

$$\mathcal{L}_{\mathrm{cNCE}}(\boldsymbol{U}^\delta)=\frac{1}{mnp^2}\sum_{i\in[m],j\in[n]}\sum_{\boldsymbol{u}^\delta\in\boldsymbol{U}_{i,j}^\delta}\sum_{\boldsymbol{v}^\delta\in\boldsymbol{U}_{i,j}^\delta}\log\left(\sum_{\boldsymbol{w}^\delta\in\boldsymbol{U}_i^\delta}\exp\left(\left(\boldsymbol{u}^\delta\right)^\top\left(\boldsymbol{w}^\delta-\boldsymbol{v}^\delta\right)/\tau\right)\right)$$

$$=\log\left((n-1)p\cdot\exp\left(\frac{1}{mn(n-1)p^2}\sum_{i\in[m],j\in[n]}\sum_{\boldsymbol{u}^\delta\in\boldsymbol{U}_{i,j}^\delta}\sum_{\boldsymbol{w}^\delta\in\boldsymbol{U}_i^\delta\setminus\boldsymbol{U}_{i,j}^\delta}(\boldsymbol{u}^\delta)^\top\boldsymbol{w}^\delta/\tau\right.\right.$$

$$\left.\left.-\frac{1}{mnp^2}\sum_{i\in[m],j\in[n]}\sum_{\boldsymbol{u}^\delta\in\boldsymbol{U}_{i,j}^\delta}\sum_{\boldsymbol{v}^\delta\in\boldsymbol{U}_{i,j}^\delta}(\boldsymbol{u}^\delta)^\top\boldsymbol{v}^\delta/\tau\right)+p\right)$$

$$=\log\left((n-1)p\cdot\exp\left(\left(1-\delta^2\frac{mn}{mn-1}\right)/\tau-1/\tau\right)+p\right)$$

$$=\log\left((n-1)p\cdot\exp\left(-\delta^2\frac{mn}{mn-1}/\tau\right)+p\right),$$

which is a monotonic decreasing function of δ. Since δ lies within the range of $\left[0,\sqrt{\frac{mn-1}{m(n-1)}}\right]$, the minimum loss is attained when $\delta=\frac{mn-1}{m(n-1)}$. Therefore, for any embedding set $\boldsymbol{U}$,

$$\mathcal{L}_{\mathrm{cNCE}}(\boldsymbol{U})\geq\mathcal{L}_{\mathrm{cNCE}}(\boldsymbol{U}^\delta)$$

holds for $\delta=\sqrt{\frac{mn-1}{m(n-1)}}$.

□

B   ADDITIONAL EXPERIMENTS ON SYNTHETIC DATA

In this section, we present additional experiments on synthetic data. We follow the training setup in Sec. 6.1, except for using the lower embedding dimension where $d = 50$.

Note that Theorem 1 assumes $d \geq mn - 1$, while these additional experiments do not strictly satisfy this condition, as $50 = d < mn - 1 = 99$. This suggests that even when the optimal embedding set of SSEM cannot exist in a lower-dimensional embedding space by Proposition 1, the average within-class variance of the learned embedding set still aligns with our theoretical analysis as shown in Fig. B.1.
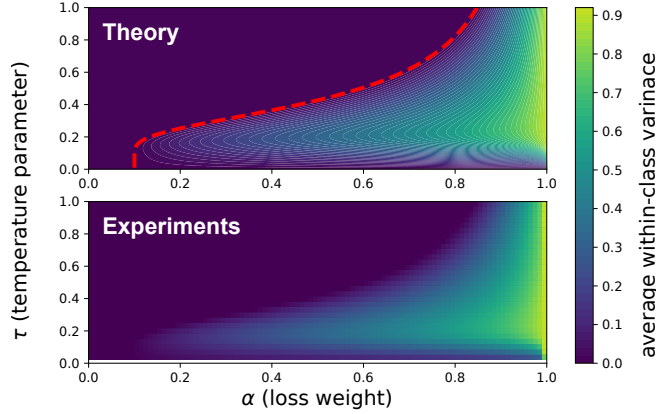

Figure B.1:   The within-class variance (averaged over different classes) of the learned embedding set U, for various loss-combining coefficient α and temperature τ. (Top): Computed from theoretical results in Sec. 5, (Bottom): Computed from the experiments on synthetic datasets in Appendix B.

C   EXPERIMENTS ON REAL DATA

C.1   Details of Datasets and Augmentation

We use the CIFAR-10 and ImageNet-100 datasets (Krizhevsky et al., 2009; Deng et al., 2009) for training the model. For CIFAR-10, we use the balanced mini-batch where the number of instances per class is equal. For ImageNet-100, we select a subset of 100 classes.

