Self-Supervised Contrastive Learning is Approximately Supervised Contrastive Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Despite its empirical success, the theoretical foundations of self-supervised contrastive learning (CL) are not yet fully established. In this work, we address this gap by showing that standard CL objectives implicitly approximate a supervised variant we call the negatives-only supervised contrastive loss (NSCL), which excludes same-class contrasts. We prove that the gap between the CL and NSCL losses vanishes as the number of semantic classes increases, under a bound that is both label-agnostic and architecture-independent.

We characterize the geometric structure of the global minimizers of the NSCL loss: the learned representations exhibit augmentation collapse, within-class collapse, and class centers that form a simplex equiangular tight frame. We further introduce a new bound on the few-shot error of linear-probing. This bound depends on two measures of feature variability—within-class dispersion and variation along the line between class centers. We show that directional variation dominates the bound and that the within-class dispersion's effect diminishes as the number of labeled samples increases. These properties enable CL and NSCL-trained representations to support accurate few-shot label recovery using simple linear probes.

Finally, we empirically validate our theoretical findings: the gap between CL and NSCL losses decays at a rate of $\mathcal{O}(\frac{1}{\#\text{classes}})$; the two losses are highly correlated; minimizing the CL loss implicitly brings the NSCL loss close to the value achieved by direct minimization; and the proposed few-shot error bound provides a tight estimate of probing performance in practice.

1 Introduction

Unsupervised representation learning refers to a class of algorithmic approaches designed to discover meaningful representations of complex data without relying on explicit supervision signals. The goal is to learn representations that preserve and expose semantic information, allowing them to be effectively leveraged in downstream supervised tasks. In recent years, these methods have proven to be effective in pre-training models on unlabeled data, enabling strong generalization on downstream computer vision tasks [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16] and in natural language processing [17, 18, 19, 20, 21, 22, 23, 24].

One of the most successful paradigms for unsupervised learning is self-supervised contrastive learning (CL), where models are trained to distinguish different augmented views of the same image (positives) from views of other images (negatives), typically by minimizing the InfoNCE loss [26, 2]. Notable methods in this category include SimCLR [5], MoCo [27, 6, 7], and CPC [2]. These approaches have led to representations that generalize remarkably well to downstream tasks, often competing with or even surpassing supervised learning. For example, an ImageNet-1M pre-trained MoCo achieves 81.5 AP_{50} on PASCAL VOC [28], while its supervised counterpart achieves 81.3 AP_{50}.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

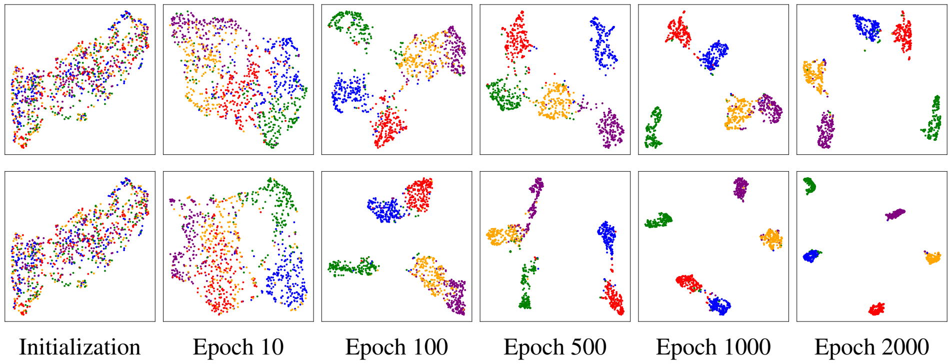


Figure 1: *DCL forms semantic clusters without label supervision, while NSCL yields tighter, more separable clusters, despite not explicitly pulling same-class samples together.* We plot UMAP visualizations for **(top)** decoupled contrastive learning (DCL) [25] and **(bottom)** negatives-only supervised contrastive learning (NSCL) training on mini-ImageNet.

Although the representations learned in supervised classification problems are relatively well understood–thanks to characterizations such as neural collapse [29, 30], which occurs under a variety of loss functions including cross-entropy [31, 32, 33], mean squared error (MSE) [30, 34, 35], and supervised contrastive loss [33, 36, 37, 38]–the theoretical understanding of SSL remains limited. Surprisingly, despite the absence of labels, SSL-trained models often produce representations that closely resemble those learned via supervised training: they tend to cluster nicely with respect to semantic classes [39, 40, 41] (see also Fig. 1) and support downstream tasks such as linear classification and clustering [5, 27, 6, 7, 2]. This raises the following question:

How does self-supervised contrastive learning learn representations similar to those learned through supervised learning, despite the lack of explicit supervision?

