Optimal Representations for Generalized Contrastive Learning with Imbalanced Datasets

Anonymous authors
Paper under double-blind review

Abstract

In this paper, we provide a computable characterization of the geometry of optimal representations in Contrastive Learning (CL) when the classes are imbalanced. When classes are balanced and the representation dimension is greater than the number of classes, it is well-known that the optimal representations exhibit Neural Collapse (NC), i.e., representations from the same class collapse to their class means and the class means form an Equiangular Tight Frame (ETF). For imbalanced classes and a large, generalized family of CL losses, we prove that the optimal representations of all samples from the same class collapse to their class means and their geometry exhibits an angular symmetry structure that is determined by the relative class proportions. In general, we show that the geometry can be determined by solving a convex optimization problem. Exploiting this symmetry structure, we analytically investigate a special case where class imbalance is extreme and prove that CL exhibits a phenomenon called Minority Collapse (MC) where all samples from the minority classes (classes with small probabilities) collapse into a single vector, whenever the class imbalance exceeds a threshold, which in turn depends on the regularity properties of the CL loss used and on the number of negative samples. Numerical results are provided to illustrate these phenomena and corroborate the theoretical results. We conclude by identifying a number of open problems.

Keywords: Contrastive Learning, Optimal Representation Geometry, Class Imbalance, Neural Collapse, Minority Collapse

1 Introduction

CL is a machine learning technique that aims to learn a representation map by pulling "similar" samples closer together while simultaneously pushing apart "different" samples in the representation space. These representations can then be directly utilized or fine-tuned for downstream tasks. Over the past decade, CL has received significant attention due to its applications ranging from computer vision, time series analysis, and natural language processing (see Jaiswal et al. (2020) for a comprehensive survey).

In CL terminology, a reference sample is called the "anchor" sample, a sample similar to it is called the "positive" sample, and a sample different from it is called the "negative" sample. If label information is not available (unsupervised setting), positive samples are usually constructed via data augmentations of the anchor, and negative samples are randomly selected from the dataset Chen et al. (2020). When label information is available (supervised setting), positive samples can be selected from the same class as the anchor while negative samples can be picked from either (a) classes other than the anchor's class Jiang et al. (2024b;a), or (b) any class (including the anchor's class) Khosla et al. (2020). Under a suitable model of the data generating the positive and negative samples in the unsupervised as well as supervised settings, the aim of this paper is to characterize the optimal representations learned via CL under an *unconstrained features model* wherein the CL map is assumed to have adequate capacity to realize any mapping. This is an important problem that sheds light on the effect of positive and negative sampling mechanisms in CL. In the next section, we will begin by reviewing related work and outline our main contributions in that context.

1.1 Limitations of related work and contributions

Loss function, sampling distribution, and number of negative samples per positive-pair k: To the best of our knowledge, most theoretical studies of CL that have aimed to understand the structure of optimum representations Fang et al. (2021); Graf et al. (2021); Kothapalli (2023); Kini et al. (2024); Behnia & Thrampoulidis (2024) have done so **only for *empirical* versions of the InfoNCE CL loss (or its variants) with norm-bounded representation constraints** where within each mini-batch b, consisting of n_b of samples, the anchor is ***uniformly*** distributed over ***all*** n_b ***samples***, the positive sample is ***uniformly*** distributed over ***all*** n_b ***samples*** (some works exclude the anchor), and for each anchor-positive pair, ***all*** n_b ***samples*** (some works exclude the anchor or/and the positive sample) are negative samples (i.e., $k = n_b$ or $n_b - 1$ or $n_b - 2$). Unraveling the impact of k is not possible with the approaches taken in extant works since they only consider empirical CL losses where k is nearly equal to the batch size.

Class proportions: In addition to heavily focusing on the empirical InfoNCE loss together with the (nearly) maximum possible range of k, almost all prior theoretical works in CL Fang et al. (2021); Graf et al. (2021); Kothapalli (2023); Behnia & Thrampoulidis (2024) have focused on the ***idealized balanced setting*** in which each sample belongs to one of $C > 1$ classes (or latent classes) and ***all classes are equally likely,*** i.e., have the same sample size in the training set. The more realistic and practically useful ***unbalanced setting*** has been analyzed primarily for ***classifier networks*** with the empirical Mean Squared Error (MSE) loss Dang et al. (2023) and empirical cross-entropy loss Hong & Ling (2024); Dang et al. (2024b) where there is an additional linear classifier layer following the representation mapping and the loss function explicitly depends on the labels of the samples. Analysis of the unbalanced case for CL is very limited and confined to the empirical InfoNCE loss Fang et al. (2021); Kini et al. (2024); Behnia & Thrampoulidis (2024).

Minority-Collapse (MC) phenomenon: When classes are not balanced, the representations of all the samples in several distinct *minority* classes (classes with small probabilities) may collapse into a single vector. This phenomenon has been studied only fairly recently, primarily within the context of *classifier* networks with either empirical MSE loss Dang et al. (2023) or empirical cross-entropy loss Hong & Ling (2024); Dang et al. (2024b). Within the CL context, the *existence* of minority-collapse was proved in Fang et al. (2021) *only in the asymptotic limit where the minority class probabilities vanish.*

This paper makes the following contributions:

1. We construct a novel lower bound (Lemma 1) that holds for the general family of CL losses that are based on functions that are strictly convex and argument-wise strictly increasing and allow any value of k (the number of negative samples per positive-pair). This subsumes and generalizes popular loss functions such as the InfoNCE loss function. The bound is a convex function of the Gram matrix whose entries are the pairwise inner products of the class mean feature vectors. We also derive the asymptotic limit of the lower bound for the InfoNCE loss function when $k \uparrow \infty$ (Corollaries 1 and 2).
2. When the representation dimension $d \geq C - 1$, we prove that ***the lower bound has a unique minimizer which is rank-deficient with a unit-constant principal diagonal*** (Lemmas 1 – 4 and Theorem 2). We also show that ***the generalized CL loss is minimized when there is intra-class variance-collapse,*** *i.e.*, when the feature vectors of all the samples from the same class are identical (Corollary 3). However, the geometry of the optimal class feature vectors need not form an Equiangular Tight Frame (ETF) as in the balanced classes scenario. We show that the optimal geometry can be numerically computed as the solution to a convex program (Remark 1).
3. We prove that the geometric structure of the optimal class means exhibits a key equiangular symmetry structure that is determined by the relative class proportions (Theorem 3 and Corollary 4). We further show that these properties are consistent with corresponding results for balanced classes and ***resolve a question that was left open in Jiang et al. (2024a)****, namely whether the ETF geometry is optimal when the positive pairs are not conditionally independent given their class label and the classes of the positive and negative samples can collide (Remark 2).*
4. We further investigate the case when the class imbalance is extreme and prove that CL exhibits the MC phenomenon in the scenario where there is one majority class and equiprobable minority classes with the minor class probability less than ***a non-asymptotic threshold*** τ that depends on the number of classes, the number of negative samples per anchor, and bounds on the norms of the

subgradients of the CL loss function (Lemmas 8 – 10 and Theorem 4). Specializing to the InfoNCE loss function yields conservative ***parameter-free*** thresholds $\tau = 0.9292$ (Corollary 6) and $\tau = 0.9438$ (Corollary 7 in Appendix A) in different negative sampling settings.

5. Finally, we prove that all the above results hold under two different negative sampling settings: (1) Unsupervised CL (UCL), where the negative samples are selected from the whole dataset including samples from the same class as that of the anchor and (2) Supervised CL (SCL), where the negative samples are selected from classes that are different from that of the anchor.

The remainder of this paper is structured as follows. Section 2 formally introduces the CL framework and formulates the core optimization problem of interest. A tight lower bound for the generalized contrastive loss (and the $k \uparrow \infty$ asymptotic limit for the InfoNCE loss) that is a function of the mean feature vectors of the classes, together with necessary and sufficient conditions for equality, is established in Section 3. That the lower bound is a strictly convex function of the Gram matrix whose entries are the pairwise inner products of unit-norm class mean feature vectors, the necessity and sufficiency of intra-class variance-collapse for optimality, and the complete characterization of the optimal rank-deficient class means when $d \geq C-1$ are all established in Section 4. Equiangular symmetry properties of the optimal class means and their implications are established in Section 5. The MC phenomenon is investigated in Section 6 where a non-asymptotic threshold for MC is derived. Numerical experiments that corroborate and illustrate our theoretical results appear in Section 7. We end with a discussion of open questions in Section 8. Proofs of theoretical results are presented in Appendix A.

Notation: For $i, j \in \mathbb{Z}$, $i < j$, we define $i:j := i, i+1, \ldots, j$ and $a_{i:j} := a_i, a_{i+1}, \ldots, a_j$. If $i > j$, $i:j$ and $a_{i:j}$ are void expressions. We will denote the "all zeros" and "all ones" column vectors by $\mathbf{0}$ and $\mathbf{1}$, respectively. The dimensions of $\mathbf{0}$ and $\mathbf{1}$ will be clarified within each context they are used.

2 Contrastive learning problem setup and notation

Let $\mathcal{X} \subseteq \mathbb{R}^{d'}$ denote the data space, $f : \mathcal{X} \to \mathcal{Z}$ a representation function from data space to representation space (or feature space) $\mathcal{Z} \subseteq \mathbb{R}^d$, and $\mathcal{F}$ a (parameterized) family of such representation functions such as a those specified by a deep neural network with a specified architecture. Contrastive Learning (CL) is based on tuples $(x, x^+, x^-_{1:k}) \sim p(x, x^+, x^-_{1:k})$, where

1. x is called the anchor (or context),
2. x^+ the positive sample (relative to the given anchor x), and
3. $x^-_{1:k}$, $k \geq 1$, the negative examples (relative to the given anchor x).

The anchor x is also regarded as a positive sample and (x, x^+) is called a positive pair. The objective of CL is to learn a mapping $f \in \mathcal{F}$ via solving the following optimization problem,

$$\arg\min_{f \in \mathcal{F}} L(f), \quad L(f) := \mathbb{E}\left[\ell_k\left(x, x^+, x^-_{1:k}, f\right)\right], \tag{1}$$

where $L(f)$ is the CL risk of a representation function f with the expectation $\mathbb{E}[\cdot]$ (or empirical average) taken with respect to the joint distribution (or empirical distribution) $p(x, x^+, x^-_{1:k})$ and $\ell_k(\cdot)$ is a CL loss function that encourages *alignment* between the positive pairs (x, x^+) in representation space, as measured by the inner product $f(x)^\top f(x^+)$, and discourages the alignment between the k negative pairs $(x, x^-_i), i = 1:k$, in representation space, as measured by the inner products $f(x)^\top f(x^-_i), i = 1:k$.[1] The representation map learned via CL is treated as a pre-trained feature extractor and is used either directly or with fine-tuning in various downstream supervised tasks, predominantly classification.

In this work, we establish results that hold in great generality for the entire family of CL loss functions proposed in (Jiang et al., 2024a) as defined below.

[1] In Contrastive Learning, the feature vectors are typically normalized to have unit Euclidean length. Then, the inner product of two feature vectors is larger if, and only if, they are closer to each other in Euclidean distance. Therefore, the inner product of two feature vectors acts as an "inverse distance" or similarity measure between them.

Definition 1 (Generalized CL Loss Function)**.** *A Generalized CL loss function is of the form*

$$\ell_k(x, x^+, x^-_{1:k}, f) := \psi\big(f(x)^\top(f(x^-_1) - f(x^+)), \ldots, f(x)^\top(f(x^-_k) - f(x^+)\big) \tag{2}$$

where $\psi : \mathbb{R}^k \to \mathbb{R}$ is a function which is strictly convex and argument-wise strictly increasing (i.e., strictly increasing with respect to each argument when the other $k-1$ arguments are held fixed).[2] *The value of k in not restricted.*

We note that this subsumes and generalizes popular loss functions with spherical-ball normalized representations including the popular InfoNCE loss function defined in Appendix A.1 and its variants (InfoLOOB, N-pair, Decoupled Contrastive Loss, etc.) which have been widely used.[3] We focus on the general family in Definition 1 to highlight that all results presented in this paper only rely on two key properties of the CL loss function, namely convexity and monotonicity, and nothing else specific to a particular loss function like InfoNCE.

Unlike prior works which are restricted to the empirical CL risk where the joint distribution of the anchor, positive and negative samples (and also their latent labels in many works) are uniform over suitable discrete subsets, we adopt a general distributional perspective throughout and work with the population risk (which subsumes the empirical risk as a special case when the distribution is empirical) with the following key modeling assumptions that are consistent with the specialized assumptions on the (empirical) distribution of samples in prior works:
A1: Class labels. The samples have associated labels given by a deterministic labeling function $y(\cdot) : \mathcal{X} \to \mathcal{C} := \{1, \ldots, C\}, C > 1$. These labels represent classes in the supervised setting and latent, i.e., hidden, classes or clusters in the unsupervised setting.
A2: Positive samples. The joint distribution of positive samples is such that they have the same label. This can be ensured by design in the supervised setting, but in the unsupervised setting this is an assumption on the method used to sample a positive pair, e.g., an augmentation mechanism.
A3: Joint distribution. Let $x, x^+ \in \mathcal{X}$ be a pair of positive samples and $y \in \mathcal{C}$ their common class label. Let $x^-_{1:k} \in \mathcal{X}$ be a set of k negative samples associated with the positive pair and $y^-_{1:k} \in \mathcal{C}$ their respective class labels. In the UCL setting where the negative samples are chosen from the entire dataset, including possibly from the class of the positive pair, the joint distribution of all $(k+2)$ samples $x, x^+, x^-_{1:k}$ and their $(k+1)$ labels $y, y^-_{1:k}$ has the following form

$$\begin{aligned} p(x, x^+, x^-_{1:k}, y, y^-_{1:k}) &= p(y, y^-_{1:k})\, p(x, x^+, x^-_{1:k} | y, y^-_{1:k}) \ , \\ p(y, y^-_{1:k}) &= \lambda_y \prod_{t=1}^{k} \lambda_{y^-_t} \ , \end{aligned} \tag{3}$$

$$p(x, x^+, x^-_{1:k} | y, y^-_{1:k}) = q(x, x^+ | y) \prod_{t=1}^{k} s(x^-_t | y^-_t), \tag{4}$$

where $\lambda_{1:C} \in (0,1), \sum_{i \in \mathcal{C}} \lambda_i = 1$, denote the probabilities (or relative sample proportions) of the C possible classes ***and they need not be balanced***, $q(x, x^+|y)$ is the conditional distribution of a positive pair given their label, and $s(x^-|y^-)$ is the conditional distribution of a negative sample given that it is from class y^-.

We note that $(x^-_1, y^-_1), \ldots, (x^-_k, y^-_k)$ are independent and identically distributed (iid) and also independent of (x, x^+, y). The $(k+1)$ labels $y, y^-_{1:k}$ are iid which implies that, with non-zero probability, negative samples could have the same label as that of the positive pair, an event referred to as "class collision". Moreover, $x^-_{1:k}$ are conditionally iid given $y^-_{1:k}$, but unlike in (Jiang et al., 2024a), we do not assume that (x, x^+) are conditionally independent given their label y.
We focus on the UCL setting to establish all results. In Appendix A.18 we discuss how *all* our theoretical results continue to hold, with minor adjustments to some expressions, in the SCL setting where the negative

[2] As a technical aside, the function ψ is a so-called *proper* convex function because its range is $\mathbb{R}$ which excludes $-\infty$.

[3] The sigmoid loss does not satisfy Definition 1. Triplet loss corresponds to choosing $\psi(t_1, \ldots, t_k) = \sum_{i=1}^{k} \max\{t_i + \alpha, 0\}, \alpha > 0$. The $\psi(\cdot)$ function here is convex, but not strictly convex. All results in this paper, except those related to the uniqueness of the minimizer, also hold for the triplet loss.

samples are chosen from classes other than that of the positive pair, i.e., $y^-_{1:k} \in \mathcal{C} \setminus \{y\}$ w.p.1 and the anchor and positive sample are conditionally iid given their class. Then, $p(y, y^-_{1:k})$ in (3) is changed to

$$p_{SCL}(y, y^-_{1:k}) := \lambda_y \prod_{t=1}^{k} \left(\frac{\lambda_{y^-_t}}{1-\lambda_y} \right). \tag{5}$$

and $q(x, x^+|y) = s(x|y)\, s(x^+|y)$.