For data augmentation strategy, we follow prior works (Chen et al., 2020; He et al., 2020), including random cropping, color jitter, random grayscale, Gaussian blur, and random horizontal flipping.

To evaluate the performance of transfer learning, we use various downstream datasets, as described in Table C.1.

Table C.1: Real datasets used in experiments.

Name	# of classes	Training size	Validation size	Test size	Evaluation metric
CIFAR10 (Krizhevsky et al., 2009)	10	45000	5000	10000	Top-1 accuracy
CIFAR100 (Krizhevsky et al., 2009)	100	45000	5000	10000	Top-1 accuracy
ImageNet100 (Russakovsky et al., 2015)	1000	126689	-	-	-
MIT67 (Quattoni and Torralba, 2009)	67	4690	670	1340	Top-1 accuracy
DTD (Cimpoi et al., 2014)	47	1880	1880	1880	Top-1 accuracy
Food (Bossard et al., 2014)	101	68175	7575	25250	Top-1 accuracy
SUN397 (Xiao et al., 2010)	397	15880	3970	19850	Top-1 accuracy
Caltech101 (Fei-Fei et al., 2004)	101	2525	505	5647	Mean per-class accuracy
CUB200 (Welinder et al., 2011)	200	4990	1000	5794	Mean per-class accuracy
Dogs (Khosla et al., 2011; Deng et al., 2009)	120	10800	1200	8580	Mean per-class accuracy
Flowers (Nilsback and Zisserman, 2008)	102	1020	1020	6149	Mean per-class accuracy
Pets (Parkhi et al., 2012)	37	2940	740	3669	Mean per-class accuracy

C.2 Details of Architecture and Training

For training on the CIFAR-10 dataset, we use the modified ResNet-18 encoder (He et al., 2016; Chen et al., 2020) followed by the 2-layer MLP projector. Specifically, we replace the first convolutional layer with a 3x3 convolution at a stride of 1, removing the initial max pooling operation. For ImageNet-100, we use the ResNet-50 encoder, also followed by the 2-layer MLP projector.

The loss configurations for each dataset are summarized in Table C.2. When training on the CIFAR-10 dataset, we first use the loss in (1) where α was fixed at 0.5 and τ ranging from 0.05 to 1.00. Next, we use the loss in (1) where τ was fixed at 0.1 and α ranging from 0.0 to 1.0. This entire training process using different loss hyperparameters is repeated for batch sizes of 100, 500, and 2000.

The models are trained utilizing a single NVIDIA RTX 4090 GPU (for CIFAR-10) or two NVIDIA RTX A5000 GPUs (for ImageNet-100), employing the SGD optimizer with a learning rate of 0.05, a momentum of 0.9, and a weight decay of 1e-4. We apply the cosine learning rate schedule (Loshchilov and Hutter, 2017). Training runs for 1,000 epochs on CIFAR-10 and 200 epochs on ImageNet-100.

Table C.2: Training configuration.

Training dataset	Batch size	α (loss-combining coefficient)	τ (temperature parameter)
CIFAR-10	100, 500, 2000	0.5 0.0, 0.1, 0.2, $\cdots$, 1.0	0.05 0.10 0.15, $\cdots$, 1.00 0.1
ImageNet-100	256	0.0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0	0.1

C.3 Details of Linear Probing Evaluation

To evaluate the average within-class variance of the learned embeddings, we begin by extracting the projector outputs of the training data, without applying any augmentations. These output vectors are subsequently normalized, after which the average within-class variance is computed across all classes.

Subsequently, we remove the projector head and evaluate the performance of the pretrained encoder on various downstream classification tasks by linear probing. Following prior works (Kornblith et al., 2019; Lee et al., 2021; Oh and Lee, 2024), we evaluate the top-1 accuracy or mean per-class accuracy depending on the downstream dataset, as shown in Table C.1. To be specific, we train the linear classifier by minimizing the L2-regularized cross-entropy loss using limited-memory BFGS (Liu and Nocedal, 1989). The best-performing classifier on the training data is subsequently used to predict the test data, followed by an evaluation of accuracy.

C.4 Additional Experiments for Evaluating Transfer Learning Performance

We use the same ResNet-50 encoders from the linear probing evaluations in Table C.1, pretrained on ImageNet-100 with the SupCL loss $\mathcal{L}(\mathbf{U})$ in (1), using $\tau = 0.1$ and α ranging from 0 to 1. We then evaluate transfer learning performance on object detection and few-shot learning tasks, following related works (He et al., 2020; Oh and Lee, 2024).

Table C.3: Transfer learning performance (%) on VOC object detection task, using the metric of COCO-style AP on the VOC07 test dataset.

α	AP
0.0	53.21
0.2	**53.27**
0.5	53.09
0.8	51.47
1.0	50.72

Table C.3 presents the results for the object detection task. We follow the experimental setup of He et al.

