Neural and Minority Collapse in Contrastive Learning with Imbalanced Datasets

Thuan Nguyen *NGUYENT11@mail.etsu.edu*
Department of Engineering Technology
East Tennessee State University

Shuchin Aeron *shuchin@ece.tufts.edu*
Department of Electrical and Computer Engineering
Tufts University

Donald Brown *drb@wpi.edu*
Department of Electrical and Computer Engineering
Worcester Polytechnic Institute

Prakash Ishwar *pi@bu.edu*
Department of Electrical and Computer Engineering
Boston University

Abstract

In this paper we provide a computable characterization of the geometry of optimal representations in Contrastive Learning (CL) when the classes are imbalanced. For balanced 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 we prove that the optimal representations of all samples from the same class collapse to their class means and their geometry can be determined by solving a convex optimization problem. We further investigate a special case when 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. Both numerical and theoretical results are provided to illustrate these results.

1 Introduction

CL is a machine learning technique that aims to learn good features (latent representations) by pulling "similar" samples closer together while simultaneously pushing apart "different" samples in a representation space. Over the past decade, CL has received significant attention due to its applications ranging from computer vision, time series analysis, text classification, and natural language processing, just to name a few (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. (2022; 2023), or (b) any class (including the anchor's class) Khosla et al. (2020).

CL learns useful features that can be used in downstream tasks, e.g., classification, by minimizing a loss that encourages the anchor and positive features to cluster close together while also pushing the features of the

anchor and negative samples further apart Jaiswal et al. (2020). Features are typically learned through a deep neural network without a classifier layer (even in SCL) and a training loss that only quantifies average or expected similarity/dissimilarity between samples.

1.1 Limitations of Related work and Contributions

Loss function, sampling distribution, and range of 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 $n_b - 2$).

Class proportions: In addition to heavily focusing on the empirical InfoNCE CL loss together with the severely restricted 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. (2023a) and empirical cross-entropy loss Hong & Ling (2024); Dang et al. (2024a) 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).

This paper makes the following contributions:

1. We construct a lower bound for the general family of CL losses that was recently proposed in Jiang et al. (2023). This family consists of losses based on loss functions that are strictly convex and argument-wise strictly increasing. This subsumes and generalizes popular loss functions with spherical-ball normalized representations including 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 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).

2. Using the lower bound, we 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 3).

3. We establish certain uniqueness and symmetry properties of the optimal geometry some under additional conditions (Lemma 5 and Corollary 4) and show that these properties are consistent with corresponding results for balanced classes (Remark 6).

4. We further investigate the case when the class imbalance is extreme and prove that CL exhibits the so-called Minority Collapse (MC) phenomenon in the scenario where there is one majority class and equiprobable minority classes with the minor class probability less than a threshold 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 6 – 8 and Theorem 3).

The remainder of this paper is structured as follows. Section 2 formally introduces the CL framework and formulates the core optimization problem of interest. Our first main result, namely a lower bound for the contrastive loss that is a function of the mean feature vectors of the classes, is established in Section 3. Our second main contribution which shows that the lower bound is a strictly convex function of the Gram

matrix whose entries are the pairwise inner products of class mean feature vectors is established in Section 4. This immediately leads to an efficient method for computing the optimal mean feature vectors. The MC phenomenon is investigated in Section 6. We provide numerical experiments that corroborate and illustrate our theoretical results in Section 7. We end with concluding remarks and a discussion of open questions in Section 8.

2 Contrastive learning problem setup and notation

We adopt the problem setup and notation used in (Jiang et al., 2023), but customize it to the specific settings germane to our work. Let $\mathcal{X} \in \mathbb{R}^{d'}$ denote the data space, $f : \mathcal{X} \to \mathcal{Z}$ a representation function from data space to representation space (or feature space) $\mathcal{Z} \in \mathbb{R}^d$, and $\mathcal{F}$ a family of such representation functions. CL seeks to select a representation function from $\mathcal{F}$ by minimizing the expected value or the sample average of a loss that encourages a large inner product between the representation of the anchor $z = f(x)$ and the representation of a positive sample $z^+ = f(x^+)$. At the same time, the loss encourages a small inner product between the representation of the anchor $z = f(x)$ and the representations of k negative samples $z_i^- = f(x_i^-)$, $i = 1, 2, \ldots, k$.[1] The anchor x is also regarded as a positive sample and (x, x^+) is called a positive pair. To simplify the 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. 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 paper, we focus on the following general family of CL losses proposed in (Jiang et al., 2023).

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

$$\ell(z, z^+, z_{1:k}^-) := \psi\big(z^\top(z_1^- - z^+), \ldots, z^\top(z_k^- - z^+)\big), \tag{1}$$

where $\psi : \mathbb{R}^k \to \mathbb{R}$ is loss 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]

As noted in (Jiang et al., 2023), the family $\mathcal{F}$ includes popular functions such as InfoNCE loss which is defined by:

$$\begin{aligned}\ell_{\text{InfoNCE}}(z, z^+, z_{1:k}^-) &= -\log\left(\frac{e^{z^\top z^+}}{e^{z^\top z^+} + \frac{1}{k}\sum_{j=1}^k e^{z^\top z_j^-}}\right) = \log\left(1 + \frac{1}{k}\sum_{j=1}^k e^{z^\top(z_j^- - z^+)}\right) \\ &= \psi\big(z^\top(z_1^- - z^+), \ldots, z^\top(z_k^- - z^+)\big),\end{aligned} \tag{2}$$

where $\psi(t_{1:k}) = \log(1 + \frac{1}{k}\sum_{i=1}^k e^{t_i})$. The InfoNCE loss function $\psi(t_{1:k}) = \psi_{\text{InfoNCE}}(t_{1:k}) := \log(1 + \frac{1}{k}\sum_{i=1}^k e^{t_i})$ is clearly argument-wise strictly increasing. Being a log-sum-exponential with a positive offset within the logarithm, it is not only convex, but strictly convex (see Appendix A.1 for a short proof based on Hölder's inequality).

As in (Jiang et al., 2023), we define the contrastive *risk* $L(f)$ of a representation function f as the expected value (or empirical average) of the contrastive loss across tuples of positive and negative samples, i.e.,

$$L(f) := \mathbb{E}\Big[\ell(f(x), f(x^+), f(x_1^-), \ldots, f(x_k^-))\Big]. \tag{3}$$

Unlike prior works which are restricted to the empirical CL loss 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 loss (which

[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.