A4: Marginal conditional distributions. As in (Jiang et al., 2024a) and for analytical simplicity we also assume that

$$\forall i \in \mathcal{C}, \forall x, x^+ \in \mathcal{X}, \quad p(x|y=i) = s(x|i), \quad p(x^+|y=i) = s(x^+|i),$$

i.e., the *marginal* conditional distributions of x^+ given $y = i$ and x given $y = i$ are both $s(\cdot|i)$ which is the marginal conditional distribution of a negative sample x^- given $y^- = i$. This assumption can be ensured in the supervised setting, since labels are available. This also holds in the unsupervised setting, if a negative sample is generated using the same sampling mechanism that was used to generate a positive sample, e.g., via an augmentation of a reference sample. Indeed, in practical implementations Chen et al. (2020); Khosla et al. (2020); Jiang et al. (2024a), all samples in a mini-batch are first augmented using the same family of random augmentations and then the anchors, positives, and negatives are selected from these. Thus, negative samples are generated using the same augmentation-based sampling mechanism used to generate the positive pair. Consequently, the ***marginal*** conditional distributions of the positives and negatives are the same. We refer the readers to Jiang et al. (2024a) for a more detailed analysis. Under this assumption, for a representation function f and all $j \in \mathcal{C}$, if we let μ_j denote the mean of class j samples in the representation space, then we have

$$\forall j \in \mathcal{C}, \forall i \in \{1:k\}, \quad \mu_j = \mathbb{E}[f(x)|y=j] = \mathbb{E}[f(x^+)|y=j] = \mathbb{E}[f(x^-_i)|y^-_i = j]. \tag{6}$$

We define M as the $d \times C$ matrix of class means in representation space, specifically,

$$M := [\mu_1\ \mu_2 \cdots \mu_C]\,.$$

A5: Spherical-ball normalized representations. All prior theoretical studies of CL constrain the norms of the representations. This is a type of feature-normalization which typically improves the performance of CL in practice Wang & Isola (2020) and also makes the inner product a truer measure of "inverse distance" (see footnote 1). This can be done by explicitly requiring all representation maps in $\mathcal{F}$ to be norm-bounded for all samples, or implicitly by adding a quadratic penalty on the representation norms of the anchor, positive, and negative samples to the loss function. In our work, we will adopt the direct approach by requiring all representation functions to have a 2-norm less than or equal to one for all samples, i.e.,

$$\mathcal{F} = \{f : \forall x \in \mathcal{X},\ \|f(x)\|^2 := f^\top(x) f(x) \leq 1\}.$$

Thus, $\mathcal{F}$ is the family of all representation functions that are norm-bounded, but otherwise unconstrained. Note that since the representation vectors are confined to the unit ball, i.e., $\forall x \in \mathcal{X}, \|f(x)\|^2 \leq 1$, from the Cauchy-Schwarz inequality (or alternatively by the convexity of the squared norm function $\|\cdot\|^2$), we must have

$$\forall j \in \mathcal{C},\ \|\mu_j\|^2 = \|\mathbb{E}[f(x^-)|y^- = j]\|^2 = \|\mathbb{E}[f(x^+)|y=j]\|^2 = \|\mathbb{E}[f(x)|y=j]\|^2 \leq \mathbb{E}[\|f(x)\|^2|y=j] \leq 1.$$

A6: Unconstrained Features Model (UFM). In practice, the family of representation functions $\mathcal{F}$ is further constrained to be representable by a neural network having a specific architecture. The optimal solutions of the optimization problem in (1) will be included in such a family if the representation capacity of the neural network is sufficiently large, i.e., the neural network can approximate an arbitrary mapping $f : \mathcal{X} \to \mathbb{R}^d$ to any desired accuracy. Almost all prior theoretical studies of CL use UFM Fang et al. (2021); Graf et al. (2021) which treats a neural network's final-layer feature vectors, denoted by $z = f(x)$, as the free optimization variables instead of the network weights. This decouples feature geometry from the complex nonlinear encoder weight parameterization. UFM is used as an analytically tractable proxy for deep neural networks with a sufficiently high representation capacity. In this work we will also use UFM with the class of generalized CL loss functions

$$\ell(z, z^+, z^-_{1:k}) = \psi(z^\top(z^-_1 - z^+), \cdots, z^\top(z^-_k - z^+))$$

where $z = f(x), z^+ = f(x^+), z_1^- = f(x_1^-), \ldots, z_k^- = f(x_k^-)$.

The optimization problem in (1) was solved for special loss functions in the balanced dataset setting, i.e., $\lambda_1 = \lambda_2 = \ldots = \lambda_C = 1/C$, in Jiang et al. (2024a); Wang & Palmer (2023), where the optimal solution was shown to exhibit NC. Characterizing and computing the optimal solutions for imbalanced datasets was left open and is the primary focus of this work.

In Section 3, we will construct a tight lower bound for the generalized contrastive risk as a function of the class means, and then optimize this lower bound to find the optimal class means in Section 4.

3 Tight lower bound for CL risk in terms of class means

Our first key result is the following lemma which shows that it is possible to lower bound the contrastive risk by a function of the class means in representation space. Furthermore, this bound can be attained by any representation function f which collapses the representations of all samples within a class to the class mean and if all class means have unit norm. The lemma also shows that in order to achieve the lower bound, "intra-class variance-collapse", i.e., the collapse of the representations of all samples from the same class to their class mean, and unit norm class means are also necessary to attain the lower bound. In the next section, we will characterize the optimal class means that minimize the lower bound.

Lemma 1. *Let* $M := [\mu_1 \ \mu_2 \ \cdots \ \mu_C] \in \mathbb{R}^{d \times C}$. *Then,*

$$
\begin{aligned}
&L(f) \geq G(M), \\
&G(M) := \sum_{i, j_{1:k} \in \mathcal{C}} \left(\lambda_i \prod_{t=1}^{k} \lambda_{j_t} \right) \psi\big(\mu_i^\top \mu_{j_1} - 1, \ldots, \mu_i^\top \mu_{j_k} - 1)\big).
\end{aligned} \tag{7}
$$

The lower bound $G(M)$ can be attained if, and only if, there is within-class variance collapse, i.e., f maps all samples belonging to any class, to the mean representation vector of the class, i.e., $\forall x \in \mathcal{X}, f(x) = \mu_{y(x)}$, and $\forall i \in \mathcal{C}, ||\mu_i||^2 = 1$.

Proof Please see Appendix A.3. ■

Specializing (7) to the InfoNCE loss function defined in Appendix A.1 we get

Corollary 1. *For the InfoNCE loss function defined in Appendix A.1,*

$$
G(M) = \sum_{i, j_{1:k} \in \mathcal{C}} \left(\lambda_i \prod_{t=1}^{k} \lambda_{j_t} \right) \log \left(1 + \frac{1}{k} \sum_{t=1}^{k} e^{\mu_i^\top \mu_{j_t} - 1} \right). \tag{8}
$$

In practice, k could be large (e.g., $k = 128, 256, 512, \ldots$). In the limit $k \to +\infty$, the expression for the lower bound in Corollary (1) simplifies substantially.

Corollary 2. *For InfoNCE loss,*

$$
\begin{aligned}
\lim_{k \to \infty} L(f) &= \mathbb{E}\left[\log \left(1 + \mathbb{E}\left[e^{f^\top(x)(f(x_1^-) - f(x^+))} \middle| x, x^+ \right] \right) \right] &(9) \\
&\geq \lim_{k \to \infty} G(M) \\
&= \sum_{i \in \mathcal{C}} \lambda_i \log \left(1 + \sum_{j \in \mathcal{C}} r(j|i)\, e^{\mu_i^\top \mu_j - 1} \right). &(10)
\end{aligned}
$$

Proof Please see Appendix A.4. ■

4 Characterizing and computing optimal class means

An optimal matrix $M^* \in \mathbb{R}^{d\times C}$ which minimizes the lower bound in Lemma 1 can be found by solving the following constrained-optimization problem:

$$\min_{M\in\mathcal{M}} G(M), \quad \text{where} \tag{11}$$

$$\mathcal{M} := \{M = [\mu_1 \cdots \mu_C] \in \mathbb{R}^{d\times C} : \forall i \in \mathcal{C}, \|\mu_i\|^2 = 1\}. \tag{12}$$

A solution to (11) exists since the objective function $G(M)$ is continuous and the constraint set $\mathcal{M}$ is compact. However, neither is the objective function in (11) convex with respect to M nor is the constraint set defined in (12) convex due to the unit norm equality constraint. This complicates the development of computational methods for finding an optimal solution. Under additional special conditions on the representations, optimal solutions can be identified. For example, if the representations are confined to the non-negative orthant of $\mathbb{R}^d$, which can be implemented through the application of a non-negative activation function, e.g., ReLU, to the final layer of the neural network of the representation map, then we have the following result.

Theorem 1. *For all $f \in \mathcal{F}$, let $f(\mathcal{X}) \subseteq \mathbb{R}^d_{\geq 0}$. Then for all $f \in \mathcal{F}$,*

$$L(f) \geq \psi(-1, \ldots, -1)$$

with equality, if, and only if, $d \geq C$, $\mu_{1:C}$ are orthonormal, and $\forall x \in \mathcal{X}$, $f(x) = \mu_{y(x)}$.

Proof Please see Appendix A.5. ■

Theorem 1 in (Kini et al., 2024) is a specialized version of Theorem 1 for a restricted form of the InfoNCE loss. These results show that with additional non-negativity constraints on the representation and $d \geq C$, the geometry of the optimum representations is an orthonormal system *irrespective of the class imbalance*. To characterize the geometry without non-negativity constraints, let

$$A := M^\top M = \begin{bmatrix} \mu_1^\top\mu_1 & \mu_1^\top\mu_2 & \ldots & \mu_1^\top\mu_C \\ \mu_2^\top\mu_1 & \mu_2^\top\mu_2 & \ldots & \mu_2^\top\mu_C \\ \ldots & \ldots & \ldots & \ldots \\ \mu_C^\top\mu_1 & \mu_C^\top\mu_2 & \ldots & \mu_C^\top\mu_C \end{bmatrix} \in \mathbb{R}^{C\times C},$$

denote the Gram matrix of class means in representation space composed of their pairwise inner products. By construction, A is symmetric, i.e., $A^\top = A$, and positive semi-definite (PSD), i.e., $A \succcurlyeq 0$, which means that $\forall u \in \mathbb{R}^C, u^\top Au \geq 0$, and additionally, $\forall i \in \mathcal{C},\ A_{ii} = 1$ since $A_{ii} = ||\mu_i||^2 = 1$ is needed to attain the lower bound in Lemma 1. Let

$$\mathcal{A}^* := \{A \in \mathbb{R}^{C\times C} : A = A^\top, A \succcurlyeq 0, \forall i \in \mathcal{C}, A_{ii} = 1\} \text{ and} \tag{13}$$

$$S(A) := \sum_{i,j_{1:k}\in\mathcal{C}} \left(\lambda_i \prod_{t=1}^{k} \lambda_{j_t}\right) \psi\big(A_{ij_1} - 1, \ldots, A_{ij_k} - 1\big). \tag{14}$$

Under certain conditions, a solution to (11) can be found by minimizing (14) over (13).

Lemma 2. *For all $M \in \mathcal{M}$, $M^\top M \in \mathcal{A}^*$ and $G(M) = S(M^\top M)$. Let $A^* \in \mathcal{A}^*$ be a solution to the following optimization problem*

$$\min_{A\in\mathcal{A}^*} S(A). \tag{15}$$

If there exists an $M^* \in \mathcal{M}$ such that $(M^*)^\top M^* = A^*$, then M^* is solution to (11).

Proof Please see Appendix A.6. ■

Lemma 2 proves that if the global minimizer A^* of (15) can be factorized as $(M^*)^\top M^*$, then M^* is a solution to the original objective (11). We note that any $M \in \mathcal{M}$ can be mapped to an $A = M^\top M \in \mathcal{A}^*$. However, if $d < C$, it may not be possible to decompose all $A \in \mathcal{A}^*$ as $A = M^\top M$ for some $M \in \mathcal{M}$.

Lemma 3. *The function $S(\cdot)$ is a strictly convex function over $\mathcal{A}^*$. The constraint set $\mathcal{A}^* \subset \mathbb{R}^{C\times C}$ is convex and compact. Therefore, the minimization problem in (15) is a convex optimization problem and has a unique solution $A^* \in \mathcal{A}^*$, i.e.,*

$$S(A^*) = \min_{A\in\mathcal{A}^*} S(A). \tag{16}$$

Proof Please see Appendix A.7. ■

If $r = \text{rank}(A^*)$, then $r \leq C$ since $A^* \in \mathbb{R}^{C\times C}$. Also note that $\text{rank}((M^*)^\top M^*) \leq \min\{d, C\}$ since $M^* \in \mathbb{R}^{d\times C}$. Now, if $d \geq C$, then *any* $C \times C$ PSD matrix (therefore also A^*) can be factorized as $(M^*)^\top M^*$ (via the eigen-decomposition of A^* truncated to r nonzero eigenvalues). There is no "low-rankness" associated with the aforementioned statement. Interestingly, the next lemma proves that the unique optimal solution to (15) is rank-deficient. Specifically, it proves that A^* is ***guaranteed*** to have rank not exceeding $C-1$ even though there is no rank constraint imposed on the optimization problem ($A^* \in \mathbb{R}^{C\times C}$).

Lemma 4. *The unique solution $A^* \in \mathcal{A}^*$ to (15) has rank$(A^*) =: r \leq C-1$. Therefore, the minimum eigenvalue of A^* is zero.*

Proof Please see Appendix A.8. ■

We note that the result of Lemma 4 is a ***consequence*** of the uniqueness of the optimal A^* proved in Lemma 3. It is ***not*** a low-rank assumption or constraint. The next theorem puts the implications of Lemmas 2–4 together and proves that as long as $d \geq C-1$, it is possible to factorize A^* as $(M^*)^\top M^*$ and it explicitly constructs M^* using A^*'s eigen-decomposition truncated to r nonzero eigenvalues.

Theorem 2 (Optimal Class Means)**.** *Let $A^* = U_r\Sigma_r U_r^\top$ be the unique solution to 15, where $r := rank(A^*) \leq C-1$, $\Sigma_r \in \mathbb{R}^{r\times r}$ is a diagonal matrix with the r strictly positive eigenvalues of A^* along the main diagonal, and $U_r \in \mathbb{R}^{C\times r}$ is the matrix of r orthonormal eigenvectors of A^* corresponding to the r positive eigenvalues. If $d \geq C-1$, then $(M^*)^\top := [U_r\sqrt{\Sigma_r} \quad 0_{C\times d-r+1}]$ is a solution to (11), where $\sqrt{\Sigma_r}$ is a diagonal matrix with the square roots of the r positive eigenvalues of A along the main diagonal and $0_{C\times d-r+1}$ is the $C \times d-r+1$ matrix of all zeros.*[4] *Moreover, $\forall i \in \mathcal{C}, \|\mu_i^*\|^2 = 1$ where μ_i^* (i^{th} column of M^*) is an optimal class mean vector in representation space for class i.*

Proof Please see Appendix A.9 ■

The solution A^* to (15) (the optimum Gram matrix) is unique. However, the solution to (11) is not unique due to the rotational invariance of the loss function. The M^* defined in Theorem 2 is just one solution to (11) when $d \geq C-1$. Still, when $d \geq C-1$, any solution $\hat{M}^*$ to (11) will also satisfy $(\hat{M}^*)^\top \hat{M}^* = A^*$ because $G(\hat{M}^*) = S((\hat{M}^*)^\top \hat{M}^*) \geq S((M^*)^\top M^*) = S(A^*)$ and A^* is the unique minimizer of $S(A)$ over $\mathcal{A}^*$.

Remark 1. *For $d \geq C-1$, Theorem 2 offers a way to find the optimal mean vectors $\mu_1^*, \mu_2^*, \ldots, \mu_C^*$ via convex optimization. In our simulations in Section 7, we utilize the convex optimization package CVX Grant & Boyd (2014) to compute A^* and then use the spectral decomposition in Theorem 2 to compute an optimal mean representation vector matrix M^*.*

Corollary 3. *Let $d \geq C-1$, $M^* = [\mu_1^*, \mu_2^*, \ldots, \mu_C^*] \in \mathcal{M}$ be a solution to (11), and $A^* = (M^*)^\top M^*$ be the unique solution to (15). Then $L(f) = G(M^*) = S(A^*)$ for an $f \in \mathcal{F}$, if, and only if, $\forall x \in \mathcal{X}, f(x) = \mu_{y(x)}^*$.*

Proof This follows immediately from the optimality of A^* and M^* and Lemma 1. ■

The condition $C-1 \leq d$ is an assumption on the number of classes (an intrinsic property of dataset or application) relative to the representation dimension (a design choice, e.g., via suitable neural net architecture). This condition is also required in many papers to show the NC phenomenon and the existence of the ETF-structure, e.g., Jiang et al. (2024a); Graf et al. (2021); Wang & Palmer (2023); Dang et al. (2023). But they are all in the setting where classes are balanced, i.e., $\forall i \in \mathcal{C}$, $\lambda_i = 1/C$. In practice, $C-1 \leq d$ in applications where the number of classes is much smaller than the dimension of the representation space, *e.g.,* $d = 512$ in

[4] If $d = r-1$, then $0_{C\times d-r+1}$ is void.

ResNet-18 compared to $C = 10$ in the CIFAR10 dataset and $C = 100$ in the CIFAR100 dataset. The case $d < C - 1$, e.g., in LLMs, is currently an unresolved open problem.

An interesting implication of Corllary 3 is that, in order to globally minimize the contrastive risk, we only require the dimension of the representation space to be $d = C - 1$. This suggests that current approaches which use a very high-dimensional representation space to learn the features, may be inefficient in terms of storage and computational resources.

5 Equiangular properties of optimal class means

In this section, we show that the optimal class of classes that are equiprobable have an equiangular geometric structure. These are consequences of the uniqueness of A^*.

Theorem 3. *Suppose that there are two distinct classes i and j with the same probability, i.e., $\lambda_i = \lambda_j$. Let $M^* = [\mu_1^*, \mu_2^*, \ldots, \mu_C^*]$ be an optimal mean vector matrix such that ${M^*}^\top M^* = A^*$. Then,*

$$\forall n \in \mathcal{C} \setminus \{i, j\},\ {\mu_i^*}^\top \mu_n^* = {\mu_j^*}^\top \mu_n^*.$$

Proof The key idea of the proof is to show that if we swap μ_i^* and μ_j^* in M^* to form a new matrix Q, then $S(Q^\top Q) = S({M^*}^\top M^*)$. The detailed proof is presented in Appendix A.10. ■

The following Corollary expands the results of Theorem 3 to the scenario where multiple classes have the same probability.

Corollary 4. *Let $\mathcal{C} := \{1, 2, \ldots, C\}$ denote the set of C classes, and $\mathcal{C}' \subseteq \mathcal{C}$ a subset of classes that have the same probability. Then,*

$$\forall i, j, \in \mathcal{C}', i \neq j,\ {\mu_i^*}^\top \mu_j^* = \textit{constant}.$$

Proof Please see Appendix A.11. ■

Corollary 5. *If all classes are equiprobable, i.e., $\mathcal{C}' = \mathcal{C}$ in Corollary 4, then for all $i, j \in \mathcal{C}, i \neq j$, we have ${\mu_i^*}^\top \mu_j^* = -1/(C-1)$, $\forall i \in \mathcal{C}$, $\|\mu_i^*\|^2 = 1$, and $\sum_{i \in \mathcal{C}} \mu_i^* = 0$, i.e., the optimal class means form an equiangular tight frame (ETF) in $\mathbb{R}^d$.*

Proof Please see Appendix A.12. ■

Remark 2. *Corollary 5 resolves a question that was left open in Jiang et al. (2024a) for balanced datasets and the general CL loss function ψ, namely whether the ETF geometry is optimal when the positive pairs are not conditionally independent given their class label and the classes of the positive and negative samples can collide.*

We note that there is no simple analytical closed-form expression available for the angles between the optimal mean vectors in the general imbalanced setting. They can, however, be computed via a convex program as we noted in Remark 1.