(2020), initializing a Faster R-CNN model with a ResNet-50 pre-trained on ImageNet-100 and fine-tuning it on the VOC07+12 training dataset (Everingham et al., 2010). Performance is evaluated using the metric of COCO-style average precision (AP) (Lin et al., 2014) on the VOC07 test dataset. As shown in Table C.3, the best results are obtained when the loss-combining coefficient α is 0.2, *i.e.,* when the pre-trained embeddings have a moderate amount of within-class variance.

Table C.4: Few-shot classification accuracy (%) evaluated across various downstream datasets.

	5-way 1-shot						5-way 5-shot					
α	Avg. accuracy	Aircraft	CUB200	FC100	Flowers102	DTD	Avg. accuracy	Aircraft	CUB200	FC100	Flowers102	DTD
0.0	51.98	31.78	49.09	43.69	78.47	56.88	68.48	45.18	66.30	62.81	93.45	74.66
0.2	51.53	30.85	47.90	45.88	75.65	57.35	67.95	44.19	64.21	65.12	91.61	74.61
0.5	**52.33**	31.23	49.73	46.09	76.90	57.72	**68.99**	44.21	66.39	65.24	92.76	76.35
0.8	49.84	30.42	45.08	41.03	75.00	57.66	66.54	43.40	61.11	60.60	91.52	76.07
1.0	46.16	29.67	40.96	35.21	69.49	55.45	61.82	39.96	55.61	51.64	88.04	73.87

Table C.4 shows empirical results for few-shot learning tasks. Following the linear probing protocol for few-shot learning (Lee et al., 2021; Oh and Lee, 2024), we first extract representations of 224×224 images (without data augmentation) from the pre-trained model and train a classification head. We then evaluate the accuracy of 5-way 1-shot and 5-way 5-shot scenarios over 2000 episodes across five downstream datasets: Aircraft (Maji et al., 2013), CUB200 (Welinder et al., 2011), FC100 (Oreshkin et al., 2018), Flowers (Nilsback and Zisserman, 2008), and DTD (Cimpoi et al., 2014). As shown in Table C.4, the model achieve optimal performance at $\alpha = 0.5$, again corresponding to a moderate level of within-class variance.

These additional evaluations demonstrate that embeddings with a moderate amount of within-class variance achieve better performance across diverse transfer learning tasks.

C.5 Comparisons with Related Methods

The SupCL loss $\mathcal{L}(\boldsymbol{U})$ in (1), which we analyze, is formulated as a convex combination of the supervised contrastive loss and the self-supervised contrastive loss. We focus on this loss because it outperforms other existing CL losses (Islam et al., 2021; Oh and Lee, 2024). To further demonstrate its effectiveness, we conducted additional experiments comparing the SupCL loss $\mathcal{L}(\boldsymbol{U})$ in (1) with existing CL methods:

- SimCLR (Chen et al., 2020): A widely adopted self-supervised CL method that does not utilize supervision.

- Vanilla SupCL (Khosla et al., 2020): The first method that integrates supervision into self-supervised CL.

- L_{spread} (Chen et al., 2022): A variant of the SupCL loss designed to mitigate class collapse.

These methods are particularly relevant to our work, as our analysis focuses on a loss that is a convex combination of the supervised contrastive loss and the self-supervised contrastive loss. We also included L_{spread} (Chen et al., 2022) in our comparison, since it was developed to address the class collapse problem.

For the comparison, we trained ResNet-50 on ImageNet-100 using a temperature of 0.1, a learning rate of 0.3, and a batch size of 256. The loss-combining weight α was set to 0.5 for both our method and L_{spread}. For consistency, the temperature parameter τ was fixed to 0.1 across all methods.

We evaluate transfer learning performance following Appendix C.3. As shown in Table C.5, using the SupCL loss $\mathcal{L}(\boldsymbol{U})$ in (1) outperforms other approaches, highlighting its superior effectiveness.

Table C.5: Classification accuracy (%) evaluated on various downstream datasets.

Method	Avg. accuracy	CIFAR10	CIFAR100	Caltech101	CUB200	Dog	DTD	Flowers102	Food101	MIT67	Pets	SUN397
SimCLR	63.06	84.15	63.17	78.64	30.40	46.19	65.00	85.71	62.02	62.91	67.60	47.89
Vanilla SupCL	68.37	88.64	69.20	87.18	35.71	62.47	66.54	88.95	58.86	63.36	80.51	50.65
L_{spread}	68.84	89.61	69.95	87.55	37.59	62.21	65.90	89.05	60.98	63.88	79.22	51.25
SupCL in (1)	**69.06**	89.43	69.45	88.35	38.48	62.78	66.33	89.49	60.36	63.43	80.68	50.88