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

subsumes the empirical loss as a special case) 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. Then the joint distribution of all $(k+2)$ samples $x, x^+, x^-_{1:k}$ and their $(k+1)$ labels $y, y^-_{1:k}$ is assumed to have the following form

$$p(x, x^+, x^-_{1:k}, y, y^-_{1:k}) = p(y, y^-_{1:k})\, p(x, x^+, x^-_{1:k} | y, y^-_{1:k}) \ ,$$

$$p(y, y^-_{1:k}) = \lambda_y \prod_{t=1}^{k} \lambda_{y^-_t} \ , \tag{4}$$

$$p(x, x^+, x^-_{1:k} | y, y^-_{1:k}) = q(x, x^+ | y) \prod_{t=1}^{k} s(x^-_t | y^-_t), \tag{5}$$

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., 2023), we do not assume that (x, x^+) are conditionally independent given their label y.

A4: Marginal conditional distributions. As in (Jiang et al., 2023) 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 $y^- = i$. As explained in (Jiang et al., 2023), this can be ensured in the supervised setting, since labels are available, and also 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. 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 $C \times d$ matrix of class means in representation space, specifically,

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

A5: Spherical-ball normalized representations. All prior theoretical studies of CL impose restrictions on 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 is accomplished either directly by explicitly requiring all representation maps in $\mathcal{F}$ to be norm-bounded, or indirectly 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 primarily adopt the direct approach by requiring all representation functions to have a 2-norm less than or equal to one, 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., $\|f(x)\|^2 \leq 1, \forall x \in \mathcal{X}$, from the Cauchy-Schwarz inequality (or alternatively by the convexity of the squared norm function $\|\cdot\|^2$), we must have

$$\forall j \in \mathcal{C},\ \|\mu_j\|^2 = \|\mathbb{E}[f(x^-)|y^- = j]\|^2 = \|\mathbb{E}[f(x^+)|y = j]\|^2 = \|\mathbb{E}[f(x)|y = j]\|^2 \leq \mathbb{E}[\|f(x)\|^2|y = j] \leq 1. \quad (7)$$

The CL problem seeks to find a representation function $f \in \mathcal{F}$ that minimizes the contrastive risk, i.e., solves the following optimization problem

$$\min_{f \in \mathcal{F}} L(f) \quad (8)$$

In practice, the family of representation functions is further constrained to be representable by a neural network having a specific architecture. The optimal solutions of the optimization problem in (8) 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.

A6: Unconstrained Features Model (UFM). 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 generalized class of 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 (8) 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. (2023); Graf et al. (2021); 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 contrastive risk in terms of class means

We first show that it is possible to lower bound the contrastive risk by a function of the class means.

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

$$L(f) \geq G(M),$$
$$G(M) := \sum_{i, j_{1:k} \in \mathcal{C}} \left(\lambda_i \prod_{t=1}^{k} \lambda_{j_t} \right) \psi\big(\mu_i^\top \mu_{j_1} - 1, \ldots, \mu_i^\top \mu_{j_k} - 1)\big). \quad (9)$$

The lower bound $G(M)$ can be attained if, and only if, 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

$$L(f) = \mathbb{E}\left[\ell\Big(f(x), f(x^+), f(x^-_1), \ldots, f(x^-_k)\Big)\right]$$
$$= \mathbb{E}\left[\psi\Big(f^\top(x)(f(x^-_1) - f(x^+)), \ldots, f^\top(x)(f(x^-_k) - f(x^+))\Big)\right] \quad (10)$$
$$= \mathbb{E}\left[\mathbb{E}\left[\psi\Big(f^\top(x)(f(x^-_1) - f(x^+)), \ldots, f^\top(x)(f(x^-_k) - f(x^+))\Big)\,\middle|\, y, y^-_1, \ldots, y^-_k\right]\right] \quad (11)$$

$$\geq \mathbb{E}\Big[\psi\Big(\mathbb{E}[f^\top(x)\, f(x_1^-)|y, y_1^-] - \mathbb{E}[f^\top(x)\, f(x^+)|y], \ldots, \mathbb{E}[f^\top(x)\, f(x_k^-)|y, y_k^-] - \mathbb{E}[f^\top(x)\, f(x^+)|y]\Big)\Big] \quad (12)$$

$$= \mathbb{E}\Big[\psi\Big(\mu_y^\top\, \mu_{y_1^-} - \mathbb{E}[f^\top(x)\, f(x^+)|y], \ldots, \mu_y^\top\, \mu_{y_k^-} - \mathbb{E}[f^\top(x)\, f(x^+)|y]\Big)\Big] \quad (13)$$

$$= \sum_{i, j_{1:k} \in \mathcal{C}} \left(\lambda_i \prod_{t=1}^{k} \lambda_{j_t}\right) \psi\big(\mu_i^\top \mu_{j_1} - \mathbb{E}[f^\top(x)\, f(x^+)|y = i], \ldots, \mu_i^\top \mu_{j_k} - \mathbb{E}[f^\top(x)\, f(x^+)|y = i]\big) \quad (14)$$

$$\geq \sum_{i, j_{1:k} \in \mathcal{C}} \left(\lambda_i \prod_{t=1}^{k} \lambda_{j_t}\right) \psi\big(\mu_i^\top \mu_{j_1} - 1, \ldots, \mu_i^\top \mu_{j_k} - 1\big), \quad (15)$$

where equality (10) follows from (1), equality (11) is the law of total expectation, inequality (12) is Jensen's inequality applied within the inner expectation conditioned on the labels of samples to the convex loss function $\psi(\cdot)$, (13) follows from the conditional independence of anchor and negative samples given their labels implied by (5), (14) follows by expanding the expectation in (13) in terms of all possible tuples of values of labels together with (4), and inequality (15) is because $\psi(\cdot)$ is an increasing function of all its arguments, all the weights $\big(\lambda_i \prod_{t=1}^{k} \lambda_{j_t}\big)$ 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 (12) and (15). Equality in (15) 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 $\big(\lambda_i \prod_{t=1}^{k} \lambda_{j_t}\big)$ 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 (12) 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 (15) 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 (15) and (12) 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 11 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 $L(f) = G(M)$. ■

Remark 1. *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 (15), 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 (15) alone without needing to analyze the conditions for equality in (12).*