6 Minority collapse

Minority collapse is a phenomenon that can be observed in imbalanced datasets. It refers to a scenario where the representations of all the samples in several distinct minority classes (classes with small probabilities) collapse into a single vector. In deep *classifier* neural networks it is known that minority collapse will occur if the class imbalance is extreme (Fang et al., 2021; Dang et al., 2023; 2024a; Hong & Ling, 2024). In this section, we show that minority collapse also occurs in contrastive learning for imbalanced datasets. To formally demonstrate the existence of this phenomenon, we consider the special scenario where $1 > \lambda_1 > \lambda_2 = \lambda_3 = \ldots = \lambda_C = \frac{1-\lambda_1}{C-1} > 0$, *i.e.*, the first class is the majority class and the remaining $C - 1$ classes are minority classes. This special

scenario is motivated by considerations of analytical tractability and the goal of deriving an explicit non-asymptotic sufficient condition under which the minority collapse phenomenon is guaranteed to manifest. We will prove that if the probability of the minority classes $\frac{1-\lambda_1}{C-1}$ is less than a certain threshold, or equivalently if λ_1 is greater than a threshold, then minority collapse will occur. We will derive an explicit formula for this threshold in terms of C, k, and bounds on the subgradients of the loss function ψ. We will then apply the formula to the InfoNCE loss function and derive a numerical threshold that holds for all $C \geq 3$ and all k.

Theorem 4 (Sufficient conditions for minority collapse)**.** *Let $C \geq 3$ and $1 > \lambda_1 > \lambda_2 = \lambda_3 = \ldots = \lambda_C = \frac{1-\lambda_1}{C-1} > 0$. Let $S(\cdot)$ be as in (14) with $\psi : \mathbb{R}^k \longrightarrow \mathbb{R}$ be strictly convex and argument-wise strictly increasing. Then ψ is Lipschitz over $\mathcal{V} := [-2, 0]^k$ with a Lipschitz constant $\Delta_2 < \infty$. For all $u \in \mathbb{R}^k_{\geq 0} \setminus \{\mathbf{0}\}$ and all $t \in [-2, 0]$, let $\phi_u(t) := \psi(t\, u)$. Then,*

$$\forall\, u \in \mathbb{R}^k_{\geq 0} \setminus \{\mathbf{0}\},\, \exists\, \delta_u \in (0, \infty)\, :\, \forall\, t, t' \in [-2, 0],\, t' \leq t, \quad (t - t')\, \delta_u \leq \phi_u(t) - \phi_u(t').$$

Let $\delta_{\mathbf{0}} := 0$ and

$$\delta_* := \min_{u \in \{0,1\}^k \setminus \{\mathbf{0}\}} \delta_u \in (0, \infty). \tag{17}$$

For all $y^-_{1:k} \in \mathcal{C}^k$, let $u(y, y^-_{1:k}) := (1(y^-_1 \neq y), \ldots, 1(y^-_k \neq y))^\top \in \{0, 1\}^k$, where $1(\cdot)$ is the indicator function. With $(y, y^-_{1:k})$ distributed as in (3), if

$$\lambda_1 \geq \frac{1}{1 + \frac{1}{\gamma_C \sqrt{k} \Delta_2}\, \mathbb{E}[\delta_{u(y, y^-_{1:k})} | \mathcal{E}_1]}, \tag{18}$$

where $\gamma_C := \frac{2(C-1)}{(C-2)}$, then for all $i \in \mathcal{C} \setminus \{1\}$, $\mu^*_i = -\mu^*_1$ with $||\mu^*_1|| = 1$, i.e., we have minority collapse. The sufficient condition for minority collapse given by (18) is satisfied if

$$\lambda_1 \in [\tau, 1), \quad \tau := \tfrac{1}{1 + \frac{\delta_*}{\gamma_C\, \sqrt{k}\, \Delta_2}} \in (0, 1). \tag{19}$$

Proof The detailed proof is long and presented in Appendix A.16. It consists of the following steps. Using Theorem 3, Corollary 4, the given class proportions, and the rank deficiency of A^* proved in Lemma 4, we first show (see Lemma 8 in Appendix A.13) that A^* belongs to a family of matrices parameterized by a single scalar $a \in [-1, 1]$ which equals the inner product between μ^*_1 and μ^*_i for any class $i \neq 1$. Next, using standard results in convex optimization theory, the fact that ψ is argument-wise strictly increasing, and the definition of subgradients and subdifferentials, we show (see Lemma 9 in Appendix A.14) that ψ is Lipschitz-Δ_2 over $[-2, 0]^k$ and also establish the properties of ϕ_u stated in the theorem. We also prove that $A^*(a)$ is element-wise Lipschtiz-γ_C (Lemma 10 in Appendix A.15). By combining these results, in Appendix A.16 we prove that if condition (18) is satisfied, then $S(A^*(a))$ is a strictly increasing function and therefore minimized at $a = -1$ which implies that for all $i \in \mathcal{C} \setminus \{1\}$, $\mu^*_i = -\mu^*_1$, i.e, we have minority collapse. Finally, we also show that the sufficient condition for minority collapse given by (18) is satisfied if condition (19) is satisfied. ■

We note that the condition $\lambda_1 \in \left(\frac{1}{1 + \delta_*/(\gamma_C\, \sqrt{k}\, \Delta_2)}, 1\right)$ is sufficient, but not necessary, for minority collapse and the threshold $\tau = \frac{1}{1 + \delta_*/(\gamma_C\, \sqrt{k}\, \Delta_2)}$ may be quite loose because it is based on δ_*, the smallest value of δ_u among all $u \neq \mathbf{0}$. Moreover, τ may depend on k and may go to 1 as k increases to infinity. For specific loss functions, such as InfoNCE, a more careful analysis of (18) can yield a non-trivial threshold that is independent of k. This is illustrated in the following corollary.

Corollary 6. *For the InfoNCE loss function defined in Appendix A.1, condition (18) for minority collapse in Theorem 4 is satisfied if*

$$\lambda_1 \in [\tau_C, 1), \text{ where } \tau_C := \frac{1 - \sqrt{1 - \beta_C^2}}{\beta_C} \text{ and } \beta_C := \frac{1}{1 + \frac{1}{4\gamma_C(1 + 3e^2)}}.$$

Moreover, for all $C \geq 3$, $\tau_C \leq \tau_3 \approx 0.9292$. Thus, $\lambda_1 \geq 0.9292$ is a sufficient condition for minority collapse for the InfoNCE loss function, irrespective of the number of classes C or the number of negative samples per anchor sample k.

Proof Please see Appendix A.17. ■

This completes the development of all our theoretical results for the UCL setting.

Remark 3. *As mentioned in Section 2, all our theoretical results in Sections 3 – 6 continue to hold, with minor adjustments to some expressions, in the SCL setting as well. This is discussed in detail in Appendix A.18. Numerical results that corroborate and illustrate the theoretical results are presented in Section 7.*

7 Computer experiments

This section provides two different types of experiments to verify the two phenomena investigated in Section 4 and Section 6, namely, **(1) intra-class variance-collapse (Section 7.1):** the representations of all the samples from the same class collapse to their class mean vector, and the optimal class mean vectors can be computed via a convex-optimization program and **(2) minority-collapse (Section 7.2):** if the probabilities of the minor classes are less than a threshold, then not only do the representations of all samples in the minor classes collapse to their class means, but also their class means collapse into a single vector. Since methods to select negative samples differ in the supervised (SCL) and unsupervised (UCL) settings, each experiment is performed under two different setups: **(a) SCL: the negative samples are selected from a class that is different from that of the positive samples**, and **(b) UCL: the negative samples are selected from the whole dataset, which may include the class of positive samples**. Although all our theoretical results are for a general loss function, we focus on the well-known InfoNCE loss for the experiments.

Since practical implementations use mini-batching, we now describe the mini-batch construction and the batch loss calculation used in our experiments. Let $\mathcal{X}_{\text{batch}} := \{x_{1:N}\}$ be a mini-batch (potentially a multiset) of N samples. For a given anchor sample $x_i \in \mathcal{X}_{\text{batch}}$, and a positive integer k, let $\mathcal{X}^-_{\text{batch},x_i} := \{x^-_{i1}, \ldots, x^-_{ik}\}$ be a multiset of k negative samples sampled from $\mathcal{X}_{\text{batch}}$ *with* replacement. In the UCL setting where the negative samples can be selected from any classes in the dataset, $x^-_{i1}, \ldots, x^-_{ik}$ are selected uniformly at random from $\mathcal{X}_{\text{batch}}$. In the SCL setting where negative samples must be selected from classes other than that of the positive samples, $x^-_{i1}, \ldots, x^-_{ik}$ are selected uniformly at random from classes different from that of x_i. Let $\mathcal{X}^+_{\text{batch},x_i}$ denote the set of samples in the batch with same label as sample x_i, *i.e.*, $\mathcal{X}^+_{\text{batch},x_i} := \{x \in \mathcal{X}_{\text{batch}} : y(x) = y(x_i)\}$. Then, our implemented loss function over a batch is:

$$
\begin{aligned}
&\mathcal{L}_{\text{batch}} \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{|\mathcal{X}^+_{\text{batch},x_i}|}\sum_{x_j\in\mathcal{X}^+_{\text{batch},x_i}} \psi_{\text{InfoNCE}}(f(x_i)^\top(f(x^-_{i1})-f(x_j)),\ldots,f(x_i)^\top(f(x^-_{ik})-f(x_j))\right] \\
&= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{|\mathcal{X}^+_{\text{batch},x_i}|}\sum_{x_j\in\mathcal{X}^+_{\text{batch},x_i}} \log\left(1+\frac{1}{k}\sum_{q=1}^{k} e^{f(x_i)^\top(f(x^-_{iq})-f(x_j))}\right)\right] \qquad (20)
\end{aligned}
$$

where $|\mathcal{X}^+_{\text{batch},x_i}|$ denotes the size of set $\mathcal{X}^+_{\text{batch},x_i}$ and ψ_{InfoNCE} is the InfoNCE loss function defined in (23).

Thus, the batch loss is computed by an outer average and an inner average of the loss function $\psi_{\text{InfoNCE}}(\cdot)$. In the outer average, $\psi_{\text{InfoNCE}}(\cdot)$ is averaged across N, $(k+1)$-tuples of anchor and negatives $(x_i, x^-_{i1}, \ldots, x^-_{ik})$, where we first select the anchor from $\mathcal{X}_{\text{batch}}$ and then k negatives associated with x_i from $\mathcal{X}_{\text{batch}}$ according to the appropriate negative sampling distribution of the SCL or UCL setting. For a given $(k+1)$-tuple $(x_i, x^-_{i1}, \ldots, x^-_{ik})$, in the inner average, $\psi_{\text{InfoNCE}}(\cdot)$ is averaged across $|\mathcal{X}^+_{\text{batch},x_i}|$ positive samples that have the same label as the anchor x_i. The batch loss can be interpreted as an empirical instantiation of the population loss via the nested (iterated) expectation

$$
\begin{aligned}
&\mathbb{E}[\psi_{\text{InfoNCE}}(f(x)^\top(f(x^-_1)-f(x^+)),\ldots,f(x)^\top(f(x^-_k)-f(x^+))] \\
&\qquad = \mathbb{E}\Big[\mathbb{E}\big[\psi_{\text{InfoNCE}}(f(x)^\top(f(x^-_1)-f(x^+)),\ldots,f(x)^\top(f(x^-_k)-f(x^+))|x, x^-_{1:k}\big]\Big]
\end{aligned}
$$

where the inner expectation is over x^+ for a given $(k+1)$-tuple $(x, x^-_1, \ldots, x^-_k)$. The overall loss in an epoch is the average of the batch loss across all the mini-batches in that epoch.

All our theoretical results were established for the batch setting. To ensure that they also hold in the mini-batch setting, as discussed in the recent work of Kini et al. (2024), mini-batches must be carefully constructed to prevent the formation of disjoint groups of non-interacting samples that remain "frozen" across epochs. One method to prevent this, proposed in Kini et al. (2024), is the so-called *batch-shuffling* method where the samples are divided into mini-batch partitions, with a random reshuffling of all samples in every epoch. We adopt this batch-shuffling method in our experiments.

7.1 Intra-class variance-collapse

In this section, we provide the numerical results to verify the intra-class variance-collapse phenomenon. We used a dataset comprising three classes extracted from the CIFAR-10 dataset. Specifically, we selected the first 1500, 750, and 750 image samples, respectively, from the first three classes (*i.e.*, $C = 3$), namely bird, automobile, and airplane, of the CIFAR10 dataset to form our dataset comprising 3000 samples. This corresponds to $\lambda_1 = 0.5$ and $\lambda_2 = \lambda_2 = 0.25$. We utilized the ResNet-50 architecture to implement the representation function f. To satisfy the condition $C = 3 \leq d + 1$ in Theorem 2, we set the dimension of the representation space to $d = 2$. We set the batch size and the number of epochs to 512 and 1000, respectively, and the number of negative samples to $k = 512$. We optimized the empirical CL risk using the Adam optimizer with a learning rate of 0.001.

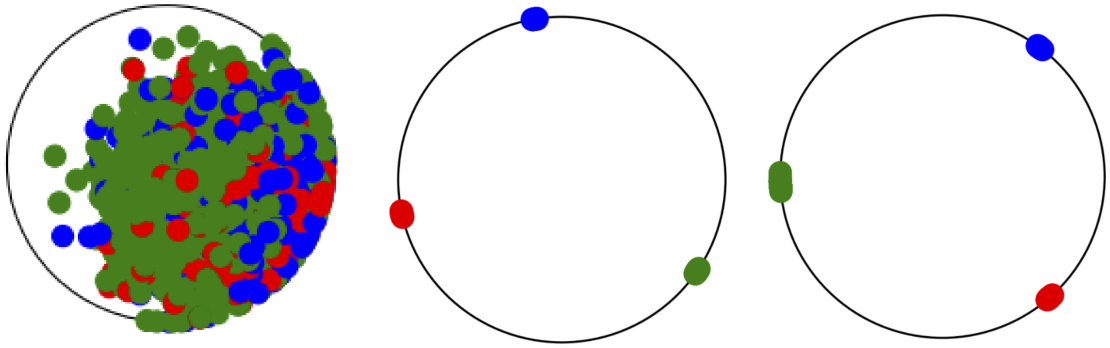


Figure 1: Intra-class variance-collapse in imbalanced datasets. Left: $\mathbb{R}^2$-space representations of 3000 images from 3 classes (indicated by color) of the CIFAR10 dataset using the initial representation function (i.e., before training). Middle: representation vectors of the same images at the conclusion of training when negative samples, within a mini-batch, are selected from classes that are different from those of the positive samples (SCL setting). Right: representation vectors of the same images at the conclusion of training when the negative samples can be selected from the entire mini-batch (UCL setting).

Figure 1 illustrates the two-dimensional representations of samples from three classes using: (a) the initial mapping before the commencement of training, (b) the optimal mapping at the conclusion of training in the SCL setting, and (c) the optimal mapping at the conclusion of training in the UCL setting. Evidently, all the samples from the same class (represented by the same color) nearly collapse to the same point, which is their class mean. As seen, when the negative samples can be selected from any classes in the dataset (UCL setting), including the class of positive samples, the distance between the two minority classes (red and blue) is much smaller compared to the setup where the negative samples are selected from classes that are different from those of the positive samples (SCL setting).

To verify that the optimal solutions obtained by the neural network are consistent with our theoretical results, we used the CVX modeling system (Grant & Boyd, 2014) to solve the convex optimization problem in (16). From Theorem 2, we know that the optimal mean vector matrix M^* is not unique, but the optimal Gram matrix A^* is unique and can be computed as the solution to a convex optimization problem. Therefore, we

compare the optimal Gram matrix provided by the neural network with the one computed using CVX. The optimal Gram matrices A^* obtained by the neural network and the CVX package are

$$A^*_{\text{Neural-Network}} = \begin{array}{c} \text{SCL setting} \\ \begin{bmatrix} 1.0000 & -0.6884 & -0.6910 \\ -0.6884 & 1.0000 & -0.0485 \\ -0.6910 & -0.0485 & 1.0000 \end{bmatrix} \end{array}, \quad \begin{array}{c} \text{UCL setting} \\ \begin{bmatrix} 1.0000 & -0.6301 & -0.6271 \\ -0.6301 & 1.0000 & -0.2097 \\ -0.6271 & -0.2097 & 1.0000 \end{bmatrix} \end{array}$$

$$A^*_{\text{CVX-package}} = \begin{bmatrix} 1.0000 & -0.6889 & -0.6889 \\ -0.6889 & 1.0000 & -0.0480 \\ -0.6889 & -0.0480 & 1.0000 \end{bmatrix}, \quad \begin{bmatrix} 1.0000 & -0.6284 & -0.6284 \\ -0.6284 & 1.0000 & -0.2105 \\ -0.6284 & -0.2105 & 1.0000 \end{bmatrix}$$

Evidently, both the neural network and CVX optimal solutions are very similar, and this empirically validates our theoretical results.

We also note that if the classes were balanced, then from Theorem 2 in Jiang et al. (2024a), the three optimal class means would form an equilateral triangle in the representation space (an equilateral triangle is an ETF in 2-D space). For our imbalanced datasets, the three class means clearly do not form an equilateral triangle. They do, however, form an isosceles triangle, and this empirically validates the result of Theorem 3 (since $\lambda_2 = \lambda_3$ in this experiment). This empirically confirms our claim that ETF is not the optimal geometric structure for imbalanced classes.