Contributions. We provide a theoretical framework that connects CL to it supervised counterpart. Our key contributions are:

- **Duality between self-supervised and supervised CL:** In Thm. 1 we formally establish a connection between decoupled contrastive learning (DCL) [25] and a supervised variant we call negatives-only supervised contrastive loss (NSCL), in which negative pairs are drawn from different classes. Our main insight is that, although DCL does not explicitly exclude same-class samples from the loss denominator, the probability of treating them as negatives vanishes as the number of classes increases. We prove that the gap between the two losses shrinks with more classes. Unlike prior analyses [42, 43, 44, 36, 45], our result holds without assumptions on model architecture, data distribution, or augmentations. Figs. 2–3 show that the two losses are highly correlated, their gap decreases as the number of classes increases, and minimizing the DCL loss implicitly minimizes the NSCL loss.
- **Estimating the few-shot error on downstream tasks:** Prop. 1 introduces a bound on the m-shot classification error based on geometric properties of the learned representation. Specifically, the bound implies that lower class-distance-normalized-variance (CDNV) (see [46] and Sec. 3.2) and directional CDNV (see Sec. 3.2) lead to lower m-shot error, with the impact of the regular CDNV diminishing as the number m of per-class labeled examples increases. Empirically, we verify this behavior for DCL-trained models (Fig. 5) and find the bound to be tight in practice (Fig. 4).
- **Characterizing the global minima of the NSCL loss:** In Thm. 2, we show that any global minimizer of the NSCL loss satisfies: **(i)** *Augmentation collapse*—all augmented views of a sample map to the same point; **(ii)** *Within-class collapse*—all samples from the same class share a representation; and **(iii)** *Simplex ETF structure*—class centers form a maximally separated, symmetric configuration. These solutions coincide with the optimal solutions of certain supervised classification settings, including those using cross-entropy loss [31, 32, 33], mean squared error (MSE) [30, 34, 35], and supervised contrastive loss [33]. Combined with Prop. 1, these properties imply that label recovery via linear probing is guaranteed with few labels.

2 Related Work

Theoretical analyses of CL. A growing body of work aims to explain the effectiveness of CL. Early studies attributed CL's success to maximizing mutual information between augmentations of the same image [3], but later work showed that enforcing strict bounds on mutual information can hurt performance [47, 48]. Another influential line of work focuses on the alignment and uniformity of learned representations [49, 50, 51]. For example, [49] showed that same-sample augmentations are aligned, while negatives are uniformly distributed on a sphere. [51] extended this to a wider family of losses, showing how alignment and uniformity can be modulated by a hyperparameter.

While these are important properties, they do not characterize how CL algorithms organize samples from different classes in the embedding space. As a result, several papers [42, 44, 52, 53, 54, 55, 56, 57, 58, 36, 59] studied CL's ability to recover meaningful clusters and latent variables in the data. However, these results often rely on strong assumptions about the data and augmentations, such as assuming augmentations of the same sample are conditionally independent given their cluster identity (see, e.g., [42, 43, 44, 36]). In order to avoid such assumptions, [60] studies function classes that induce a similar bias towards preserving clusters without any assumption on the connectedness of augmentation sets. However, their analysis is focused on minimizing a spectral contrastive loss which serves as a surrogate for the more practically used InfoNCE loss. [45] instead proves that InfoNCE learns cluster-preserving embeddings by capping the representation class's capacity so that any function that tries to carve a semantic cluster must also split an image from its own augmentation set, making cluster integrity the loss-minimizing choice.

Another way to understand SSL is by examining its connection to supervised learning. For example, [61] showed that, in linear models, certain SSL objectives resembling VicReg are equivalent to supervised quadratic losses. Building on this perspective, we establish a bound between the contrastive loss and a supervised variant of it. Unlike previous work, our bound is both architecture-independent and label-agnostic.

Other theoretical studies have analyzed SSL from different angles, such as feature learning dynamics in linear models and two-layer networks [62, 63, 64, 65], the role of augmentations [12, 66], the projection head [67, 68, 69, 70], CL's sample complexity [71] and how to relax the dependence on large mini-batches [72]. Other work explores connections between self-supervised contrastive and non-contrastive learning [73, 74, 75, 76, 77].