Corollary 1. *For InfoNCE loss,*

$$G(M) = \sum_{i, j_{1:k} \in \mathcal{C}} \left(\lambda_i \prod_{t=1}^{k} \lambda_{j_t}\right) \log\left(1 + \frac{1}{k} \sum_{t=1}^{k} e^{\mu_i^\top \mu_{j_t} - 1}\right). \quad (16)$$

Proof This follows from (15) by replacing the generalized loss with the InfoNCE loss defined in (2). ■

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

Corollary 2. *For InfoNCE loss,*

$$\lim_{k\to\infty} L(f) = \mathbb{E}\left[\log\left(1 + \mathbb{E}\left[e^{f^\top(x)(f(x_1^-)-f(x^+))}\middle| x, x^+\right]\right)\right] \tag{17}$$

$$\geq \lim_{k\to\infty} G(M)$$

$$= \sum_{i\in\mathcal{C}} \lambda_i \log\left(1 + \sum_{j\in\mathcal{C}} r(j|i)\, e^{\mu_i^\top \mu_j - 1}\right). \tag{18}$$

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 (10), the definition of the InfoNCE loss in (1), and (12), (15), and (9) we have

$$L(f) = \mathbb{E}\left[\mathbb{E}\left[\log\left(1 + \frac{1}{k}\sum_{t=1}^{k} U_t(x, x^+, x_t^-)\right)\middle| x, x^+\right]\right], \tag{19}$$

$$G(M) = \mathbb{E}\left[\mathbb{E}\left[\log\left(1 + \frac{1}{k}\sum_{t=1}^{k} V_t(y, y_t^-)\right)\middle| y\right]\right]. \tag{20}$$

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.,

$$\frac{1}{k}\sum_{t=1}^{k} U_t \xrightarrow[k\to\infty]{\text{w.p.1}} \mathbb{E}[U_1|x, x^+] = \mathbb{E}\left[e^{f^\top(x)(f(x_1^-)-f(x^+))}\middle| x, x^+\right], \tag{21}$$

$$\frac{1}{k}\sum_{t=1}^{k} V_t \xrightarrow[k\to\infty]{\text{w.p.1}} \mathbb{E}[V_1|x, x^+] = \mathbb{E}\left[e^{(\mu_y^\top \mu_{y_1^-})-1}\middle| y\right] = \sum_{j\in\mathcal{C}_y} r(j|y)\, e^{(\mu_y^\top \mu_j)-1}. \tag{22}$$

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 (17) and (18) then follow from (19), (20), (21), (22), the Dominated Convergence Theorem, and the fact that $L(f) \geq G(M)$ proved in Lemma 1. ■

In this section, we showed that the contrastive risk can be lower bounded by a function of the class means in representation space and 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. We also showed 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 which minimize the lower bound.

4 Characterizing and computing optimal class means

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

$$\min_{M\in\mathcal{M}} G(M), \quad \text{where} \tag{23}$$

$$\mathcal{M} := \{M = [\mu_1 \cdots \mu_C]^\top \in \mathbb{R}^{C\times d} : \forall i \in \mathcal{C}, \|\mu_i\|^2 = 1\}. \tag{24}$$

A solution to (23) exists since the objective function $G(M)$ is continuous and the constraint set $\mathcal{M}$ is compact. However, neither is the objective function in (23) convex with respect to M nor is the constraint set defined in (24) 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 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 (9), $\psi(\cdot)$ 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 $M \in \mathcal{M} \cap \mathbb{R}^d_{\geq 0}$ if, and only if, $\mu_{1:C}$ are orthonormal. This requires $d \geq C$. ■

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

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

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 $\forall u \in \mathbb{R}^C, u^\top A u \geq 0$, and additionally, $\forall i \in \mathcal{C},\ A_{ii} = 1$ since $A_{ii} = ||\mu_i||^2 = 1$ to attain the lower bound in Lemma 1. Let

$$\mathcal{A}^* := \{A \in \mathbb{R}^{C\times C} : A = A^\top, A \succcurlyeq 0, \forall i \in \mathcal{C}, A_{ii} = 1\} \text{ and} \tag{25}$$

$$S(A) := \sum_{i, j_{1:k} \in \mathcal{C}} \left(\lambda_i \prod_{t=1}^k \lambda_{j_t}\right) \psi\big(A_{ij_1} - 1, \ldots, A_{ij_k} - 1\big). \tag{26}$$

Under certain conditions, a solution to (23) can be found by minimizing (26) over (25).

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

$$\min_{A \in \mathcal{A}^*} S(A). \tag{27}$$

If there exists an $M^* \in \mathcal{M}$ such that $M^*(M^*)^\top = A^*$, then M^* is solution to (23).

Proof If $A := MM^\top$, then clearly, $A = A^\top$ and $A \succcurlyeq 0$ (since for all $u \in \mathbb{R}^C, u^\top MM^\top u = ||M^\top u||^2 \geq 0$), and $\forall i \in \mathcal{C}, A_{ii} = ||\mu_i||^2 = 1$. Therefore $A = MM^\top \in \mathcal{A}^*$. From (23) and (26), which define $G(\cdot)$ and $S(\cdot)$ respectively, it follows that $G(M) = S(A) = S(MM^\top)$. Therefore, for any $M \in \mathcal{M}$ we have $G(M) = S(MM^\top) \geq S(A^*) = S(M^*(M^*)^\top) = G(M^*)$. This shows that M^* is a solution to (23). ■

We note that any $M \in \mathcal{M}$ can be mapped to an $A = MM^\top \in \mathcal{A}^*$. However, if $d < C$, it may not be possible to decompose all $A \in \mathcal{A}^*$ as $A = MM^\top$ 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 (27) is a convex optimization problem and has a unique solution $A^* \in \mathcal{A}^*$, i.e.,*

$$S(A^*) = \min_{A\in\mathcal{A}^*} S(A). \tag{28}$$

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 (26), for all $i, j_{1:k} \in \mathcal{C}$, the k-tuples $\big(A_{ij_1} - 1, \ldots, A_{ij_k} - 1\big)$ are linear functions of A and the weights $\big(\lambda_i \prod_{t=1}^k \lambda_{j_t}\big)$ 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 (26) are all strictly positive, it follows that $S(\cdot)$ is a strictly convex function over $\mathcal{A}^*$. ■

The next lemma proves that the unique solution to (27) is rank-deficient.