7.2 Minority collapse

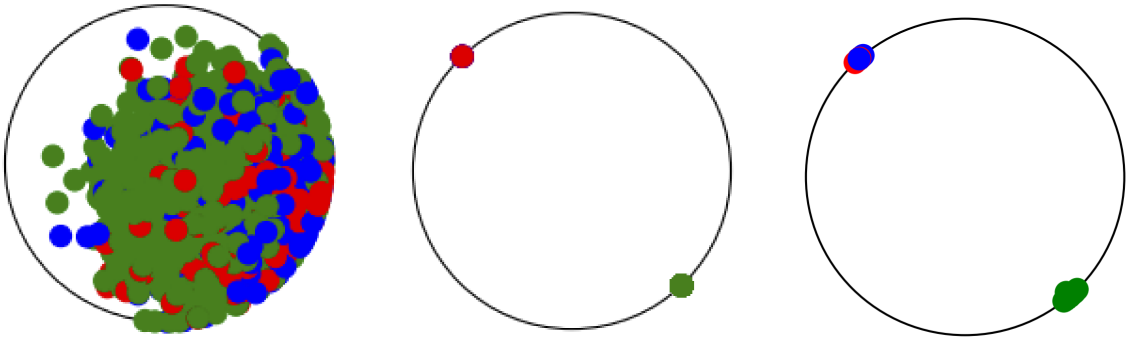
In this section, we provide the numerical results to verify the minority-collapse phenomenon. To do so, we constructed a three-class dataset with 2700, 150, and 150 image samples from the first three classes of the CIFAR-10 dataset, respectively, to form our second dataset of 3000 samples. This setup makes $\lambda_1 = 0.9$ and $\lambda_2 = \lambda_3 = 0.05$, which is the case when the data is heavily imbalanced. We utilized the ResNet-50 architecture to implement the representation function f. Similarly to the setup in Section 7.1, to satisfy the condition $C = 3 \leq d+1$ in Theorem 2, we set the dimension of the representation space to $d = 2$. We also set the batch size and the number of epochs to 512 and 1000, respectively, and the number of negative samples to $k = 512$. We optimized the empirical CL risk using the Adam optimizer with a learning rate of 0.001.

Figure 2 shows the representation vectors of all 3000 samples in the dataset at the beginning and at the end of training. Evidently, the representations of the two minor classes (blue and red) have collapsed (or nearly collapsed) into one vector (shown in red color), and the representations of these two classes are diametrically opposite on the unit circle to the representations of the major class (shown in green color). These results empirically validate the main conclusions of Section 6. We further note that $\lambda_1 = 0.9$ in this experiment is below the threshold of 0.9239 in Corollary 6 for UCL and 0.9438 in Corollary 7 for SCL, which guarantee minority collapse. This empirically bolsters our remarks before Corollary 6 that the threshold for minority collapse in Theorem 4 is sufficient for minority collapse, but may not be necessary.

The optimal Gram matrices A^* obtained by the neural network and the CVX package are

$$A^*_{\text{minority-collapse}} = \begin{array}{c} \text{SCL setting} \\ \begin{bmatrix} 1.0000 & -1.0000 & -1.0000 \\ -1.0000 & 1.0000 & 1.0000 \\ -1.0000 & 1.0000 & 1.0000 \end{bmatrix} \end{array}, \quad \begin{array}{c} \text{UCL setting} \\ \begin{bmatrix} 1.0000 & -0.9997 & -0.9997 \\ -0.9997 & 1.0000 & 0.9999 \\ -0.9997 & 0.9999 & 1.0000 \end{bmatrix} \end{array}$$

$$A^*_{\text{CVX-package}} = \begin{bmatrix} 1.0000 & -1.0000 & -1.0000 \\ -1.0000 & 1.0000 & 1.0000 \\ -1.0000 & 1.0000 & 1.0000 \end{bmatrix}, \quad \begin{bmatrix} 1.0000 & -1.0000 & -1.0000 \\ -1.0000 & 1.0000 & 1.0000 \\ -1.0000 & 1.0000 & 1.0000 \end{bmatrix}$$



Before training SCL post-training UCL post-training

Figure 2: Minority collapse in heavily imbalanced datasets. The $\mathbb{R}^2$-space representations of 3000 images from the first three classes of the CIFAR10 dataset before training (left sub-figure) collapse after training to two diametrically opposite points on the unit circle (middle and right sub-figures). The coincident or nearly coincident red/blue points in the middle and right sub-figures represent 300 samples from the second and third (minority) classes combined, whereas the green points represent the 2700 samples from the first (majority) class.

respectively, and they are identical up to the displayed numerical precision. This empirically corroborates our theoretical results that the optimal Gram matrix can be found efficiently using convex optimization.

8 Summary and Open Problems

In this paper, we proved that for a general family of CL losses (including the widely used InfoNCE loss) which are based on loss functions which are strictly convex and argument-wise strictly increasing, the optimal representations, will exhibit the intra-class variance-collapse phenomenon (representations of all samples from the same class must collapse to their class mean when globally minimizing the risk).

Even though there is no specific optimal structure or closed-form expression available for the optimal class means in the general imbalanced case, we derived an efficient method based on convex optimization to compute these optimal class means. We also established some equiangular properties of the optimal class means of equiprobable classes.

We further investigated a special case of extreme class imbalance and showed that CL also exhibits a phenomenon called minority collapse, wherein the optimal representations of all samples from the minority classes (classes with small probabilities) collapse into a single vector. Our key theoretical results were empirically validated through computer experiments.

Our work opens up several new problems that are of practical importance: (a) investigating the optimal geometry of neural collapse when the number of classes is more than the dimension of the representation space plus one – this scenario is particularly relevant to many large language models where embedding dimensions are typically on the order of hundreds and the the number of classes range in thousands during pre-training, (b) analyzing the neural collapse phenomenon with hard-negative samples – this is relevant to CL since it has been shown that hard-negative sampling alleviates issues with CL Jiang et al. (2024a); Robinson et al. (2020), and (c) characterizing non-asymptotic thresholds for the minority-collapse phenomenon for more than one major class.

References

Tina Behnia and Christos Thrampoulidis. Supervised contrastive representation learning: Landscape analysis with unconstrained features. In 2024 IEEE International Symposium on Information Theory (ISIT), pp. 575–580. IEEE, 2024.

Dimitri P. Bertsekas. Nonlinear Programming. Athena Scientific, Belmont, MA, 2nd edition, 2002.

D.P. Bertsekas. Convex Optimization Theory. Universities Press, 2010. ISBN 9788173717147. URL https://books.google.com/books?id=c80_nQAACAAJ.

Stephen P Boyd and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004.

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

Hien Dang, Tho Tran Huu, Stanley Osher, Hung The Tran, Nhat Ho, and Tan Minh Nguyen. Neural collapse in deep linear networks: From balanced to imbalanced data. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), Proceedings of the 40th International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, pp. 6873–6947. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/dang23b.html.

Hien Dang, Tho Tran Huu, Tan Minh Nguyen, and Nhat Ho. Neural collapse for cross-entropy class-imbalanced learning with unconstrained relu features model. In International Conference on Machine Learning, pp. 10017–10040. PMLR, 2024a.

Hien Dang, Tho Tran Huu, Tan Minh Nguyen, and Nhat Ho. Neural collapse for cross-entropy class-imbalanced learning with unconstrained ReLU features model. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pp. 10017–10040. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/dang24a.html.

Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. Proceedings of the National Academy of Sciences, 118 (43):e2103091118, 2021.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In International Conference on Machine Learning, pp. 3821–3830. PMLR, 2021.

Michael Grant and Stephen Boyd. Cvx: Matlab software for disciplined convex programming, version 2.1, 2014.

Wanli Hong and Shuyang Ling. Neural collapse for unconstrained feature model under cross-entropy loss with imbalanced data. Journal of Machine Learning Research, 25(192):1–48, 2024.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. Technologies, 9(1):2, 2020.

Ruijie Jiang, Thuan Nguyen, Shuchin Aeron, and Prakash Ishwar. Hard-negative sampling for contrastive learning: Optimal representation geometry and neural-vs dimensional-collapse. Transactions on Machine Learning Research, 2024a.

Ruijie Jiang, Thuan Nguyen, Prakash Ishwar, and Shuchin Aeron. Supervised contrastive learning with hard negative samples. In 2024 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE, 2024b.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. Advances in Neural Information Processing Systems, 33:18661–18673, 2020.

Ganesh Ramachandra Kini, Vala Vakilian, Tina Behnia, Jaidev Gill, and Christos Thrampoulidis. Symmetric neural-collapse representations with supervised contrastive loss: The impact of reLU and batching. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=AyXIDfvYg8.

Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. Transactions on Machine Learning Research, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=QTXocpAP9p.

Frank Nielsen and Gaëtan Hadjeres. Monte carlo information geometry: The dually flat case. arXiv preprint arXiv:1803.07225, 2018.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In International Conference on Learning Representations, 2020.

H. L. Royden. Real Analysis 3rd Ed. Macmillan Publishing Company, New York, NY, 1988.

Siwei Wang and Stephanie E Palmer. Towards understanding neural collapse in supervised contrastive learning with the information bottleneck method. arXiv preprint arXiv:2305.11957, 2023.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In International Conference on Machine Learning, pp. 9929–9939. PMLR, 2020.

A Proofs and additional supporting results

A.1 Strict convexity of the InfoNCE loss function

Lemma 5. *For all $i = 0, 1, \ldots, k$, let $\alpha_i > 0$. Then the generalized log-sum-exponential (GLSE) function*

$$\psi_{GLSE}(t_{1:k}) := \log\left(\alpha_0 + \sum_{i=1}^{k} \alpha_i e^{t_i}\right) \tag{21}$$

is strictly convex.

Proof The function $\psi_{\mathrm{GLSE}}(t_1, \ldots, t_k)$ is similar to the well-known "standard" log-sum-exponential function Boyd & Vandenberghe (2004). The standard log-sum-exponential function is known to be convex, but not strictly convex. Even though the result in Lemma 5 seems to be well-known, we are only able to find one reference that briefly mentions this result without a detailed proof Nielsen & Hadjeres (2018). Therefore, to make the paper self-contained, we provide the proof of Lemma 5 below.

For all $i \in \{1 : k\}$, let $u_i, v_i \in \mathbb{R}$, and $w_i := (1-\lambda)u_i + \lambda v_i$, where $\lambda \in (0, 1)$. Let $u_0 = v_0 = w_0 = 0$ and for some $i \in \{1 : k\}$, let $u_i \neq v_i$. If $p := \frac{1}{(1-\lambda)}$ and $q := \frac{1}{\lambda}$, then $p, q \in (1, \infty)$, $\frac{1}{p} + \frac{1}{q} = 1$, and we have

$$\begin{aligned}
\psi_{\mathrm{GLSE}}(w_{1:k}) &= \log\left(\alpha_0 + \sum_{i=1}^{k} \alpha_i e^{(1-\lambda)u_i + \lambda v_i}\right) \\
&= \log\left(\sum_{i=0}^{k} (\alpha_i e^{u_i})^{1/p} (\alpha_i e^{v_i})^{1/q}\right) \\
&\overset{\text{Hölder}}{\leq} \log\left(\Big(\sum_{i=0}^{k} \alpha_i e^{u_i}\Big)^{1/p} \Big(\sum_{i=0}^{k} \alpha_i e^{v_i}\Big)^{1/q}\right) \\
&= (1-\lambda)\log\left(\alpha_0 + \sum_{i=1}^{k} \alpha_i e^{u_i}\right) + \lambda \log\left(\alpha_0 + \sum_{i=1}^{k} \alpha_i e^{v_i}\right) \\
&= (1-\lambda)\psi_{\mathrm{GLSE}}(u_{1:k}) + \lambda \psi_{\mathrm{GLSE}}(v_{1:k}).
\end{aligned} \tag{22}$$

This shows that $\psi_{\mathrm{GLSE}}(\cdot)$ is a convex function. Equality holds in Hölder's inequality if, and only if, for all $i \in \{0 : k\}$, we have $((\alpha_i e^{u_i})^{1/p})^p = c\,((\alpha_i e^{v_i})^{1/q})^q$ for some constant c, i.e., $e^{u_i} = c\, e^{v_i}$, since $\alpha_i > 0$ for all $i \in \{0 : k\}$ and $1/p, 1/q \in (0, 1)$. Since $u_0 = v_0 = 0$, equality can occur if, and only if, $c = 1$. This would imply that $u_i = v_i$ for all $i \in \{1 : k\}$ which would contradict the assumption that for some $i \in \{1 : k\}$, $u_i \neq v_i$. This proves that the inequality in (22) is strict and therefore $\psi_{\mathrm{GLSE}}(\cdot)$ is a *strictly* convex function. ■

The InfoNCE loss function

$$\psi(t_{1:k}) = \psi_{\mathrm{InfoNCE}}(t_{1:k}) := \log\left(1 + \frac{1}{k}\sum_{i=1}^{k} e^{t_i}\right) \tag{23}$$

is argument-wise strictly increasing and is not only convex (being a log-sum-exponential with a positive offset within the logarithm), but also strictly convex since it is a GLSE function with $\alpha_0 = 1$ and $\alpha_1 = \ldots = \alpha_k = \frac{1}{k} > 0$.

A.2 Lemmas for proving variance collapse

Lemma 6. *Let u, v be iid random vectors in $\mathbb{R}^d$ with probability distribution $p(\cdot)$. If $u^\top v \overset{w.p.1}{=} 0$, then, $u \overset{w.p.1}{=} v \overset{w.p.1}{=} 0$.*

Proof Let $\mathcal{D} := \{1, \ldots, d+1\}$ and $w_1, \ldots, w_{d+1} \sim$ iid $p(\cdot)$. Since any $d+1$ vectors in d-dimensional space are linearly dependent,

$$\text{w.p.1. } \exists i \in \mathcal{D} : w_i \in \text{Span}(w_{1:d+1} \setminus \{w_i\}).$$

But for all $i \in \mathcal{D}$ and all $j \in \mathcal{D} \setminus \{i\}$, we have $w_i^\top w_j \overset{\text{w.p.1}}{=} 0$ since $w_i, w_j \sim$ iid $p(\cdot)$. This implies that

$$\exists i \in \mathcal{D} : w_i \overset{\text{w.p.1}}{=} 0.$$

But $u, v, w_1, \ldots, w_{d+1}$ all have the same distribution $p(\cdot)$. Therefore, $u \overset{\text{w.p.1}}{=} v \overset{\text{w.p.1}}{=} 0$. ■

Remark 4. *The result of Lemma 6 is false if u, v are independent, but not identically distributed, e.g., if $u = (u_1, 0)^\top, v = (0, v_2), \in \mathbb{R}^2, u_1, v_2$ iid standard normal. Clearly $u \overset{w.p.1}{\neq} 0$ and $v \overset{w.p.1}{\neq} 0$, but $u^\top v \overset{w.p.1}{=} 0$. Here, u, v are independent, but they are not identically distributed because the first component of $u \overset{w.p.1}{\neq} 0$ but the second is and it is reversed for v.*

Lemma 7. *Let z_1, z_2 be iid random vectors in $\mathbb{R}^d$ and $\mu := \mathbb{E}[z_1] = \mathbb{E}[z_2]$. If $z_1^\top z_2 \overset{w.p.1}{=} \gamma$, a constant, then, $z_1 \overset{w.p.1}{=} z_2 \overset{w.p.1}{=} \mu$ and $\gamma = ||\mu||^2$.*

Proof

$$z_1^\top z_2 \overset{\text{w.p.1}}{=} \gamma \Rightarrow \mathbb{E}[z_1^\top z_2 | z_1] \overset{\text{w.p.1}}{=} \gamma \Rightarrow z_1^\top \mathbb{E}[z_2|z_1] \overset{\text{w.p.1}}{=} \gamma \Rightarrow z_1^\top \mathbb{E}[z_2] \overset{\text{w.p.1}}{=} \gamma \Rightarrow z_1^\top \mu \overset{\text{w.p.1}}{=} \gamma$$

where the last but one implication is because z_1 and z_2 are independent. Since z_1 and z_2 are also identically distributed, we have

$$z_1^\top \mu \overset{\text{w.p.1}}{=} z_2^\top \mu \overset{\text{w.p.1}}{=} \gamma$$

Therefore, $\mathbb{E}[z_1^\top \mu] = \gamma \Rightarrow \mu^\top \mu = ||\mu||^2 = \gamma$. Next, define $u := z_1 - \mu$ and $v := z_2 - \mu$. Then u, v are iid random vectors in $\mathbb{R}^d$ and $u^\top v = z_1^\top z_2 - z_1^\top \mu - \mu^\top z_2 + \mu^\top \mu \overset{\text{w.p.1}}{=} \gamma - \gamma - \gamma + \gamma = 0$. By Lemma 6, $z_1 - \mu \overset{\text{w.p.1}}{=} z_2 - \mu \overset{\text{w.p.1}}{=} 0$ which implies that $z_1 \overset{\text{w.p.1}}{=} z_2 \overset{\text{w.p.1}}{=} \mu$. ■

A.3 Proof of Lemma 1

The proof makes use of the results in Appendix A.2 pertaining to variance collapse.