Neural collapse. Neural collapse (NC) [29, 30] is a phenomenon that occurs in the final stages in training of overparameterized networks for classification. In this regime, we observe: (i) *within-class collapse*, where the embeddings of samples within each class converge to a single vector; (ii) *class separation*, where these class means spread out and often form a Simplex Equiangular Tight Frame (ETF); and a form of (iii) *alignment* between the feature space and the classifier's weights.

To understand the emergence of neural collapse, several papers studied the emergence of these properties in different learning settings. In [78, 79] they introduced the "unconstrained features model" or "layer peeled model"—where each sample's embedding is treated as a free parameter in $\mathbb{R}^d$. Under such relaxations, research has demonstrated that neural collapse is a characteristic of global minimizers for many commonly used supervised loss functions [80], including cross-entropy [31, 32, 33], mean squared error (MSE) [30, 34, 35], and supervised contrastive loss [33, 36, 37, 38]. In Thm. 2 we use a similar approach in order to characterize the global minima of the CL and NSCL losses, connecting them to downstream performance using tools from [46, 81, 82].

3 Theoretical Analysis

In this section, we explore *why* CL—despite being agnostic to class labels—tends to organize data by class and support few-shot transfer. We first relate the self-supervised loss to a label-aware loss function, then describe the geometric conditions under which a representation supports few-shot transfer, and finally show how this loss function induces these properties.

Setup. We consider training an embedding function $f \in \mathcal{F} \subset \{f \mid f : \mathcal{X} \to \mathbb{R}^d\}$ via self-supervised contrastive learning, using a dataset $S = \{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times [C]$, where the algorithm only sees inputs $x_i \in \mathcal{X}$ (e.g., images), but not the corresponding class labels $y_i \in [C]$. We assume that each class contains at most $n_{\max}$ samples. We wish to learn a "meaningful" (see Sec. 3.2) function $f : \mathcal{X} \to \mathbb{R}^d$ that maps samples to d-dimensional embedding vectors.

Concretely, for each sample x_i, we also construct K *augmented* versions $x_i^l = \alpha_l(x_i)$ (via data augmentations $\alpha_1, \dots, \alpha_K$, e.g. identity, random cropping, jitter, etc.). Then, we define $z_i^l = f(\alpha_l(x_i))$ as the embeddings of the augmented samples.

We focus on the global decoupled contrastive loss (DCL) [25], which is given by

$$\mathcal{L}^{\mathrm{DCL}}(f) = -\frac{1}{K^2 N} \sum_{l_1, l_2=1}^{K} \sum_{i=1}^{N} \log \left(\frac{\exp(\mathrm{sim}(z_i^{l_1}, z_i^{l_2}))}{\sum_{l_3=1}^{K} \sum_{j \in [N] \setminus \{i\}} \exp(\mathrm{sim}(z_i^{l_1}, z_j^{l_3}))} \right), \tag{1}$$

where $\mathrm{sim}(\cdot, \cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the cosine similarity function, which is defined as follows: $\mathrm{sim}(x_1, x_2) = \frac{\langle x_1, x_2 \rangle}{\|x_1\|_2 \cdot \|x_2\|_2}$. We focus on the DCL loss since its a slight improvement to the original global CL loss [2, 5, 27, 72, 5], where $\sum_{j=1}^{N}$ is replaced with $\sum_{j \neq i}^{N}$ in the denominator. While here we focus on the global [72, 49], for completeness, in Appendix B.1 we extend Thm. 1 to mini-batch CL losses as done in practice.

Despite the lack of label supervision, contrastive models often learn class-aware features. This is surprising: **Why should a label-agnostic loss promote semantic structure?**

3.1 Self-Supervised vs. Supervised Contrastive Learning

To answer this, we compare DCL with a supervised contrastive loss that removes same-class contrasts. Namely, we consider the *negatives-only supervised contrastive loss* (NSCL), $\mathcal{L}^{\mathrm{NSCL}}(f)$, which is defined the same as DCL in (1) but with $\sum_{j \in [N] \setminus \{i\}}$ in the denominator replaced by $\sum_{j : y_j \neq y_i}$.