Lemma 4. *The unique solution $A^* \in \mathcal{A}^*$ to (27) has rank$(A^*) =: r \leq C - 1$. Therefore, the minimum eigenvalue of A^* is zero.*

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 (26) 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 $t = t' := \frac{\nu(A^*)}{2(C-1)}$. This would imply that $B(t') \in \mathcal{A}^*$ and contradict the optimality of A^*. By the Courant-Fischer min-max theorem,

$$\underline{\nu}(B(t)) = \min_{u\neq 0} \frac{u^\top B(t) u}{||u||^2} = \min_{u\neq 0} \frac{u^\top (A^* - t\,\mathbf{1}\mathbf{1}^\top + t\,I) u}{||u||^2} = \min_{u\neq 0} \frac{u^\top A^* u - t\,(u^\top \mathbf{1})^2 + t||u||^2}{||u||^2}$$

$$\geq \min_{u \neq 0} \frac{u^\top A^* u - t\, C||u||^2 + t||u||^2}{||u||^2} \tag{29}$$
$$= \min_{u \neq 0} \frac{u^\top A^* u}{||u||^2} - (C-1)t = \underline{\nu}(A^*) - (C-1)t,$$

where (29) 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 $\text{rank}(A^*) < C$. ■

Theorem 2 (Main Theorem 1)**.** *Let $A^* = U_r \Sigma_r U_r^\top$ be the unique solution to 27, 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 $M^* := [U_r\sqrt{\Sigma_r} \quad 0_{C \times d-r+1}]$ is a solution to (23), 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.*[3] *Moreover, $\forall i \in \mathcal{C}, \|\mu_i^*\|^2 = 1$ where μ_i^* (i^{th} column of $(M^*)^\top$) is an optimal class mean vector in representation space for class i.*

Proof Lemma 4 proved that (23) 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 eigendecomposition given by $A^* = U_r \Sigma_r U_r^\top$. For all $d \geq C-1$, the matrix $M^* := [U_r\sqrt{\Sigma_r} \quad 0_{C \times d-r+1}]$ is well defined and $M^*(M^*)^\top = U_r\sqrt{\Sigma}(\sqrt{\Sigma})^\top U_r^\top + 0_{C \times d-r+1} 0_{C \times d-r+1}^\top = U_r \Sigma_r U_r^\top = A^*$. From Lemma 2 it follows that M^* is a solution to (23). Moreover, for all $i \in \mathcal{C}$, we have $||\mu_i^*||^2 = A_{ii}^* = 1$. ■

Remark 2. *The solution A^* to (27) is unique. However, the solution to (23) is not unique. The M^* defined in Theorem 2 is just one solution to (23) when $d \geq C-1$. Still, when $d \geq C-1$, any solution $\hat{M}^*$ to (23) will also satisfy $\hat{M}^*(\hat{M}^*)^\top = A^*$ because $G(\hat{M}^*) = S(\hat{M}^*(\hat{M}^*)^\top) \geq S(M^*(M^*)^\top) = S(A^*)$ and A^* is the unique minimizer of $S(A)$ over $\mathcal{A}^*$.*

Remark 3. *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 $d \geq C-1$, $M^* = [\mu_1^*, \mu_2^*, \ldots, \mu_C^*]^\top \in \mathcal{M}$ be a solution to (23), and $A^* = M^*(M^*)^\top$ be the unique solution to (27). 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. ■

Remark 4. *The condition $C-1 \leq d$ is also required in many papers to show the NC phenomenon and the existence of the ETF-structure , e.g., Jiang et al. (2023); Graf et al. (2021); Wang & Palmer (2023); Dang et al. (2023b) . But they are all in the setting where classes are balanced, i.e., $\forall i \in \mathcal{C}$, $\lambda_i = 1/C$. In practice, it is quite reasonable to assume that $C-1 \leq d$ since the number of classes is usually much smaller than the dimension of the representation space, e.g., $d = 512$ in ResNet-18 compared to $C = 10$ in the CIFAR10 dataset and $C = 100$ in the CIFAR100 dataset.*

Remark 5. *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.*

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

5 Some equiangular properties of optimal class means

In this section we present certain equiangular properties of the optimal class means of classes that are equiprobable. These are consequences of the the uniqueness of A^*.

Lemma 5. *Suppose that there are two distinct classes i and j with the same probability, i.e., $\lambda_i = \lambda_j$. Let $M^* = [\mu_1^*, \mu_2^*, \ldots, \mu_C^*]^\top$ be an optimal mean vector matrix such that $M^* {M^*}^\top = A^*$. 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 $S(QQ^\top) = S(M^* {M^*}^\top)$. By construction, the gram matrix $B := QQ^\top \in \mathcal{A}^*$ since $A^* = M^* {M^*}^\top \in \mathcal{A}^*$. Since the optimal Gram matrix is unique, $B = QQ^\top = M^* {M^*}^\top = A^*$ and therefore for all $n \in \mathcal{C} \setminus \{i, j\}$, we must have $Q_{jn} = {\mu_i^*}^\top \mu_n^* = A_{jn}^* = {\mu_j^*}^\top \mu_n^*$.
It remains to show that $S(QQ^\top) = S(M^* {M^*}^\top)$, 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

$$\forall j_1, j_2 \in \mathcal{C},\ A_{j_1' j_2'}^* = B_{\sigma(j_1')\sigma(j_2')} = B_{j_1 j_2}. \tag{30}$$

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

$$\forall n \in \mathcal{C},\ \lambda_{n'} = \lambda_{\sigma(n')} = \lambda_n. \tag{31}$$

Therefore,

$$\begin{aligned}
S(A^*) &= \sum_{j_0, j_{1:k} \in \mathcal{C}} \left(\lambda_{j_0} \prod_{t=1}^{k} \lambda_{j_t} \right) \psi\big(A_{j_0 j_1}^* - A_{j_0 j_0}^*, \ldots, A_{j_0 j_k}^* - A_{j_0 j_0}^*\big). && (32)\\
&= \sum_{j_0', j_{1:k}' \in \mathcal{C}} \left(\lambda_{j_0'} \prod_{t=1}^{k} \lambda_{j_t'} \right) \psi\big(A_{j_0' j_1'}^* - A_{j_0' j_0'}^*, \ldots, A_{j_0' j_k'}^* - A_{j_0' j_0'}^*\big). && (33)\\
&= \sum_{j_0', j_{1:k}' \in \mathcal{C}} \left(\lambda_{j_0} \prod_{t=1}^{k} \lambda_{j_t} \right) \psi\big(B_{j_0 j_1} - B_{j_0 j_0}, \ldots, B_{j_0 j_k} - B_{j_0 j_0}\big). && (34)\\
&= \sum_{j_0, j_{1:k} \in \mathcal{C}} \left(\lambda_{j_0} \prod_{t=1}^{k} \lambda_{j_t} \right) \psi\big(B_{j_0 j_1} - B_{j_0 j_0}, \ldots, B_{j_0 j_k} - B_{j_0 j_0}\big). && (35)\\
&= S(B), && (36)
\end{aligned}$$

where (32) follows from the definition of $S(\cdot)$ in (26), equality (33) holds because $\sigma(\cdot)$ is a bijection, (34) is due to (30) and (31), equality (35) holds because $\sigma(\cdot)$ is a bijection, and (36) again follows from the definition of $S(\cdot)$ in (26). ■