Proof

$$\begin{aligned}
&L(f) = \\
&= \mathbb{E}\left[\ell\Big(f(x), f(x^+), f(x_1^-), \ldots, f(x_k^-)\Big)\right] \\
&= \mathbb{E}\left[\psi\Big(f^\top(x)(f(x_1^-) - f(x^+)), \ldots, f^\top(x)(f(x_k^-) - f(x^+))\Big)\right] && (24) \\
&= \mathbb{E}\left[\mathbb{E}\left[\psi\Big(f^\top(x)(f(x_1^-) - f(x^+)), \ldots, f^\top(x)(f(x_k^-) - f(x^+))\Big)\,\middle|\, y, y_1^-, \ldots, y_k^-\right]\right] && (25) \\
&\geq \mathbb{E}\left[\psi\Big(\mathbb{E}[f^\top(x)\, f(x_1^-)|y, y_1^-] - \mathbb{E}[f^\top(x)\, f(x^+)|y], \ldots, \mathbb{E}[f^\top(x)\, f(x_k^-)|y, y_k^-] - \mathbb{E}[f^\top(x)\, f(x^+)|y]\Big)\right] && (26) \\
&= \mathbb{E}\left[\psi\Big(\mu_y^\top\, \mu_{y_1^-} - \mathbb{E}[f^\top(x)\, f(x^+)|y], \ldots, \mu_y^\top\, \mu_{y_k^-} - \mathbb{E}[f^\top(x)\, f(x^+)|y]\Big)\right] && (27) \\
&= \sum_{i, j_{1:k} \in \mathcal{C}} \left(\lambda_i \prod_{t=1}^{k} \lambda_{j_t}\right) \psi\big(\mu_i^\top \mu_{j_1} - \mathbb{E}[f^\top(x)\, f(x^+)|y = i], \ldots, \mu_i^\top \mu_{j_k} - \mathbb{E}[f^\top(x)\, f(x^+)|y = i]\big) && (28)
\end{aligned}$$

$$\geq \sum_{i,j_{1:k}\in\mathcal{C}} \left(\lambda_i \prod_{t=1}^{k} \lambda_{j_t}\right) \psi\big(\mu_i^\top \mu_{j_1} - 1, \ldots, \mu_i^\top \mu_{j_k} - 1\big), \tag{29}$$

where equality (24) follows from (2), equality (25) is the law of total expectation, inequality (26) is Jensen's inequality applied within the inner expectation conditioned on the labels of samples to the convex loss function $\psi(\cdot)$, (27) follows from the conditional independence of anchor and negative samples given their labels implied by (4), (28) follows by expanding the expectation in (27) in terms of all possible tuples of values of labels together with (3), and inequality (29) is because $\psi(\cdot)$ is an increasing function of all its arguments, all the weights $\left(\lambda_i \prod_{t=1}^{k} \lambda_{j_t}\right)$ are positive, and $f^\top(x)\, f(x^+) \leq ||f(x)|| \cdot ||f(x^+)|| \leq 1$ since the representations are constrained to be within the unit ball.

Clearly, if f is such that $\forall x \in \mathcal{X}, f(x) = \mu_{y(x)}$ and $\forall i \in \mathcal{C}, ||\mu_i||^2 = 1$, then $L(f) = G(M)$. We will now prove that these conditions are also necessary for equality. If $L(f) = G(M)$, then we must have equality in (26) and (29). Equality in (29) can be attained only if $\forall i \in \mathcal{C}$, with probability one (w.p.1) given $y = i$, i.e., under the distribution $q(x, x^+|i)$, we have $f^\top(x)\, f(x^+) = 1$. This is because $\psi(\cdot)$ is a *strictly* increasing function of its arguments, all the weights $\left(\lambda_i \prod_{t=1}^{k} \lambda_{j_t}\right)$ are *strictly* positive, and the norms of all representations are bounded by one. Therefore, w.p.1 given $y = i$, we must have $f(x) = f(x^+)$ and $||f(x)|| = 1$. Next, equality in the conditional Jensen's inequality (26) can be attained only if $\forall i \in \{1:k\}$, w.p.1 given y, y_i^-, we have $f^\top(x) f(x_i^-) - f^\top(x) f(x^+) = \mu_y^\top\, \mu_{y_i^-} - \mathbb{E}[f^\top(x)\, f(x^+)|y]$. This is because $\psi(\cdot)$ is a *strictly* convex function of its arguments and for all label tuples, $p(y, y_{1:k}^-) > 0$. This implies that $\forall i \in \{1:k\}$ and all $j, l \in \mathcal{C}$, w.p.1 given $y = j, y_i^- = l$, we have $f^\top(x) f(x_i^-) = \mu_j^\top\, \mu_l$ since, as we previously proved, equality in (29) implies that w.p.1 given $y = j$, we must have $f(x) = f(x^+)$ and $||f(x)|| = 1$. Taking $j = l$, we conclude that equality in (29) and (26) imply that $\forall i \in \{1:k\}$ and all $j \in \mathcal{C}$, w.p.1 given $y = y_i^- = j$, we have $f^\top(x) f(x_i^-) = ||\mu_j||^2$. But x, x_i^- are conditionally iid with distribution $s(\cdot|j)$ given $y = y_i^- = j$. From Lemma 7 in Appendix A.2, it then follows that for all $j \in \mathcal{C}$, w.p.1 given $y = j$, $f(x) = \mu_j$, or more compactly, $\forall x \in \mathcal{X}, f(x) = \mu_{y(x)}$. Thus we have shown that the conditions $\forall x \in \mathcal{X}, f(x) = \mu_{y(x)}$ and $\forall i \in \mathcal{C}, ||\mu_i||^2 = 1$ are both sufficient and necessary for the lower bound $G(M)$ to be attained, i.e., for $L(f) = G(M)$. ■

In the proof of necessity of within-class variance collapse for the attainment of the lower bound in Lemma 1, as an intermediate step we first proved that if we have equality in (29), then for each $i \in \mathcal{C}$, w.p.1 given $y = i$, we must have $f(x) = f(x^+)$ and $||f(x)|| = 1$. Without making any additional assumptions on the joint distribution of the positive pair, specifically, $q(x, x^+|y)$, we cannot conclude from here that we must have within-class variance collapse. For example, if $x^+ = x$ w.p.1, or if the samples in each class are grouped into non-overlapping pairs and x, x^+ are confined to be within a pair. But if, for example, the support of $q(x, x^+|y)$ is the Cartesian product of the supports of $s(x|y)$ and $s(x^+|y)$, then indeed we can conclude within-class variance collapse directly from equality in (29) alone without needing to analyze the conditions for equality in (26).

A.4 Proof of Corollary 2

Proof For $t = 1:k$, let

$$U_t(x, x^+, x_t^-) := e^{f^\top(x)(f(x_t^-) - f(x^+))},$$
$$V_t(y, y_t^-) := e^{(\mu_y^\top \mu_{y_t^-}) - 1}.$$

Then, from (24), the definition of the InfoNCE loss function in (23), and (26), (29), and (7) we have

$$L(f) = \mathbb{E}\left[\mathbb{E}\left[\log\left(1 + \frac{1}{k}\sum_{t=1}^{k} U_t(x, x^+, x_t^-)\right)\middle| x, x^+\right]\right], \tag{30}$$

$$G(M) = \mathbb{E}\left[\mathbb{E}\left[\log\left(1 + \frac{1}{k}\sum_{t=1}^{k} V_t(y, y_t^-)\right)\middle| y\right]\right]. \tag{31}$$

Since for all $x \in \mathcal{X}$, $\|f(x)\| \leq 1$, it follows from the convexity of the Euclidean norm and Jensen's inequality that for all $j \in \mathcal{C}$, $\|\mu_j\| = \|\mathbb{E}[f(x)|y = j]\| \leq \mathbb{E}[\|f(x)\|\,|y = j] \leq 1$ and therefore (by the Cauchy-Schwartz

inequality) $|f^\top(x)f(x_t^-)|, |f^\top(x)f(x^+)| \leq 1$. This proves that for all $t = 1:k$, $|U_t|, |V_t| \leq e^2$, i.e., they are bounded random variables. Now, $U_{1:k}|x, x^+$ and $V_{1:k}|y$ are conditionally iid. Thus, by the Strong Law of Large Numbers, their averages converge w.p.1 to their respective conditional expectations, i.e.,

$$\frac{1}{k}\sum_{t=1}^{k} U_t \xrightarrow[k\to\infty]{\text{w.p.1}} \mathbb{E}[U_1|x, x^+] = \mathbb{E}\Big[e^{f^\top(x)(f(x_1^-)-f(x^+))}\Big|x, x^+\Big], \tag{32}$$

$$\frac{1}{k}\sum_{t=1}^{k} V_t \xrightarrow[k\to\infty]{\text{w.p.1}} \mathbb{E}[V_1|x, x^+] = \mathbb{E}\Big[e^{(\mu_y^\top \mu_{y_1^-})-1}\Big|y\Big] = \sum_{j\in\mathcal{C}_y} r(j|y)\, e^{(\mu_y^\top \mu_j)-1}. \tag{33}$$

Since $U_{1:k}$ and $V_{1:k}$ are bounded by e^2 so are $(\sum_{t=1}^k U_t)/k$ and $(\sum_{t=1}^k V_t)/k$. The results (9) and (10) then follow from (30), (31), (32), (33), the Dominated Convergence Theorem, and the fact that $L(f) \geq G(M)$ proved in Lemma 1. ■

A.5 Proof of Theorem 1

Proof From Lemma 1, $L(f) \geq G(M)$ with equality if, and only if, $\forall x \in \mathcal{X}$, $f(x) = \mu_{y(x)}$ and $\mu_{1:C} \in \mathcal{M}$. For any $u, v \in \mathbb{R}^d_{\geq 0}$, $u^\top v \geq 0$ with equality only if u and v are orthogonal. In (7), $\psi(\cdot)$ is a *strictly* increasing function of its arguments and all the weights $\left(\lambda_i \prod_{t=1}^k \lambda_{j_t}\right)$ are *strictly* positive and sum to one. Therefore, $G(M)$ is minimized over $M \in \mathcal{M} \cap \mathbb{R}^{d\times C}_{\geq 0}$ if, and only if, $\mu_{1:C}$ are orthonormal. This requires $d \geq C$. ■

A.6 Proof of Lemma 2

Proof If $A := M^\top M$, then clearly, $A = A^\top$ and $A \succcurlyeq 0$ (since for all $u \in \mathbb{R}^C, u^\top M^\top M u = ||Mu||^2 \geq 0$), and $\forall i \in \mathcal{C}, A_{ii} = ||\mu_i||^2 = 1$. Therefore $A = M^\top M \in \mathcal{A}^*$. From (11) and (14), which define $G(\cdot)$ and $S(\cdot)$ respectively, it follows that $G(M) = S(A) = S(M^\top M)$. Therefore, for any $M \in \mathcal{M}$ we have $G(M) = S(M^\top M) \geq S(A^*) = S((M^*)^\top M^*) = G(M^*)$. This shows that M^* is a solution to (11). ■

A.7 Proof of Lemma 3

Proof *Convexity and compactness of $\mathcal{A}^*$:* The set $\mathcal{A}^*$ is clearly convex, since the set of all symmetric PSD matrices in $\mathbb{R}^{C\times C}$ satisfying the specified unit diagonal equality constraints is convex. The set $\mathcal{A}^*$ is also compact since $\mathcal{A}^* \subset \mathbb{R}^{C\times C}$ and for any $A \in \mathcal{A}^*$ and all $i, j \in \mathcal{C}$, $|A_{ij}| \leq |A_{ii}| \cdot |A_{jj}| = 1$, as we prove next. Since A is real, symmetric, and PSD, by the Real Spectral Theorem it has an eigendecomposition given by $A = U\Sigma U^\top$. If $\sqrt{A} := U\sqrt{\Sigma}U^\top$, where $\sqrt{\Sigma} \in \mathbb{R}^{C\times C}$ is a diagonal matrix with the square roots of C non-negative eigenvalues of A along the main diagonal, then $\sqrt{A}\cdot\sqrt{A} = A$. If $e_{1:C}$ is the standard basis for $\mathbb{R}^C$, then $|A_{ij}| = |e_i^\top A e_j| = |e_i^\top \sqrt{A}\sqrt{A} e_j| \leq ||\sqrt{A}e_i|| \cdot ||\sqrt{A}e_j|| = \sqrt{e_i^\top \sqrt{A}\sqrt{A}e_i} \cdot \sqrt{e_j^\top \sqrt{A}\sqrt{A}e_j} = \sqrt{e_i^\top A e_i} \cdot \sqrt{e_j^\top A e_j} = A_{ii}\cdot A_{jj} = 1$, where the first inequality is the Cauchy-Schwartz inequality. Thus, for all $i, j \in \mathcal{C}$, we have $|A_{i,j}| \leq 1$. This shows that $\mathcal{A}^*$ is a compact set.

Strict convexity of $S(\cdot)$ over $\mathcal{A}^*$: Let $A \in \mathcal{A}^*$. In (14), for all $i, j_{1:k} \in \mathcal{C}$, the k-tuples $\left(A_{ij_1} - 1, \ldots, A_{ij_k} - 1\right)$ are linear functions of A and the weights $\left(\lambda_i \prod_{t=1}^k \lambda_{j_t}\right)$ are all non-negative (in fact, they are all strictly positive). Since the function $\psi(\cdot)$ is convex (in fact, it is strictly convex), and $S(A)$ is a positive linear combination of convex functions of linear functions of A, it follows that $S(A)$ is a convex function of A. To prove that $S(\cdot)$ is strictly convex over $\mathcal{A}^*$, let $A, B \in \mathcal{A}^*, A \neq B$. Since $\forall i \in \mathcal{C}$, $A_{ii} = B_{ii} = 1$, we must have $A_{ij} \neq B_{ij}$ for at least one $i \neq j, i, j \in \mathcal{C}$. For any $t \in (0, 1)$, let $W := (1-t)A + tB$. Then, $W \in \mathcal{A}^*$ since $\mathcal{A}^*$ is a convex set and $A, B \in \mathcal{A}^*$, and $\forall i \in \mathcal{C}, W_{ii} = (1-t)A_{ii} + tB_{ii} = 1$. Since $\psi(\cdot)$ is a convex function of its

arguments, for all tuples $(i, j_{1:k}) \in \mathcal{C}^{k+1}$, we will have

$$\begin{aligned}(1-t)\,\psi\big(A_{ij_1} - A_{ii}, \ldots, A_{ij_k} - A_{ii}\big) + t\,\psi\big(B_{ij_1} - B_{ii}, \ldots, B_{ij_k} - B_{ii}\big) \\ &= (1-t)\,\psi\big(A_{ij_1} - 1, \ldots, A_{ij_k} - 1\big) + t\,\psi\big(B_{ij_1} - 1, \ldots, B_{ij_k} - 1\big) \\ &\geq \psi\big((1-t)\,(A_{ij_1} - 1) + t\,(B_{ij_1} - 1), \ldots, (1-t)\,(A_{ij_k} - 1) + t\,(B_{ij_k} - 1)\big) \\ &= \psi\big(W_{ij_1} - 1, \ldots, W_{ij_k} - 1\big) \\ &= \psi\big(W_{ij_1} - W_{ii}, \ldots, W_{ij_k} - W_{ii}\big)\end{aligned}$$

and the inequality is strict for at least one tuple $(i, j_{1:k}) \in \mathcal{C}^{k+1}$ because $\psi(\cdot)$ is a *strictly* convex function of its arguments, $A \neq B$, and $t \notin \{0, 1\}$. Since the weights $\big(\lambda_i \prod_{t=1}^{k} \lambda_{j_t}\big)$ in (14) are all strictly positive, it follows that $S(\cdot)$ is a strictly convex function over $\mathcal{A}^*$. ■

A.8 Proof of Lemma 4

Proof Let $\underline{\nu}(\cdot)$ denote the minimum eigenvalue of a matrix. We will prove that $\underline{\nu}(A^*) = 0$. For all $t > 0$, let

$$B(t) := A^* - t\,\mathbf{1}\mathbf{1}^\top + t\,I$$

where $\mathbf{1}$ is the $C \times 1$ vector of all ones and I is the $C \times C$ identity matrix. For all t, $B(t)$ is symmetric since $A^*, \mathbf{1}\mathbf{1}^\top$, and I are symmetric matrices. For all $i \in \mathcal{C}$, $B_{ii}(t) = A^*_{ii} - t + t = 1$ and for all $i, j \in \mathcal{C}, i \neq j$, $B_{ij}(t) = A^*_{ij} - t + 0 < A^*_{ij}$. Since $\psi(\cdot)$ is a *strictly* increasing function of all its arguments and all the weights $\lambda_i \prod_{t=1}^{k} (\lambda_{j_t}/(1-\lambda_i))$ in (14) are strictly positive, it follows that $S(B) < S(A^*)$. We now show that if $\underline{\nu}(A^*) > 0$, then $B(t)$ is PSD for $t = t' := \frac{\underline{\nu}(A^*)}{2(C-1)}$. This would imply that $B(t') \in \mathcal{A}^*$ and contradict the optimality of A^*. By the Courant-Fischer min-max theorem,

$$\begin{aligned}\underline{\nu}(B(t)) &= \min_{u \neq 0} \frac{u^\top B(t) u}{||u||^2} = \min_{u \neq 0} \frac{u^\top (A^* - t\,\mathbf{1}\mathbf{1}^\top + t\,I) u}{||u||^2} = \min_{u \neq 0} \frac{u^\top A^* u - t\,(u^\top \mathbf{1})^2 + t||u||^2}{||u||^2} \\ &\geq \min_{u \neq 0} \frac{u^\top A^* u - t\,C||u||^2 + t||u||^2}{||u||^2} \qquad (34) \\ &= \min_{u \neq 0} \frac{u^\top A^* u}{||u||^2} - (C-1)t = \underline{\nu}(A^*) - (C-1)t,\end{aligned}$$

where (34) is due to the Cauchy-Schwartz inequality. Therefore, $\underline{\nu}(B(t')) \geq \frac{\underline{\nu}(A^*)}{2}$. Thus, if $\underline{\nu}(A^*) > 0$, then $\underline{\nu}(B(t')) > 0$ which would make $B(t')$ a PSD matrix and contradict the optimality of A^*. We must therefore conclude that $\underline{\nu}(A^*) = 0$ which implies that $\text{rank}(A^*) < C$. ■

A.9 Proof of Theorem 2

Proof Lemma 4 proved that (11) has a unique solution A^* in $\mathcal{A}^*$ with rank r less than or equal to $C-1$. Since A^* is also a real, symmetric, PSD matrix, by the Real Spectral Theorem, it has a reduced eigen-decomposition given by $A^* = U_r \Sigma_r U_r^\top$. For all $d \geq C-1$, the matrix $(M^*)^\top := [U_r \sqrt{\Sigma_r} \quad 0_{C \times d-r+1}]$ is well defined and $(M^*)^\top M^* = U_r \sqrt{\Sigma} (\sqrt{\Sigma})^\top U_r^\top + 0_{C \times d-r+1} 0_{C \times d-r+1}^\top = U_r \Sigma_r U_r^\top = A^*$. From Lemma 2 it follows that M^* is a solution to (11). Moreover, for all $i \in \mathcal{C}$, we have $||\mu_i^*||^2 = A^*_{ii} = 1$. ■

A.10 Proof of Theorem 3

Proof The key idea of the proof is to show that if we swap μ_i^* and μ_j^* in M^* to form a new matrix Q, then $S(Q^\top Q) = S(M^{*\top} M^*)$. By construction, the gram matrix $B := Q^\top Q \in \mathcal{A}^*$ since $A^* = M^{*\top} M^* \in \mathcal{A}^*$.

Since the optimal Gram matrix is unique, $B = Q^\top Q = {M^*}^\top M^* = A^*$ and therefore for all $n \in \mathcal{C} \setminus \{i, j\}$, we must have $B_{jn} = {\mu_i^*}^\top \mu_n^* = A^*_{jn} = {\mu_j^*}^\top \mu_n^*$.
It remains to show that $S(Q^\top Q) = S({M^*}^\top M^*)$, i.e., $S(A^*) = S(B)$. To this end, let $\sigma : \mathcal{C} \to \mathcal{C}$ denote the bijection (specifically, a transposition permutation) where $\sigma(i) = j, \sigma(j) = i$, and for all $n \in \mathcal{C} \setminus \{i, j\}, \sigma(n) = n$. Then, $\sigma(\cdot)$ is its own inverse, i.e., $\forall n \in \mathcal{C}$, $\sigma(\sigma(n)) = n$. For notational convenience, let primed-indices denote the image under $\sigma(\cdot)$, i.e., $n' := \sigma(n)$. By construction of Q and the definition of $\sigma(\cdot)$, we have

$$\forall j_1, j_2 \in \mathcal{C},\ A^*_{j_1' j_2'} = B_{j_1 j_2}. \tag{35}$$

Since $\lambda_i = \lambda_j$, it follows from the definition of $\sigma(\cdot)$ that

$$\forall n \in \mathcal{C},\ \lambda_{n'} = \lambda_{\sigma(n')} = \lambda_n. \tag{36}$$

Therefore,

$$\begin{aligned}
S(A^*) &= \sum_{j_0, j_{1:k} \in \mathcal{C}} \left(\lambda_{j_0} \prod_{t=1}^{k} \lambda_{j_t} \right) \psi\big(A^*_{j_0 j_1} - A^*_{j_0 j_0}, \ldots, A^*_{j_0 j_k} - A^*_{j_0 j_0}\big). && (37)\\
&= \sum_{j_0', j_{1:k}' \in \mathcal{C}} \left(\lambda_{j_0'} \prod_{t=1}^{k} \lambda_{j_t'} \right) \psi\big(A^*_{j_0' j_1'} - A^*_{j_0' j_0'}, \ldots, A^*_{j_0' j_k'} - A^*_{j_0' j_0'}\big). && (38)\\
&= \sum_{j_0', j_{1:k}' \in \mathcal{C}} \left(\lambda_{j_0} \prod_{t=1}^{k} \lambda_{j_t} \right) \psi\big(B_{j_0 j_1} - B_{j_0 j_0}, \ldots, B_{j_0 j_k} - B_{j_0 j_0}\big). && (39)\\
&= \sum_{j_0, j_{1:k} \in \mathcal{C}} \left(\lambda_{j_0} \prod_{t=1}^{k} \lambda_{j_t} \right) \psi\big(B_{j_0 j_1} - B_{j_0 j_0}, \ldots, B_{j_0 j_k} - B_{j_0 j_0}\big). && (40)\\
&= S(B), && (41)
\end{aligned}$$

where (37) follows from the definition of $S(\cdot)$ in (14), equality (38) holds because $\sigma(\cdot)$ is a bijection, (39) is due to (35) and (36), equality (40) holds because $\sigma(\cdot)$ is a bijection, and (41) again follows from the definition of $S(\cdot)$ in (14). ■

A.11 Proof of Corollary 4

Proof The Corollary follows directly by applying the result in Theorem 3 to different pairs of $(i, j) \in \mathcal{C}'$ as follows. If $\mathcal{C}'$ contains only two classes, then the proof is immediate. If $\mathcal{C}'$ contains more than two classes, consider any three distinct classes $i, j, n \in \mathcal{C}'$. Then, from Theorem 3 we have (1) ${\mu_n^*}^\top \mu_j^* = {\mu_n^*}^\top \mu_i^*$ since $\lambda_i = \lambda_j$ and (2) ${\mu_i^*}^\top \mu_j^* = {\mu_n^*}^\top \mu_j^*$ since $\lambda_i = \lambda_n$. Therefore, ${\mu_i^*}^\top \mu_j^* = {\mu_n^*}^\top \mu_j^* = {\mu_n^*}^\top \mu_i^*$. In other words, any pair of class means has the same inner product. ■

A.12 Proof of Corollary 5

Proof If $\mathcal{C}' = \mathcal{C}$ in Corollary 4, then for all $i, j \in \mathcal{C}, i \neq j$, we have ${\mu_i^*}^\top \mu_j^* = b$ for some constant b. This implies that $A^* \in \mathcal{A}^*$ has the following form

$$A^* = (1 - b)I + b\mathbf{1}\,\mathbf{1}^\top = \begin{bmatrix} 1 & b & b & \cdots & b \\ b & 1 & b & \cdots & b \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b & b & b & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{C \times C}, \tag{42}$$

where I is the $C \times C$ identity matrix and $\mathbf{1} \in \mathbb{R}^C$ is the all-ones column vector. A matrix A^* having the above form has $(C-1)$ eigenvalues equal to $(1-b)$ and one eigenvalue equal to $(C-1)b+1$. Since A^* is PSD, $b \in [-1/(C-1), 1]$. By Lemma 4, the smallest eigenvalue of A^* is zero which implies that either $1-b=0 \Rightarrow b=1$ or $(C-1)b+1=0 \Rightarrow b=-1/(C-1)$. For both choices of b, A^* is PSD, but for the choice $b=-1/(C-1)$ (the smaller choice), the value of $S(A^*)$ is smaller because for A^* having the form in (42),

$$S(A^*) = \sum_{i, j_{1:k} \in \mathcal{C}} \left(\lambda_i \prod_{t=1}^{k} \lambda_{j_t} \right) \psi\big((b-1)1(j_1 \neq i), \ldots, (b-1)1(j_k \neq i)\big),$$

where $1(\cdot)$ is the indicator function, and ψ is a strictly increasing function of all it arguments. Thus $b=-1/(C-1)$. Finally, $||\sum_{i \in \mathcal{C}} \mu_i^*||^2 = \sum_{i \in \mathcal{C}} ||\mu_i^*||^2 + \sum_{i \neq j, i,j \in \mathcal{C}} (\mu_i^*)^\top \mu_j^* = C - C(C-1)/(C-1) = 0$. ■

A.13 Structure of Optimum A^*

Lemma 8. *Let $C \geq 3$ and $1 > \lambda_1 > \lambda_2 = \lambda_3 = \ldots = \lambda_C = \frac{1-\lambda_1}{C-1} > 0$. Then*

$$A^* = \begin{bmatrix} 1 & a & a & \cdots & a \\ a & 1 & b & \cdots & b \\ a & b & 1 & \cdots & b \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a & b & b & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{C \times C}, \tag{43}$$

with $a \in [-1, 1]$ and $b = (a^2(C-1)-1)/(C-2)$.

The form of A^* in (43) follows from Theorem 3 and Corollary 4. The condition on b follows from the rank deficiency of A^* proved in Lemma 4. This requires a careful analysis of the eigenstructure of PSD matrices having the form in (43). The detailed proof is presented below.

Proof From Theorem 3 and Corollary 4, it follows that $\forall i \in \mathcal{C} \setminus \{1\}$, ${\mu_i^*}^\top \mu_1^* = a$ and $\forall i, j \in \mathcal{C} \setminus \{1\}$, $i \neq j$, ${\mu_i^*}^\top \mu_j^* = b$ for some constants $a, b \in [-1, 1]$. Thus, $A^* \in \mathbb{R}^{C \times C}$ is of the form

$$A^* = \begin{bmatrix} 1 & a & a & \cdots & a \\ a & 1 & b & \cdots & b \\ a & b & 1 & \cdots & b \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a & b & b & \cdots & 1 \end{bmatrix}. \tag{44}$$

Since $A^* \in \mathcal{A}^*$, it is PSD and all its eigenvalues are non-negative. From Lemma 4, the minimum eigenvalue of A^* is zero. We will show that this implies either $a=0$ and $b=1$ or $a \in [-1, 1]$ and $b = (a^2(C-1)-1)/(C-2)$.

To this end, let $\mathbf{1} \in \mathbb{R}^C$ denote the all-ones column vector and $e_1 \in \mathbb{R}^C$ the standard basis vector whose first component is one and the remaining components are zero. Let $u := \mathbf{1} - e_1$. Then, $u \perp e_1$ and

$$A^* = (1-b)\, I + b\, u\, u^\top + b\, e_1\, e_1^\top + a\, e_1\, u^\top + a\, u\, e_1^\top, \tag{45}$$

where I is the $C \times C$ identity matrix. Let $v_{1:C}$ be any orthonormal basis for $\mathbb{R}^C$ with $v_1 := e_1$, $v_2 := u/||u||$, and $v_{3:C} \in \text{Span}^\perp(e_1, u)$. Then using (45), it follows that for all $i \geq 3$,

$$A^* v_i = (1-b) v_i = 0.$$

This shows that $v_{3:C}$ are $(C-2)$ orthonormal eigenvectors of A^* with eigenvalue $(1-b)$. The remaining two eigenvectors of A^* must therefore belong to $\text{Span}(e_1, u)$. Let $v = \alpha e_1 + \beta u$ be an eigenvector of A^* in $\text{Span}(e_1, u)$ with eigenvalue $\nu \geq 0$. Then $v = (\alpha\ \beta \ldots \beta)^\top$ and either $\alpha \neq 0$ or $\beta \neq 0$ because, by definition, an eigenvector is a non-zero vector. Since $A^*\, v = \nu\, v$ and A^* has the form shown in (44), we have

$$\alpha + (C-1)\, a\, \beta = \nu\, \alpha \Rightarrow (C-1)\, a\, \beta = -(1-\nu)\, \alpha \tag{46}$$

$$a\,\alpha + \beta + (C-2)\,b\,\beta = \nu\,\beta \Rightarrow \beta((1-\nu) + (C-2)b) = -a\,\alpha \tag{47}$$

Case $a = 0$. Then, $b \neq 0$ since otherwise we would have $A^* = I$ which has C eigenvalues all equal to one and this would contradict the result of Lemma 4. With $a = 0, b \neq 0$, (46) would imply that $(1-\nu)\,\alpha = 0$ which would imply that either $\nu = 1$ or $\alpha = 0$. If $\nu = 1$, then (47) together with $a = 0$ and $b \neq 0$ would imply that $\beta = 0$ which would, in turn, imply that $\alpha \neq 0$ since both α and β cannot be simultaneously zero. Thus, when $a = 0$, one eigenvalue is $\nu = 1$ with eigenvector given by $\alpha \neq 0, \beta = 0$. If $a = 0$ and we have $\nu \neq 1$, then $\alpha = 0$, $\beta \neq 0$, and $(1-\nu) + (C-2)b = 0 \Rightarrow \nu = 1 + (C-2)b$. In summary, if $a = 0$ then $b \neq 0$ and A^* would have $(C-2)$ eigenvalues equal to $(1-b)$, one eigenvalue equal to 1, and one eigenvalue equal to $1 + (C-2)b$. Since the smallest eigenvalue of A^* is zero, this would imply that either $b = 1$ or $b = -1/(C-2)$.

Case $a \neq 0$. In this case we must have $\nu \neq 1$ because otherwise (46) and $C \geq 3$ would imply that $\beta = 0$ and then (47) would imply that $\alpha = 0$ which would contradict the assumption that both α and β cannot be zero simultaneously. Thus, $\nu \neq 1$. Then, (46) would imply that $\alpha = -(C-1)a\,\beta/(1-\nu)$. Substituting this into (47) gives us

$$\beta((1-\nu) + (C-2)b) = \beta\,\frac{(C-1)\,a^2}{(1-\nu)} \Rightarrow (1-\nu)^2 + (C-2)\,b\,(1-\nu) - (C-1)\,a^2 = 0,$$

where we could cancel the common factor β in the first equation because $\beta \neq 0$ (if $\beta = 0$ then with $\nu \neq 1$, (46) would imply that $\alpha = 0$, a contradiction). Solving for the roots of the quadratic equation in $(1-\nu)$ we get

$$\nu = 1 + \frac{(C-2)\,b}{2} \pm \sqrt{\frac{(C-2)^2 b^2}{4} + (C-1)a^2} \tag{48}$$

In summary, if $a \neq 0$, then A^* would have $(C-2)$ eigenvalues equal to $(1-b)$ and two eigenvalues given by (48). Since the smallest eigenvalue of A^* is zero, this would imply that either $b = 1$ or

$$1 + \frac{(C-2)\,b}{2} - \sqrt{\frac{(C-2)^2 b^2}{4} + (C-1)a^2} = 0 \Rightarrow b = \frac{(C-1)\,a^2 - 1}{(C-2)}. \tag{49}$$

Observe that if we substitute $a = 0$ into the expression for b in terms of a given by (49), we get $b = -1/(C-2)$, which is consistent with one of the two possibilities that we obtained when we previously analyzed the case $a = 0$. Combining the analysis of both cases, we conclude that we must have either $a = 0, b = 1$ or $a \in [-1, 1]$, $b = ((C-1)\,a^2 - 1)/(C-2)$.

Since $\psi(\cdot)$ is a *strictly* increasing function of all its arguments and all the weights $\lambda_i \prod_{t=1}^{k}(\lambda_{j_t}/(1-\lambda_i))$ in (14) are strictly positive, $S(A^*)$ will have a strictly smaller value when $a = 0, b = -1/(C-1)$ than when $a = 0, b = 1$. Therefore, we must have $a \in [-1, 1]$, $b = ((C-1)\,a^2 - 1)/(C-2)$. ■

A.14 Subgradients of strictly convex and argument-wise strictly increasing functions

Lemma 9. *Let $\psi : \mathbb{R}^k \longrightarrow \mathbb{R}$ be strictly convex and argument-wise strictly increasing. Then for all $v \in \mathbb{R}^k$, the subdifferential set $\partial\psi(v)$ is non-empty, convex, and compact. Moreover, if $\mathcal{V} := [-2, 0]^k$, then $\mathcal{S}(\psi) := \cup_{v\in\mathcal{V}}\partial\psi(v)$ is bounded and ψ is Lipschitz over $\mathcal{V}$. Specifically, if*

$$\Delta_2 := \sup_{w\,\in\,\mathcal{S}(\psi)} ||w||_2, \quad \textit{then } \Delta_2 < \infty, \quad \mathcal{S}(\psi) \subseteq (0, \Delta_2]^k, \textit{ and}$$
$$\forall\, v, v' \in \mathcal{V}, \quad |\psi(v) - \psi(v')| \leq \Delta_2\,||v - v'||_2. \tag{50}$$

For all $u \in \mathbb{R}^k_{\geq 0} \setminus \{\mathbf{0}\}$ and all $t \in [-2, 0]$, let $\phi_u(t) := \psi(t\,u)$. Then,

$$\forall\, u \in \mathbb{R}^k_{\geq 0} \setminus \{\mathbf{0}\},\, \exists\, \delta_u \in (0, \infty) \,:\, \forall\, t, t' \in [-2, 0],\, t' \leq t, \quad (t - t')\,\delta_u \leq \phi_u(t) - \phi_u(t'). \tag{51}$$

If $u = \mathbf{0}$, then for all t, $\phi_{\mathbf{0}}(t) = \psi(\mathbf{0})$ and we define $\delta_{\mathbf{0}} := 0$. If $\nabla\psi(v)$ exits for all $v \in \mathcal{V}$, then $\Delta_2 = \sup_{v\in\mathcal{V}} ||\nabla\psi(v)||_2$ and $\forall\, u \in \mathbb{R}^k_{\geq 0}$, $\delta_u = u^\top \nabla\psi(-2u)$.

The proof essentially follows from standard results in convex optimization theory, the fact that ψ is argument-wise strictly increasing, and the definition of subgradients and subdifferentials. The detailed proof is presented below.

Proof Proposition 5.4.2 in (Bertsekas, 2010) and Proposition B.24 in Appendix B of (Bertsekas, 2002) prove that the subdifferential set $\partial\psi(v)$ at any point $v \in \mathbb{R}^k$ of any real-valued convex function $\psi : \mathbb{R}^k \longrightarrow \mathbb{R}$, is non-empty, convex, and compact. Moreover, the union of subdifferential sets of all points belonging to any non-empty compact set $\mathcal{V}$ is also bounded, i.e., $\cup_{v\in\mathcal{V}}\partial\psi(v)$ is bounded.