The following Corollary expands the results of Lemma 5 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 The Corollary follows directly by applying the result in Lemma 5 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 Lemma 5 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. ■

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 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

$$A^* = (1-b)I + b\mathbf{1}\,\mathbf{1}^\top = \begin{bmatrix} 1 & b & b & \cdots & b \\ b & 1 & b & \cdots & b \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b & b & b & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{C\times C}, \tag{37}$$

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 (37),

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

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

Remark 6. *Corollary 5 resolves a question that was left open in Jiang et al. (2023) 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.*

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., 2023b; 2024b; 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.

Lemma 6. *Let $C \geq 3$ and $1 > \lambda_1 > \lambda_2 = \lambda_3 = \ldots = \lambda_C = \frac{1-\lambda_1}{C-1} > 0$. Then*

$$A^* = \begin{bmatrix} 1 & a & a & \cdots & a \\ a & 1 & b & \cdots & b \\ a & b & 1 & \cdots & b \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a & b & b & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{C\times C}, \tag{38}$$

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

Proof The form of A^* in (38) follows from Lemma 5 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 (38). The detailed proof is presented in Appendix A.3. ■

Lemma 7. *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*

$$\Delta_2 := \sup_{w \,\in\, \mathcal{S}(\psi)} ||w||_2, \quad \textit{then } \Delta_2 < \infty, \quad \mathcal{S}(\psi) \subseteq (0, \Delta_2]^k, \textit{ and}$$
$$\forall\, v, v' \in \mathcal{V}, \quad |\psi(v) - \psi(v')| \leq \Delta_2\, ||v - v'||_2. \tag{39}$$

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,

$$\forall\, u \in \mathbb{R}^k_{\geq 0} \setminus \{\mathbf{0}\},\, \exists\, \delta_u \in (0, \infty) \,:\, \forall\, t, t' \in [-2, 0],\, t' \leq t, \quad (t - t')\, \delta_u \leq \phi_u(t) - \phi_u(t'). \tag{40}$$

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 $\forall\, u \in \mathbb{R}^k_{\geq 0}$, $\delta_u = u^\top \nabla\psi(-2u)$.

Proof 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 in Appendix A.4. ■

Lemma 8. *Let $C \geq 3$ and let $A^*(a)$ denote the matrix A^* in (38) 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 7.

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 7, we get

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

■

Theorem 3 (Sufficient conditions for minority collapse)**.** *Let $C, \lambda_{1:C}$ be as in Lemma 6, $S(\cdot)$ be as in (26), Δ_2, ϕ, and δ_u be as in Lemma 7 with $\delta_{\mathbf{0}} := 0$ and*

$$\delta_* := \min_{u \in \{0,1\}^k \setminus \{\mathbf{0}\}} \delta_u \in (0, \infty), \tag{41}$$

and let $a, b, A^*(a)$, and γ_C be as in Lemma 8. Let $\mathcal{E}_{1\bar{1}} := \{y = 1$ and for some i, $y_i^- \neq 1\}$, $\mathcal{E}_{\bar{1}1} := \{y \neq 1$ and for all i, $y_i^- = 1\}$, and $\mathcal{E}_1 := \mathcal{E}_{1\bar{1}} \cup \mathcal{E}_{\bar{1}1}$. 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 (4), if

$$\lambda_1 \geq \frac{1}{1 + \frac{1}{\gamma_C \sqrt{k} \Delta_2} \mathbb{E}[\delta_{u(y, y_{1:k}^-)} | \mathcal{E}_1]}, \tag{42}$$

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. The sufficient condition for minority collapse given by (42) is satisfied if

$$\lambda_1 \in [\tau, 1), \quad \tau := \frac{1}{1 + \frac{\delta_*}{\gamma_C \sqrt{k} \Delta_2}} \in (0, 1). \tag{43}$$

Proof 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

$$\begin{aligned}
&\Pr(\mathcal{E}_=) = \lambda_1^{k+1} + (C-1)\frac{(1-\lambda_1)^{k+1}}{(C-1)^{k+1}} = \lambda_1^{k+1} + \frac{(1-\lambda_1)^{k+1}}{(C-1)^k} \\
&\Pr(\mathcal{E}_1) = \Pr(\mathcal{E}_{1\bar{1}}) + \Pr(\mathcal{E}_{\bar{1}1}) = \lambda_1(1-\lambda_1)^k + (1-\lambda_1)\lambda_1^k = \lambda_1 (1-\lambda_1)\left(\frac{(1-\lambda_1)^k}{1-\lambda_1} + \lambda_1^{k-1}\right), \qquad (44) \\
&\Pr(\mathcal{E}_2) = 1 - \Pr(\mathcal{E}_=) - \Pr(\mathcal{E}1) = (1-\lambda_1)^2\left(\frac{(1-\lambda_1)^k}{1-\lambda_1} - \frac{(1-\lambda_1)^{k-1}}{(C-1)^k}\right), \\
&\frac{\Pr(\mathcal{E}_1)}{\Pr(\mathcal{E}_2)} \geq \frac{\lambda_1}{1-\lambda_1}. \qquad (45)
\end{aligned}$$

Next, noting the definition of ϕ in Lemma 7 and that for all $j \in \mathcal{C}$, $(A^*_{1j} - 1) = (A^*_{j1} - 1) = (a-1)\, 1(j \neq 1)$, we have

$$\forall\, (y, y_{1:k}^-) \in \mathcal{E}_= \cup\, \mathcal{E}_1, \quad \psi(A^*_{y y_1^-}(a) - 1, \ldots, A^*_{y y_1^-}(a) - 1) = \psi((a-1)\, u(y, y_{1:k}^-)) = \phi_{u(y, y_{1:k}^-)}(a-1), \tag{46}$$