In the lemma, we have $\mathcal{V} = [-2, 0]^k$ which is a non-empty compact set. Therefore, $\mathcal{S}(\psi) := \cup_{v\in\mathcal{V}}\partial\psi(v)$ is bounded and $\Delta_2 := \sup_{w\in\mathcal{S}(\psi)} ||w||_2 < \infty$. For any vector $w \in \mathbb{R}^k$ we have $||w||_\infty \leq ||w||_2$. This implies that for all $w \in \mathcal{S}(\psi)$, we have $||w||_\infty \leq \Delta_2$. Since ψ is also strictly increasing over $\mathbb{R}^k$, all components of any subgradient vector at any point are strictly positive. Specifically, for all $v \in \mathcal{V}$, all subgradients $w \in \partial\psi(v)$, all $i \in \mathcal{C}$, and all $t > 0$, we have (by the definition of a subgradient)

$$-t(e_i^\top w) + \psi(v) \leq \psi(v - t\, e_i)$$

where e_i is the i^{th} standard basis vector of $\mathbb{R}^k$. Thus, the i^{th} component of w is bounded from below as follows

$$(e_i^\top w) \geq \frac{\psi(v) - \psi(v - t\, e_i)}{t} > 0$$

where the last inequality is strict since ψ is argument-wise strictly increasing and $t > 0$. Therefore, we conclude that $\mathcal{S}(\psi) \subseteq (0, \Delta_2]^k$.

Next, for all $v, v \in [-2, 0]^k$, all $w \in \partial\psi(v)$, and all $w' \in \partial\psi(v')$, by the definition of a subgradient, the fact that $||w||_2, ||w'||_2 \leq \Delta_2 < \infty$, and the Cauchy-Schwartz inequality, we have

$$\begin{aligned}-\Delta_2||v - v'||_2 \leq -||w'||_2 \cdot ||v' - v||_2 \leq (v - v')^\top w' &\leq \psi(v) - \psi(v') \\ &\leq (v' - v)^\top w \leq ||w||_2 \cdot ||v' - v||_2 \leq \Delta_2||v' - v||_2.\end{aligned}$$

Thus, $|\psi(v) - \psi(v')| \leq \Delta_2\, ||v - v'||_2$. If $\nabla\psi(v)$ exists for all $v \in \mathcal{V}$, then $\mathcal{S}(\psi) = \{\nabla\psi(v) : v \in \mathcal{V}\}$ and $\Delta_2 = \sup_{v\in\mathcal{V}} ||\nabla\psi(v)||_2$.

Since ψ is strictly convex and argument-wise strictly increasing over $\mathbb{R}^k$, it follows that $\forall u \in \mathbb{R}^k_{\geq 0} \setminus \{\mathbf{0}\}$, $\phi_u(t) := \psi(t\, u)$ is also strictly convex and strictly increasing over $\mathbb{R}$ (strictly, because at least one component of u is strictly positive). According to the "chord-slopes inequality" for convex functions (see (Royden, 1988), Chapter 5, Section 5), if $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ is convex, then for all $s_1, s_2, s_1', s_2' \in \mathbb{R}$ such that $s_1 \leq s_1' < s_2'$ and $s_1 < s_2 \leq s_2'$, we have

$$\frac{\phi(s_2) - \phi(s_1)}{s_2 - s_1} \leq \frac{\phi(s_2') - \phi(s_1')}{s_2' - s_1'}.$$

Applying this inequality to ϕ_u with $s_2' = t$, $s_1' = t'$, with $-2 \leq t' < t \leq 0$, and $s_2 = -2$, and $s_1 = -2 - \epsilon$, where $\epsilon > 0$, we get

$$\frac{\phi_u(-2) - \phi_u(-2 - \epsilon)}{(-2) - (-2 - \epsilon)} \leq \frac{\phi_u(t) - \phi_u(t')}{t - t'}.$$

Since ϕ_u is a strictly increasing function, we get

$$0 < \delta_u := \sup_{\epsilon>0}\left[\frac{\phi_u(-2) - \phi_u(-2 - \epsilon)}{\epsilon}\right] \leq \frac{\phi_u(t) - \phi_u(t')}{t - t'}.$$

Thus, for all $t, t' \in [-2, 0]$, with $t' < t$, we have

$$(t - t')\, \delta_u \leq \phi_u(t) - \phi_u(t').$$

The last inequality clearly holds when $t' = t$ as well.

If $\nabla\psi(v)$ exists for all $v \in \mathcal{V}$, then

$$\delta_u = u^\top \nabla\psi(-2\, u),$$

since for all $\epsilon > 0$, the convexity of ϕ_u implies that $-\epsilon\,\nabla\phi_u(-2) + \phi_u(-2) \leq \phi_u(-2-\epsilon) \Rightarrow \frac{\phi_u(-2)-\phi_u(-2-\epsilon)}{\epsilon} \leq \nabla\phi_u(-2) = u^\top \nabla\psi(-2\,u)$, and $\lim_{\epsilon\downarrow 0} \frac{\phi_u(-2)-\phi_u(-2-\epsilon)}{\epsilon} = \nabla\phi_u(-2)$. ■

A.15 $A^*(a)$ is Lipschitz

Lemma 10. *Let $C \geq 3$ and let $A^*(a)$ denote the matrix A^* in Equation (43) of Lemma 8 with $a \in [-1,1]$ and $b = (a^2(C-1)-1)/(C-2)$. Then, for all $a, a' \in [-1,1]$ such that $a' \leq a$, and all $i, j \in \mathcal{C}$, we have*

$$|A^*_{ij}(a) - A^*_{ij}(a')| \leq \gamma_C \cdot (a - a'),$$

where $\gamma_C := \frac{2(C-1)}{(C-2)}$, and for all $i \in \mathcal{C}$ and all $j_{1:k} \in \mathcal{C} \setminus \{i\}$,

$$|\psi(A^*_{ij_1}(a) - A^*_{ii}(a), \ldots, A^*_{ij_k}(a) - A^*_{ii}(a)) - \psi(A^*_{ij_1}(a') - A^*_{ii}(a'), \ldots, A^*_{ij_k}(a') - A^*_{ii}(a'))| \leq \gamma_C\,\Delta_2\sqrt{k}(a - a'),$$

where ψ and Δ_2 are as in Lemma 9.

Proof For all $i = j \in \mathcal{C}$, $A^*_{ii}(a) = 1$, a constant, irrespective of the value of $a \in [-1,1]$. Therefore, for all $a, a' \in [-1,1]$ such that $a' \leq a$, we have $|A^*_{ii}(a) - A^*_{ii}(a')| = 0 \leq \frac{2(C-1)}{(C-2)}(a - a') = \gamma_C \cdot (a - a')$. Note that $1 < \frac{\gamma_C}{2} = \frac{(C-1)}{(C-2)} < \infty$ since $C \geq 3$. Now consider any $i, j \in \mathcal{C}$ with $i \neq j$. If either $i = 1$ or $j = 1$, then for all $a \in [-1,1]$, $A^*_{ij}(a) = a$ and therefore $|A^*_{ij}(a) - A^*_{ij}(a')| = |a - a'| = (a - a') \leq \gamma_C \cdot (a - a')$. If $i \neq 1$ and $j \neq 1$ and $i \neq j$, then for all $a \in [-1,1]$, $A^*_{ij}(a) = b = (a^2(C-1)-1)/(C-2)$ and then,

$$|A^*_{ij}(a) - A^*_{ij}(a')| = \frac{|a^2 - (a')^2|(C-1)}{(C-2)} = \frac{(a + a')(a - a')(C-1)}{(C-2)} \leq \tfrac{2(C-1)}{(C-2)}(a - a') = \gamma_C \cdot (a - a').$$

This proves that for all $a' \leq a$ with $a, a' \in [-1,1]$, and all $i, j \in \mathcal{C}$, we have $|A^*_{ij}(a) - A^*_{ij}(a')| \leq \gamma_C \cdot (a - a')$. Next, for all $i \in \mathcal{C}$, all $j_{1:k} \in \mathcal{C} \setminus \{i\}$, and all $a \in [-1,1]$, let

$$v(a) := (A^*_{ij_1}(a) - A^*_{ii}(a), \ldots, A^*_{ij_k}(a) - A^*_{ii}(a))^\top = (A^*_{ij_1}(a) - 1, \ldots, A^*_{ij_k}(a) - 1)^\top.$$

Then, for all $a, a' \in [-1,1]$ with $a' \leq a$, the bound on $|A^*_{ij}(a) - A^*_{ij}(a')|$ that we just proved implies that

$$||v(a) - v(a')||_2 = \sqrt{\sum_{m=1}^{k} |A^*_{ij_m}(a) - A^*_{ij_m}(a')|^2} \leq \sqrt{\sum_{m=1}^{k} (\gamma_C \cdot\ (a - a'))^2} = \gamma_C\,\sqrt{k}\ (a - a').$$

Therefore, from Lemma 9, we get

$$|\psi(v(a)) - \psi(v(a'))| \leq \Delta_2 ||v(a) - v(a')||_2 \leq \gamma_C\,\sqrt{k}\,\Delta_2(a - a').$$

■

A.16 Proof of Theorem 4

The proof makes use of the results in Lemma 8, Lemma 9, and Lemma 10 which appear in Appendix A.13, Appendix A.14, and Appendix A.15, respectively.

Proof Let $\mathcal{E}_{1\bar{1}} := \{y = 1 \text{ and for some } i,\ y_i^- \neq 1\}$, $\mathcal{E}_{\bar{1}1} := \{y \neq 1 \text{ and for all } i,\ y_i^- = 1\}$, and $\mathcal{E}_1 := \mathcal{E}_{1\bar{1}} \cup \mathcal{E}_{\bar{1}1}$. Let $\mathcal{E}_= := \{y = y_1^- = \ldots = y_k^-\}$ and $\mathcal{E}_2 := (\mathcal{E}_= \cup \mathcal{E}_1)^c$. Then, $\mathcal{E}_=$, $\mathcal{E}_1$, and $\mathcal{E}_2$ are mutually exclusive and exhaustive events with

$$\Pr(\mathcal{E}_=) = \lambda_1^{k+1} + (C-1)\,\frac{(1-\lambda_1)^{k+1}}{(C-1)^{k+1}} = \lambda_1^{k+1} + \frac{(1-\lambda_1)^{k+1}}{(C-1)^k}$$

$$\Pr(\mathcal{E}_1) = \Pr(\mathcal{E}_{1\bar{1}}) + \Pr(\mathcal{E}_{\bar{1}1}) = \lambda_1(1-\lambda_1)^k + (1-\lambda_1)\lambda_1^k = \lambda_1\,(1-\lambda_1)\left(\frac{(1-\lambda_1)^k}{1-\lambda_1} + \lambda_1^{k-1}\right), \tag{52}$$

$$\Pr(\mathcal{E}_2) = 1 - \Pr(\mathcal{E}_=) - \Pr(\mathcal{E}1) = (1-\lambda_1)^2\left(\frac{(1-\lambda_1)^k}{1-\lambda_1} - \frac{(1-\lambda_1)^{k-1}}{(C-1)^k}\right),$$

$$\frac{\Pr(\mathcal{E}_1)}{\Pr(\mathcal{E}_2)} \geq \frac{\lambda_1}{1-\lambda_1}. \tag{53}$$

Next, noting the definition of ϕ in Lemma 9 and that for all $j \in \mathcal{C}$, $(A^*_{1j} - 1) = (A^*_{j1} - 1) = (a-1)\,1(j \neq 1)$, we have $\forall\,(y, y^-_{1:k}) \in \mathcal{E}_= \uplus\, \mathcal{E}_1$,

$$\psi(A^*_{yy^-_1}(a) - 1, \ldots, A^*_{yy^-_1}(a) - 1) = \psi((a-1)\,u(y, y^-_{1:k})) = \phi_{u(y,y^-_{1:k})}(a-1), \tag{54}$$

and in particular for all $(y, y^-_{1:k}) \in \mathcal{E}_=$, $u(y, y^-_{1:k}) = \mathbf{0}$ and $\phi_{u(y,y^-_{1:k})}(a-1) = \psi(0, \ldots, 0) = \phi_{\mathbf{0}}(0)$, a constant. We also note that for all $(y, y^-_{1:k}) \in \mathcal{E}_1$, $u(y, y^-_{1:k}) \neq \mathbf{0}$. For all $a, a' \in [-1, 1]$ with $a' < a$ and $y, y^-_{1:k}$ distributed as in (3), we have

$$\begin{aligned} S(A^*(a)) &= \mathbb{E}\Big[\psi(A^*_{yy^-_1}(a) - 1, \ldots, A^*_{yy^-_1}(a) - 1)\Big] \\ &= \Pr(\mathcal{E}_=)\,\phi_{\mathbf{0}}(0) + \Pr(\mathcal{E}_1)\,\mathbb{E}[\phi_{u(y,y^-_{1:k})}(a-1)|\mathcal{E}_1] \quad + \\ &\qquad \Pr(\mathcal{E}_2)\,\mathbb{E}\Big[\psi(A^*_{yy^-_1}(a) - 1, \ldots, A^*_{yy^-_1}(a) - 1)\Big|\mathcal{E}_2\Big]. \end{aligned}$$

Therefore,

$$\begin{aligned} &S(A^*(a)) - S(A^*(a')) \\ &= \Pr(\mathcal{E}_1)\,\mathbb{E}\Big[\phi_{u(y,y^-_{1:k})}(a-1) - \phi_{u(y,y^-_{1:k})}(a'-1)|\mathcal{E}_1\Big] \quad + \\ &\quad \Pr(\mathcal{E}_2)\,\mathbb{E}\Big[\psi(A^*_{yy^-_1}(a) - 1, \ldots, A^*_{yy^-_1}(a) - 1) - \psi(A^*_{yy^-_1}(a') - 1, \ldots, A^*_{yy^-_1}(a') - 1)\Big|\mathcal{E}_2\Big] \\ &\geq \Pr(\mathcal{E}_1)\,\mathbb{E}\big[\delta_{u(y,y^-_{1:k})}\,(a-a')|\mathcal{E}_1\big] - \Pr(\mathcal{E}_2)\,\gamma_C\,\sqrt{k}\,\Delta_2\,(a-a') \qquad (55) \\ &= (a-a')\,\Pr(\mathcal{E}_2)\,\left\{\frac{\Pr(\mathcal{E}_1)}{\Pr(\mathcal{E}_2)}\,\mathbb{E}\big[\delta_{u(y,y^-_{1:k})}|\mathcal{E}_1\big] - \gamma_C\sqrt{k}\Delta_2\right\} \\ &\geq (a-a')\,\frac{\Pr(\mathcal{E}_2)}{1-\lambda_1}\,\left\{\lambda_1\,\mathbb{E}\big[\delta_{u(y,y^-_{1:k})}|\mathcal{E}_1\big] - (1-\lambda_1)\,\gamma_C\sqrt{k}\Delta_2\right\} \qquad (56) \\ &= (a-a')\,\frac{\gamma_C\sqrt{k}\Delta_2\,\Pr(\mathcal{E}_2)}{1-\lambda_1}\,\left\{\lambda_1\left(1 + \frac{1}{\gamma_C\sqrt{k}\Delta_2}\mathbb{E}\big[\delta_{u(y,y^-_{1:k})}|\mathcal{E}_1\big]\right) - 1\right\} \\ &\geq 0. \qquad (57) \end{aligned}$$

Inequality (55) follows from (51) and Lemma 10 together with the fact that $s \geq -|s|$ for all $s \in \mathbb{R}$. Inequality (56) follows from (53). Inequality (57) follows from condition (18) and the assumption that $a' < a$.

Thus, if condition (18) is satisfied, then for all $a \in [-1, 1]$, $S(A^*(a))$ is a strictly increasing function of the variable a and is minimized when $a = -1$. When $a = -1$, $b = (a^2(C-1)-1)/(C-2) = 1$. Then, $\forall i \in \mathcal{C} \setminus \{1\}$, $(\mu^*_i)^\top \mu^*_1 = a = -1$. Since for all $j \in \mathcal{C}$ we have $||\mu^*_j||^2 = 1$, it follows from the alignment conditions for equality in the Cauchy-Schwartz inequality that for all $i \in \mathcal{C} \setminus \{1\}$, $\mu^*_i = -\mu^*_1$. Finally, if condition (19) is satisfied, then condition (18) is also satisfied because

$$\lambda_1 \geq \tau \Rightarrow \lambda_1 \geq \tfrac{1}{1+\frac{\delta_*}{\gamma_C\,\sqrt{k}\,\Delta_2}} \geq \tfrac{1}{1+\frac{1}{\gamma_C\,\sqrt{k}\,\Delta_2}\mathbb{E}\left[\delta_{u(y,y^-_{1:k})}|\mathcal{E}_1\right]}$$

and the last inequality holds because for all $(y, y^-_{1:k}) \in \mathcal{E}_1$, $u(y, y^-_{1:k}) \neq \mathbf{0}$, and by the definition of δ_* in (17), for all $u \neq \mathbf{0}$, $\delta_u \geq \delta_* > 0$. ■

A.17 Proof of Corollary 6

Proof From (18), a sufficient condition for minority collapse is given by

$$\lambda_1 \geq \frac{1}{1+\frac{1}{\gamma_C\sqrt{k}\Delta_2}\,\mathbb{E}[\delta_{u(y,y^-_{1:k})}|\mathcal{E}_1]}.$$

For the InfoNCE loss function, we will show that $\Delta_2 = 1/(2\sqrt{k})$ and develop a lower bound for $\mathbb{E}\big[\delta_{u(y,y^-_{1:k})}|\mathcal{E}_1\big]$ which is independent of k. This would yield a sufficient threshold for minority collapse. For the InfoNCE loss function,

$$\psi(t_{1:k}) = \log\Big(1+\tfrac{1}{k}\sum_{i=1}^{k} e^{t_i}\Big) \Rightarrow \nabla\psi^\top(t_{1:k}) = \tfrac{1}{k+\sum_{i=1}^{k} e^{t_i}}(e^{t_1},\ldots,e^{t_k})$$
$$\Rightarrow ||\nabla\psi(t_{1:k})||_2 = \sqrt{\frac{\sum_{i=1}^{k}(e^{t_i})^2}{\big(k+\sum_{i=1}^{k} e^{t_i}\big)^2}}.$$

For all $v_{1:k} \in \mathbb{R}$ we have

$$0 \leq \Big(k-\sum_{i=1}^{k} v_i\Big)^2 \Rightarrow 2k\Big(\sum_{i=1}^{k} v_i\Big) \leq k^2 + \Big(\sum_{i=1}^{k} v_i\Big)^2 \Rightarrow 4k\Big(\sum_{i=1}^{k} v_i\Big) \leq \Big(k+\sum_{i=1}^{k} v_i\Big)^2.$$

Therefore, for all $v_{1:k} \in [0,1]$,

$$4k\Big(\sum_{i=1}^{k} v_i^2\Big) \leq 4k\Big(\sum_{i=1}^{k} v_i\Big) \leq \Big(k+\sum_{i=1}^{k} v_i\Big)^2 \Rightarrow \sqrt{\frac{\sum_{i=1}^{k} v_i^2}{\big(k+\sum_{i=1}^{k} v_i\big)^2}} \leq \frac{1}{2\sqrt{k}}$$

with equality if, and only if, $\forall i, v_i = 1$. Thus, for all $t_{1:k} \in [-2,0]$, with $v_i := e^{t_i} \in [e^{-2},1]$, we get

$$\Delta_2 = \sup_{t_{1:k}\in[-2,0]} ||\nabla\psi(t_{1:k})||_2 = \sup_{v_{1:k}\in[e^{-2},1]} \sqrt{\frac{\sum_{i=1}^{k} v_i^2}{\big(k+\sum_{i=1}^{k} v_i\big)^2}} = \frac{1}{2\sqrt{k}} \Rightarrow \gamma_C\,\sqrt{k}\,\Delta_2 = \tfrac{1}{2}\gamma_C.$$

Thus, a sufficient condition for minority collapse is given by

$$\lambda_1 \geq \frac{1}{1+\frac{2}{\gamma_C}\,\mathbb{E}\big[\delta_{u(y,y^-_{1:k})}|\mathcal{E}_1\big]}.$$

We will now develop a lower bound for $\mathbb{E}\big[\delta_{u(y,y^-_{1:k})}|\mathcal{E}_1\big]$ which is independent of k. By Lemma 9, for all $u \in \{0,1\}^k$,

$$\begin{aligned}
\delta_u &= u^\top\,\nabla\psi(-2\,u)\\
&= \sum_{i=1}^{k} u_i \frac{e^{-2u_i}}{k+\sum_{j=1}^{k} e^{-2u_j}}\\
&= \frac{e^{-2}||u||_1}{k+e^{-2}||u||_1+(k-||u||_1)}\\
&= \frac{||u||_1}{2ke^2-(e^2-1)||u||_1}\\
&=: g(||u||_1),
\end{aligned}$$

and we note that $\delta_u = g(||u||_1)$ is an increasing function of $||u||_1$. From Theorem 4, $\mathcal{E}_1 := \mathcal{E}_{1\bar{1}} \uplus \mathcal{E}_{\bar{1}1}$, for all $(y, y^-_{1:k}) \in \mathcal{E}_1$, $u(y, y^-_{1:k}) \neq \mathbf{0}$, and for all $(y, y^-_{1:k}) \in \mathcal{E}_{1\bar{1}}$, $y = 1$ and $(y^-_{1:k}) \neq (1,\ldots,1)$. Moreover, from (52),

$$\Pr(\mathcal{E}_1) = \lambda_1\,(1-\lambda_1)^k + \lambda_1^k\,(1-\lambda_1) \leq \lambda_1\,(1-\lambda_1) + \lambda_1\,(1-\lambda_1) = 2\lambda_1\,(1-\lambda_1) \leq \tfrac{1}{2}.$$

Therefore,

$$
\begin{aligned}
&2\lambda_1\,(1-\lambda_1)\,\mathbb{E}\big[\delta_{u(y,y^-_{1:k})}|\mathcal{E}_1\big] \\
&\quad\geq \Pr(\mathcal{E}_1)\,\mathbb{E}\big[\delta_{u(y,y^-_{1:k})}|\mathcal{E}_1\big]\,, \\
&\quad= \sum_{(y,y^-_{1:k})\in\mathcal{E}_1} p(y,y^-_{1:k})\,\delta_{u(y,y^-_{1:k})}\,, \\
&\quad\geq \sum_{(y,y^-_{1:k})\in\mathcal{E}_{1\bar{1}}} p(y,y^-_{1:k})\,\delta_{u(y,y^-_{1:k})}\,, \\
&\quad= \sum_{(y,y^-_{1:k})\in\mathcal{E}_{1\bar{1}}} \lambda_1\Big(\prod_{i=1}^{k}\lambda_{y^-_i}\Big)\,\delta_{u(y,y^-_{1:k})}\,, \\
&\quad= \lambda_1 \sum_{(y^-_{1:k})\in\mathcal{C}^k\setminus\{\mathbf{1}\}} \lambda_1^{k-||u(1,y^-_{1:k})||_1}\Big(\frac{1-\lambda_1}{C-1}\Big)^{||u(1,y^-_{1:k})||_1}\,g(||u(1,y^-_{1:k})||_1)\,, \\
&\quad= \lambda_1 \sum_{w\in\{0,1\}^k\setminus\{\mathbf{0}\}}\ \sum_{y^-_{1:k}:u(1,y^-_{1:k})=w} \lambda_1^{k-||w||_1}\Big(\frac{1-\lambda_1}{C-1}\Big)^{||w||_1}\,g(||w||_1)\,, \\
&\quad= \lambda_1 \sum_{w\in\{0,1\}^k\setminus\{\mathbf{0}\}} (C-1)^{||w||_1}\,\lambda_1^{k-||w||_1}\Big(\frac{1-\lambda_1}{C-1}\Big)^{||w||_1}\,g(||w||_1)\,, \\
&\quad= \lambda_1 \sum_{w\in\{0,1\}^k\setminus\{\mathbf{0}\}} \lambda_1^{k-||w||_1}\,(1-\lambda_1)^{||w||_1}\,g(||w||_1)\,, \\
&\quad= \lambda_1 \sum_{l=1}^{k}\binom{k}{l}\lambda_1^{k-l}\,(1-\lambda_1)^l\,g(l)\,, \\
&\quad= \lambda_1 \sum_{l=0}^{k}\binom{k}{l}\lambda_1^{k-l}\,(1-\lambda_1)^l\,g(l), \quad \text{since } g(0)=0\,, \\
&\quad= \lambda_1\,\mathbb{E}[g(l)], \quad l\sim\text{Binomial}(k,1-\lambda_1)\,, \\
&\quad\geq \lambda_1\,\mathbb{E}[1(l\geq k/2)\,g(l)], \quad l\sim\text{Binomial}(k,1-\lambda_1)\,, \quad \text{since } \forall l, g(l)\geq 0, \\
&\quad\geq \lambda_1\,\mathbb{E}[1(l\geq k/2)\,g(k/2)], \quad l\sim\text{Binomial}(k,1-\lambda_1)\,, \quad \text{since } g(l) \text{ increases with } l, \\
&\quad= \lambda_1\,g(k/2)\,\Pr(l\geq k/2), \quad l\sim\text{Binomial}(k,1-\lambda_1)\,, \\
&\quad\geq \lambda_1\,\frac{1}{2}\,g(k/2),
\end{aligned}
$$

$$
\begin{aligned}
\Rightarrow \mathbb{E}\big[\delta_{u(y,y^-_{1:k})}|\mathcal{E}_1\big] &\geq \frac{1}{4(1-\lambda_1)}\,g(k/2) \\
&= \frac{1}{4(1+3e^2)(1-\lambda_1)}
\end{aligned}
$$

Therefore, a sufficient condition for minority collapse is given by

$$
\begin{aligned}
\lambda_1 \geq \frac{1}{1+\frac{2}{\gamma_C}\frac{1}{4(1+3e^2)(1-\lambda_1)}} = \frac{1}{1+\frac{1}{2\,\gamma_C\,(1+3e^2)(1-\lambda_1)}} &\Rightarrow 0\geq \lambda_1^2 - 2\frac{\lambda_1}{\beta_C} + 1 \\
&\Rightarrow \lambda_1 \geq \frac{1-\sqrt{1-\beta_C^2}}{\beta_C} =: \tau_C,
\end{aligned}
$$

where $\beta_C := \frac{1}{1+\frac{1}{4\gamma_C(1+3e^2)}}$. We note that $\tau_C\in(0,1)$ since $\beta_C\in(0,1)$.
Since $(C-1)=(C-2)+1$ and $C\geq 3$, we have $\gamma_C=\frac{2(C-1)}{(C-2)}\in(2,4]$. The most conservative (maximum)

value of τ_C occurs when β_C is maximum (since τ_C is an increasing function of β_C) which occurs when γ_C is maximum (since β_C is an increasing function of γ_C), which occurs when C is minimum, i.e., $C = 3$. When $C = 3$, $\gamma_C = 4, \beta_C \approx 0.9973$, and $\tau_C \approx 0.9292$. Thus, $\lambda_1 \in \Big(0.9292, 1\Big)$ is a sufficient condition for minority collapse for the InfoNCE loss, which holds for all $C \geq 3$ and all k. ■

A.18 Extension of theoretical results to the SCL setting

In the SCL setting, for all $y \in \mathcal{C}$ we have $y^-_{1:k} \in \mathcal{C} \setminus \{y\}$ w.p.1,

$$p_{SCL}(y, y^-_{1:k}) := \lambda_y \prod_{t=1}^{k} \left(\frac{\lambda_{y^-_t}}{1 - \lambda_y} \right), \tag{58}$$

and for all $x, x^+ \in \mathcal{X}, y \in \mathcal{C}$,

$$q(x, x^+|y) = s(x|y)\, s(x^+|y).$$

With the above changes, all results in Section 3, Section 4, and Section 5 hold with all summations $\sum_{y, y^-_{1:k} \in \mathcal{C}}$ replaced by $\sum_{y \in \mathcal{C}, y^-_{1:k} \in \mathcal{C} \setminus \{y\}}$ and all products $\prod_{t=1}^{k} \lambda_{y^-_t}$ replaced by $\prod_{t=1}^{k} \frac{\lambda_{y^-_t}}{1-\lambda_y}$. With these changes, the proofs of all results in Section 3, Section 4, and Section 5 go through in a straightforward manner with the exception of the proof of necessity of within-class variance collapse in Lemma 1 which requires additional elaboration.

As in the proof of the UCL setting, equality in (29) can be attained only if $\forall i \in \mathcal{C}$, w.p.1 given $y = i$, we must have $f(x) = f(x^+)$ and $||f(x)|| = 1$ (here, all the weights $\lambda_y \prod_{t=1}^{k} \left(\frac{\lambda_{y^-_t}}{1-\lambda_y} \right)$ are *strictly* positive). In the SCL setting, x, x^+ are conditionally iid with distribution $s(\cdot|j)$ given $y = j$. From Lemma 7 in Appendix A.2, it then follows that for all $j \in \mathcal{C}$, w.p.1 given $y = j$, $f(x) = \mu_j$, or more compactly, $\forall x \in \mathcal{X}, f(x) = \mu_{y(x)}$ completing the proof of necessity in the SCL setting.

The proofs of all subsequent results in Section 4 and Section 5 go through straightforwardly since they only make use of the lower bound $G(M)$ in Lemma 1.

In the SCL setting, Lemma 8, Lemma 9, and Lemma 10 in Section 6 and their proofs in the appendices hold without any changes. However, Theorem 4 and Corollary 1 and their proofs change slightly in the SCL setting as described below.

Theorem 5 (Sufficient conditions for minority collapse in the SCL setting)**.** *Let $C, \lambda_{1:C}$ be as in Lemma 8, $S(\cdot)$ be as in (14), Δ_2, ϕ, and δ_u be as in Lemma 9 and let $a, b, A^*(a)$, and γ_C be as in Lemma 10. With $(y, y^-_{1:k})$ distributed as in (5), if*

$$\lambda_1 \geq \tau := \frac{1}{1 + \frac{1}{\gamma_C \sqrt{k} \Delta_2} \delta_{\mathbf{1}}} \in (0, 1), \tag{59}$$

where $\mathbf{1}$ is the $C \times 1$ vector of all ones, then for all $a \in [-1, 1]$, $S(A^*(a))$ is a strictly increasing function of the variable a and is minimized when $a = -1 \Rightarrow b = 1$ and then for all $i \in \mathcal{C} \setminus \{1\}$, $\mu^*_i = -\mu^*_1$ with $||\mu^*_1|| = 1$, i.e., we have minority collapse.

Proof For all $a, a' \in [-1, 1]$ with $a' < a$ and $(y, y^-_{1:k})$ distributed as in (5) we have

$$\begin{aligned}
S(A^*(a)) &= \mathbb{E}\Big[\psi(A^*_{yy^-_1}(a) - A^*_{yy}(a), \ldots, A^*_{yy^-_1}(a) - A^*_{yy}(a))\Big] \\
&= \mathbb{E}\Big[\psi(A^*_{yy^-_1}(a) - 1, \ldots, A^*_{yy^-_1}(a) - 1)\Big] \\
&= \lambda_1\, \mathbb{E}\Big[\psi(A^*_{1y^-_1}(a) - 1, \ldots, A^*_{1y^-_1}(a) - 1)\Big| y = 1\Big] \quad +
\end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad (1-\lambda_1)\,\mathbb{E}\Big[\psi(A^*_{yy_1^-}(a)-1,\ldots,A^*_{yy_1^-}(a)-1)\Big|y\neq 1\Big]\\
&=\lambda_1\,\mathbb{E}\Big[\psi(a-1,\ldots,a-1)\Big|y=1\Big]\quad+\\
&\qquad\qquad (1-\lambda_1)\,\mathbb{E}\Big[\psi(A^*_{yy_1^-}(a)-1,\ldots,A^*_{yy_1^-}(a)-1)\Big|y\neq 1\Big]\\
&=\lambda_1\,\phi_{\mathbf{1}}(a-1)+(1-\lambda_1)\,\mathbb{E}\Big[\psi(A^*_{yy_1^-}(a)-1,\ldots,A^*_{yy_1^-}(a)-1)\Big|y\neq 1\Big],
\end{aligned}
\tag{60}
$$

where the second and third equalities are because all the diagonal entries of the matrix $A^*(a)$ in Equation (43) of Lemma 8 are equal to one and if $y=1$ then for all $m\in\{1:k\}$, $y_m^-\neq 1$ which would imply that $A^*_{yy_m^-}(a)=A^*_{1y_m^-}(a)=a$. Equation (60) follows from the definition of ϕ_u in Lemma 9. Therefore,

$$
\begin{aligned}
&S(A^*(a))-S(A^*(a'))\\
&\quad=\lambda_1\left(\phi_{\mathbf{1}}(a-1)-\phi_{\mathbf{1}}(a'-1)\right)\;+(1-\lambda_1)\,\mathbb{E}\Big[\psi(A^*_{yy_1^-}(a)-1,\ldots,A^*_{yy_1^-}(a)-1)-\\
&\qquad\qquad\qquad\qquad\psi(A^*_{yy_1^-}(a')-1,\ldots,A^*_{yy_1^-}(a')-1)\Big|y\neq 1\Big]\\
&\quad\geq\lambda_1\left(\delta_{\mathbf{1}}\,(a-a')\right)-(1-\lambda_1)\,\gamma_C\,\sqrt{k}\,\Delta_2\,(a-a') \qquad (61)\\
&\quad=(a-a')\left(-\gamma_C\,\sqrt{k}\,\Delta_2+\lambda_1\left(\delta_{\mathbf{1}}+\gamma_C\,\sqrt{k}\,\Delta_2\right)\right)\\
&\quad>(a-a')\left(-\gamma_C\,\sqrt{k}\,\Delta_2+\gamma_C\,\sqrt{k}\,\Delta_2\right) \qquad (62)\\
&\quad=0,
\end{aligned}
$$

where (61) follows from (51) and Lemma 10 together with the fact that $s\geq -|s|$ for all $s\in\mathbb{R}$, and (62) follows from (59). Thus, for all $a\in[-1,1]$, $S(A^*(a))$ is a strictly increasing function of the variable a and is minimized when $a=-1$. When $a=-1$, $b=(a^2(C-1)-1)/(C-2)=1$. Then, $\forall i\in\mathcal{C}\setminus\{1\}$, $(\mu_i^*)^\top\mu_1^*=a=-1$. Since for all $j\in\mathcal{C}$ we have $||\mu_j^*||^2=1$, it follows from the alignment conditions for equality in the Cauchy-Schwartz inequality that for all $i\in\mathcal{C}\setminus\{1\}$, $\mu_i^*=-\mu_1^*$. ■

Corollary 7. *For the InfoNCE loss function, condition (59) for minority collapse in Theorem 5 is satisfied if*

$$
\lambda_1\in[\tau_C,1),\ \textit{where}\ \tau_C:=\frac{1}{1+\frac{2}{\gamma_C(1+e^2)}}.
$$

Moreover, for all $C\geq 3$, $\tau_C\leq\tau_3\approx 0.9438$. Thus, $\lambda_1\geq 0.9438$ is a sufficient condition for minority collapse in the SCL setting for the InfoNCE loss function, irrespective of the number of classes C or the number of negative samples per anchor sample k.

Proof As in the proof of Corollary 6 in Appendix A.17,

$$
\Delta_2=\frac{1}{2\sqrt{2}}.
$$

Moreover,

$$
\delta_{\mathbf{1}}=\mathbf{1}^\top\nabla\psi(-2\,\mathbf{1})=\frac{ke^{-2}}{k+ke^{-2}}=\frac{1}{1+e^2}.
$$

Plugging these into (59) we get

$$
\tau=\tau_C:=\frac{1}{1+\frac{2}{\gamma_C}\frac{1}{1+e^2}}.
$$

The most conservative (maximum) value of γ_C occurs when C is minimum, i.e., $C=3$. When $C=3$, $\gamma_C=4$, and $\tau_C\approx 0.9438$. ■