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 (4), we have

$$\begin{aligned}
S(A^*(a)) &= \mathbb{E}\Big[\psi(A^*_{y y_1^-}(a) - 1, \ldots, A^*_{y y_1^-}(a) - 1)\Big] \\
&= \Pr(\mathcal{E}_=)\, \phi_{\mathbf{0}}(0) + \Pr(\mathcal{E}_1)\, \mathbb{E}[\phi_{u(y, y_{1:k}^-)}(a-1) | \mathcal{E}_1] + \Pr(\mathcal{E}_2)\, \mathbb{E}\Big[\psi(A^*_{y y_1^-}(a) - 1, \ldots, A^*_{y y_1^-}(a) - 1)\Big| \mathcal{E}_2\Big].
\end{aligned}$$

Therefore,

$S(A^*(a)) - S(A^*(a'))$

$$= \Pr(\mathcal{E}_1)\,\mathbb{E}\Big[\phi_{u(y,y^-_{1:k})}(a-1) - \phi_{u(y,y^-_{1:k})}(a'-1)|\mathcal{E}_1\Big] + \Pr(\mathcal{E}_2)\,\mathbb{E}\Big[\psi(A^*_{yy^-_1}(a)-1,\ldots,A^*_{yy^-_1}(a)-1) - \psi(A^*_{yy^-_1}(a')-1,\ldots,A^*_{yy^-_1}(a')-1)\Big|\mathcal{E}_2\Big]$$

$$\geq \Pr(\mathcal{E}_1)\,\mathbb{E}\big[\delta_{u(y,y^-_{1:k})}(a-a')|\mathcal{E}_1\big] - \Pr(\mathcal{E}_2)\,\gamma_C\sqrt{k}\,\Delta_2\,(a-a') \tag{47}$$

$$= (a-a')\Pr(\mathcal{E}_2)\left\{\frac{\Pr(\mathcal{E}_1)}{\Pr(\mathcal{E}_2)}\,\mathbb{E}\big[\delta_{u(y,y^-_{1:k})}|\mathcal{E}_1\big] - \gamma_C\sqrt{k}\Delta_2\right\}$$

$$\geq (a-a')\,\frac{\Pr(\mathcal{E}_2)}{1-\lambda_1}\left\{\lambda_1\,\mathbb{E}\big[\delta_{u(y,y^-_{1:k})}|\mathcal{E}_1\big] - (1-\lambda_1)\,\gamma_C\sqrt{k}\Delta_2\right\} \tag{48}$$

$$= (a-a')\,\frac{\gamma_C\sqrt{k}\Delta_2\Pr(\mathcal{E}_2)}{1-\lambda_1}\left\{\lambda_1\left(1+\frac{1}{\gamma_C\sqrt{k}\Delta_2}\mathbb{E}\big[\delta_{u(y,y^-_{1:k})}|\mathcal{E}_1\big]\right)-1\right\}$$

$$\geq 0. \tag{49}$$

Inequality (47) follows from (40) and Lemma 8 together with the fact that $s \geq -|s|$ for all $s \in \mathbb{R}$. Inequality (48) follows from (45). Inequality (49) follows from condition (42) and the assumption that $a' < a$. Thus, if condition (42) 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 (43) is satisfied, then condition (42) is also satisfied because

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

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 (41), for all $u \neq \mathbf{0}$, $\delta_u \geq \delta_* > 0$. ■

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 (42) 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, condition (42) for minority collapse in Theorem 3 is satisfied if*

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

Proof From (42), 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,

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

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

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

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

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

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

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

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 7, for all $u \in \{0,1\}^k$,

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

and we note that $\delta_u = g(||u||_1)$ is an increasing function of $||u||_1$. From Theorem 3, $\mathcal{E}_1 := \mathcal{E}_{1\bar{1}} \uplus \mathcal{E}_{\bar{1}1}$, 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 (44),

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

Therefore,

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

$$
\begin{aligned}
&= \lambda_1 \, \mathbb{E}[g(l)], \quad l \sim \text{Binomial}(k, 1-\lambda_1)\,, \\
&\geq \lambda_1 \, \mathbb{E}[1(l \geq k/2)\, g(l)], \quad l \sim \text{Binomial}(k, 1-\lambda_1)\,, \quad \text{since } \forall l, g(l) \geq 0, \\
&\geq \lambda_1 \, \mathbb{E}[1(l \geq k/2)\, g(k/2)], \quad l \sim \text{Binomial}(k, 1-\lambda_1)\,, \quad \text{since } g(l) \text{ increases with } l, \\
&= \lambda_1 \, g(k/2) \Pr(l \geq k/2), \quad l \sim \text{Binomial}(k, 1-\lambda_1)\,, \\
&\geq \lambda_1 \, \frac{1}{2} \, g(k/2), \\
\Rightarrow \mathbb{E}\big[\delta_{u(y, y^-_{1:k})} | \mathcal{E}_1\big] &\geq \frac{1}{4(1-\lambda_1)} \, g(k/2) \\
&= \frac{1}{4(1+3e^2)(1-\lambda_1)}
\end{aligned}
$$

Therefore, a sufficient condition for minority collapse is given by

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

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. .

7 Computer experiments

This section provides two different experiments to verify two phenomena investigated in Section 4 and Section 6, namely, (1) intra-class variance-collapse: 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: if the probabilities of 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 those class means also collapse to a single vector. We focus on the well-known InfoNCE loss in all our experiments.

Intra-class variance-collapse. To verify the intra-class variance-collapse phenomenon, we used a dataset comprising three classes extracted from the CIFAR10 dataset. Specifically, we selected the first 2100, 450, and 450 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 $\lambda_1 = 0.7$ and $\lambda_2 = \lambda_2 = 0.15$. 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 200, 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 illustrates the samples from three classes in the representation space using: (1) the initial mapping before the commencement of training, and (2) the optimal mapping at the conclusion of training. Evidently, all the samples from the same class (represented by the same color) nearly collapse to the same point which is their class mean.

To verify the optimal solutions obtained by the neural network, we used the CVX modeling system (Grant & Boyd, 2014) to solve the convex optimization problem in (28). 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 computed using CVX. The optimal Gram matrices A^* obtained by the neural

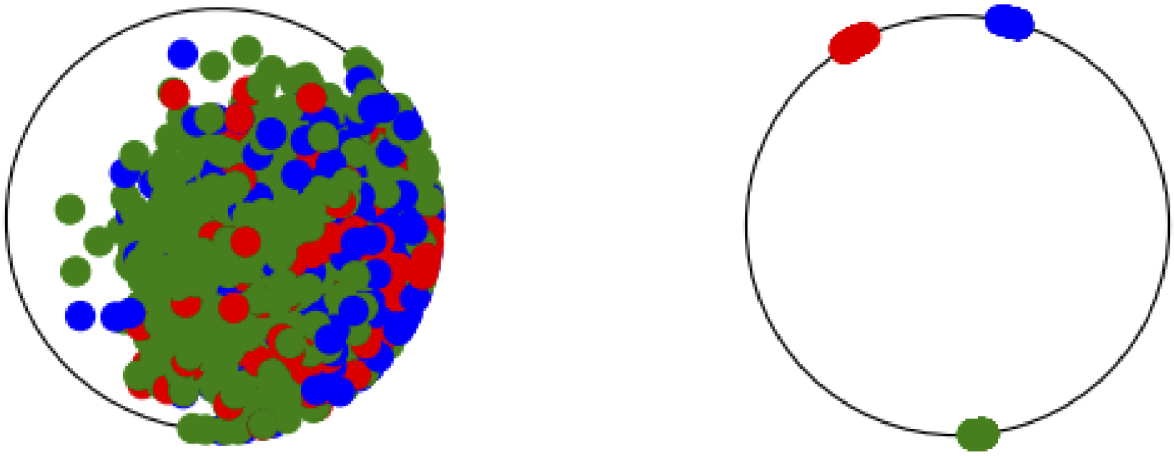

Figure 1: Intra-class variance-collapse in imbalanced datasets. Left: $\mathbb{R}^2$-space representations of 3000 images from 3 classes (indicated by color) of the CIFAR10 dataset using the initial representation function (i.e., before training). Right: representation vectors of the same images at the conclusion of training.

network and the CVX package are

$$A^*_{\text{Neural-Network}} = \begin{bmatrix} 1.0000 & -0.9264 & -0.9264 \\ -0.9264 & 1.0000 & 0.7165 \\ -0.9264 & 0.7165 & 1.0000 \end{bmatrix}, \quad A^*_{\text{CVX-package}} = \begin{bmatrix} 1.0000 & -0.9261 & -0.9261 \\ -0.9261 & 1.0000 & 0.7153 \\ -0.9261 & 0.7153 & 1.0000 \end{bmatrix}.$$

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

We note that if the classes were balanced, then from Theorem 2 in Jiang et al. (2023), 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 Lemma 5 (since $\lambda_2 = \lambda_3$ in this experiment). This confirms our claim that ETF is not the optimal structure for imbalanced cases.

***Minority collapse*:** In Fig. 1, where the class probabilities are $\lambda_1 = 0.7$ and $\lambda_2 = \lambda_3 = 0.15$, we do not observe minority collapse. To empirically validate the minority collapse phenomenon, we constructed a three class dataset that is more severely imbalanced. Specifically, we selected the first 2700, 150, and 150 image samples, respectively, from the first three classes (*i.e.*, $C = 3$), namely bird, automobile, and airplane, of the CIFAR10 dataset to form our second dataset of 3000 samples. This corresponds to the case where $\lambda_1 = 0.9$ and $\lambda_2 = \lambda_3 = 0.05$. 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 $d = 2$. The batch size and the number of epochs were set to 512 and 200, respectively, while the number of negative samples was set 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 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 0.97 in Corollary 6 which guarantees minority collapse. This empirically bolsters our remarks before Corollary 6 that the threshold for minority collapse in Theorem 3 is sufficient for minority collapse, but may not be necessary.

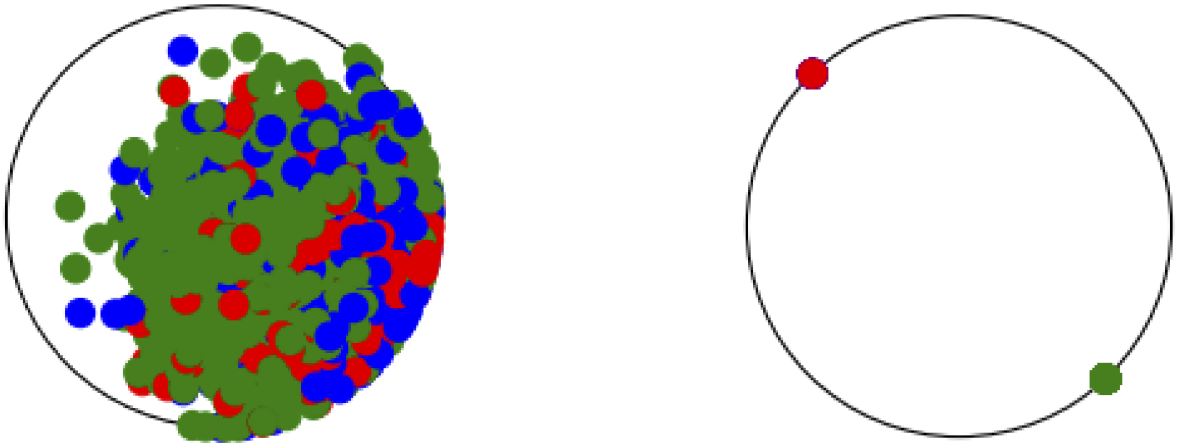

Figure 2: Minority collapse in heavily imbalanced datasets: $\mathbb{R}^2$-space representations of 3000 images from the first three classes of the CIFAR10 dataset (left subfigure) collapse after training to two diametrically opposite points on the unit circle (right subfigure). The red point in the right subfigure represents all 300 samples from the second and third (minority) classes combined whereas the green point represents the 2700 samples from the first (majority) class.

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

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

and they are identical up to the displayed numerical precision. This further confirms our theoretical results that the optimal Gram matrix can be found efficiently using convex optimization. For the reproducibility of our numerical results, we have made our code available at this link[4].

8 Conclusions

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, in the supervised setting and imbalanced datasets, 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 to determine 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.

[4] https://drive.google.com/file/d/1PsCOqrVkxcjL025xs1Dz_UDBrqr_2q0D/view?usp=sharing

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 2023a. URL `https://proceedings.mlr.press/v202/dang23b.html`.

Hien Dang, Tan Nguyen, Tho Tran, Hung Tran, and Nhat Ho. Neural collapse in deep linear network: From balanced to imbalanced data. *arXiv preprint arXiv:2301.00437*, 2023b.

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 2024a. URL `https://proceedings.mlr.press/v235/dang24a.html`.

Hien Dang, Tho Tran, Tan Nguyen, and Nhat Ho. Neural collapse for cross-entropy class-imbalanced learning with unconstrained relu feature model. *arXiv preprint arXiv:2401.02058*, 2024b.

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, Prakash Ishwar, and Shuchin Aeron. Supervised contrastive learning with hard negative samples. *arXiv preprint arXiv:2209.00078*, 2022.

Ruijie Jiang, Thuan Nguyen, Shuchin Aeron, and Prakash Ishwar. On neural and dimensional collapse in supervised and unsupervised contrastive learning with hard negative sampling. *arXiv preprint arXiv:2311.05139*, 2023.

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.

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 Appendix

A.1 Proof of strict convexity of the InfoNCE loss function

Lemma 9. *For all $i = 0, 1, \ldots, k$, let $\alpha_i > 0$. Then the generalized log-sum-exponential (GLSE) function*

$$\psi_{GLSE}(t_{1:k}) := \log\left(\alpha_0 + \sum_{i=1}^{k} \alpha_i e^{t_i}\right) \tag{50}$$

is strictly convex.

Proof The function $\psi_{\text{GLSE}}(t_1, \ldots, t_k)$ 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 9 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 9 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

$$\begin{aligned}
\psi_{\text{GLSE}}(w_{1:k}) &= \log\left(\alpha_0 + \sum_{i=1}^{k} \alpha_i e^{(1-\lambda)u_i + \lambda v_i}\right) \\
&= \log\left(\sum_{i=0}^{k} (\alpha_i e^{u_i})^{1/p} (\alpha_i e^{v_i})^{1/q}\right) \\
&\overset{\text{Hölder}}{\leq} \log\left(\Big(\sum_{i=0}^{k} \alpha_i e^{u_i}\Big)^{1/p} \Big(\sum_{i=0}^{k} \alpha_i e^{v_i}\Big)^{1/q}\right) \\
&= (1-\lambda)\log\left(\alpha_0 + \sum_{i=1}^{k} \alpha_i e^{u_i}\right) + \lambda \log\left(\alpha_0 + \sum_{i=1}^{k} \alpha_i e^{v_i}\right) \\
&= (1-\lambda)\psi_{\text{GLSE}}(u_{1:k}) + \lambda \psi_{\text{GLSE}}(v_{1:k}).
\end{aligned} \tag{51}$$

This shows that $\psi_{\text{GLSE}}(\cdot)$ 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 (51) is strict and therefore $\psi_{\text{GLSE}}(\cdot)$ is a *strictly* convex function. ■

Since the InfoNCE loss function is a GLSE function with $\alpha_0 = 1$ and $\alpha_1 = \ldots = \alpha_k = \frac{1}{k} > 0$, it is a strictly convex function.

A.2 Lemmas for proving variance collapse

Lemma 10. *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 7. *The result of Lemma 10 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 11. *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 10, $z_1 - \mu \overset{\text{w.p.1}}{=} z_2 - \mu \overset{\text{w.p.1}}{=} 0$ which implies that $z_1 \overset{\text{w.p.1}}{=} z_2 \overset{\text{w.p.1}}{=} \mu$. ■

A.3 Proof of Lemma 6

Proof From Lemma 5 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

$$A^* = \begin{bmatrix} 1 & a & a & \cdots & a \\ a & 1 & b & \cdots & b \\ a & b & 1 & \cdots & b \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a & b & b & \cdots & 1 \end{bmatrix}. \tag{52}$$

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

$$A^* = (1-b)\, I + b\, u\, u^\top + b\, e_1\, e_1^\top + a\, e_1\, u^\top + a\, u\, e_1^\top, \tag{53}$$

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 (53), 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 (52), we have

$$\alpha + (C-1)\, a\, \beta = \nu\, \alpha \Rightarrow (C-1)\, a\, \beta = -(1-\nu)\, \alpha \tag{54}$$

$$a\,\alpha + \beta + (C-2)\,b\,\beta = \nu\,\beta \Rightarrow \beta((1-\nu) + (C-2)b) = -a\,\alpha \tag{55}$$

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 $a = 0, b \neq 0$, (54) would imply that $(1-\nu)\,\alpha = 0$ which would imply that either $\nu = 1$ or $\alpha = 0$. If $\nu = 1$, then (55) 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 (54) and $C \geq 3$ would imply that $\beta = 0$ and then (55) would imply that $\alpha = 0$ which would contradict the assumption that both α and β cannot be zero simultaneously. Thus, $\nu \neq 1$. Then, (54) would imply that $\alpha = -(C-1)a\,\beta/(1-\nu)$. Substituting this into (55) 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 $\nu \neq 1$, (54) would imply that $\alpha = 0$, a contradiction). Solving for the roots of the quadratic equation in $(1-\nu)$ we get

$$\nu = 1 + \frac{(C-2)\,b}{2} \pm \sqrt{\frac{(C-2)^2 b^2}{4} + (C-1)a^2} \tag{56}$$

In summary, if $a \neq 0$, then A^* would have $(C-2)$ eigenvalues equal to $(1-b)$ and two eigenvalues given by (56). Since the smallest eigenvalue of A^* is zero, this would imply that either $b = 1$ or

$$1 + \frac{(C-2)\,b}{2} - \sqrt{\frac{(C-2)^2 b^2}{4} + (C-1)a^2} = 0 \Rightarrow b = \frac{(C-1)\,a^2 - 1}{(C-2)}. \tag{57}$$

Observe that if we substitute $a = 0$ into the expression for b in terms of a given by (57), 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 (26) 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.4 Proof of Lemma 7

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

$$-\Delta_2||v-v'||_2 \leq -||w'||_2 \cdot ||v'-v||_2 \leq (v-v')^\top w' \leq \psi(v)-\psi(v') \leq (v'-v)^\top w \leq ||w||_2 \cdot ||v'-v||_2 \leq \Delta_2||v'-v||_2.$$

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 $\lim_{\epsilon\downarrow 0} \frac{\phi_u(-2)-\phi_u(-2-\epsilon)}{\epsilon} = \nabla\phi_u(-2)$. ■