When comparing the two losses, the unsupervised denominator includes at most extra $K(n_{\max} - 1)$ terms corresponding to data from the same class. Therefore, if $n_{\max}$ is small relative to the number of classes C (i.e., when C is large), the extra $K(n_{\max} - 1)$ terms are expected to be negligible compared to the total number of negatives $\geq K(N - n_{\max})$, and the two objectives, $\mathcal{L}^{\mathrm{DCL}}(f)$ and $\mathcal{L}^{\mathrm{NSCL}}(f)$, become nearly equivalent. For instance, assume for simplicity that $\exp(\mathrm{sim}(z_i, z_j)) \approx \gamma$ for all $i \neq j$ (i.e., the similarity values are approximately constant). Then the unsupervised denominator is approximately $K(N-1)\gamma$ while the supervised denominator (excluding same-class data) becomes approximately $K(N - n_{\max})\gamma$. Overall, we obtain that $|\mathcal{L}^{\mathrm{DCL}}(f) - \mathcal{L}^{\mathrm{NSCL}}(f)| \approx \log\left(1 + \frac{n_{\max}}{N - n_{\max}}\right) \leq \frac{n_{\max}}{N - n_{\max}}$. Hence, the gap between the two losses shrinks as C grows.

In essence, this example shows that if $\exp(\mathrm{sim}(z_i^{l_1}, z_j^{l_2}))$ is nearly constant, removing same-class negatives has little effect on the loss. The following theorem shows that one can achieve a slightly worse bound on $|\mathcal{L}^{\mathrm{DCL}}(f) - \mathcal{L}^{\mathrm{NSCL}}(f)|$ without assuming that $\exp(\mathrm{sim}(z_i^{l_1}, z_j^{l_2}))$ is constant.

Theorem 1. *Let $S = \{(x_i, y_i)\}_{i=1}^{N} \subset \mathcal{X} \times [C]$ be a labeled dataset with C classes, each containing at most $n_{\max}$ distinct samples. Let $f : \mathcal{X} \to \mathbb{R}^d$ be any function. Then, we have*

$$\mathcal{L}^{\mathrm{NSCL}}(f) \leq \mathcal{L}^{\mathrm{DCL}}(f) \leq \mathcal{L}^{\mathrm{NSCL}}(f) + \log\left(1 + \frac{n_{\max} \mathrm{e}^2}{N - n_{\max}}\right) \leq \mathcal{L}^{\mathrm{NSCL}}(f) + \frac{n_{\max} \mathrm{e}^2}{N - n_{\max}},$$

where e *denotes Euler's constant. For a balanced classification problem,* $\frac{n_{\max}}{N - n_{\max}} = \frac{1}{C-1}$.

The above theorem gives a justification for why contrastive SSL yields class-aware features without ever seeing the labels. Interestingly, since the right-hand side depends only on the number of samples N and the maximal number of samples per class $n_{\max}$—not on the specific assignment Y—the inequality is label-agnostic. Namely, for any labeling Y with $n_{\max}$ samples per class, we have,

$$0 \leq \mathcal{L}^{\mathrm{DCL}}(f) - \mathcal{L}^{\mathrm{NSCL}}_{(Y)}(f) \leq \log(1 + \tfrac{n_{\max} \mathrm{e}^2}{N - n_{\max}}), \tag{2}$$

where $\mathcal{L}^{\mathrm{NSCL}}_{(Y)}(f)$ is the NSCL loss under that labeling. In particular, $C \to \infty$ the term $\log(1 + \frac{n_{\max} \mathrm{e}^2}{N - n_{\max}})$ vanishes, the probability of drawing a same-class contrasting drops to zero, and DCL effectively collapses onto the NSCL for all labelings simultaneously.

3.2 Characterizing Good Representations for Few-Shot Learning

The primary goal of SSL is to learn representations that are easily adaptable to downstream tasks. We now describe the conditions under which an SSL-trained representation f can be adapted to a specific downstream task. For this purpose, we consider the *recoverability* of labels from a learned representation $f\colon \mathcal{X} \to \mathbb{R}^d$. We call f *"good"* if it supports accurate downstream classification with only a few labeled samples per class.

We formalize this via the expected m-shot classification error. Suppose we have a set of class-conditional distributions $D_1, \dots, D_{C'}$ and define $D = \frac{1}{C'} \sum_{c=1}^{C'} D_c$ as their uniform mixture. For example, $D_1, \dots, D_{C'}$ can represent a subset of $C' \le C$ classes from the pre-training task, where all D_i are either training datasets or data distributions corresponding to their respective classes. Given an m-shot training set $\widehat{S}_i \sim D_i^m$ for each class i, we define:

$$\mathrm{err}_{m,D}(f) \;:=\; \mathbb{E}_{\widehat{S}_1,\dots\widehat{S}_{C'}}[\mathrm{err}_D(h_{f,\widehat{S}})],$$

where $h_{f,\widehat{S}}$ is a classifier trained using only the small labeled support $\widehat{S} = \cup_{c=1}^{C'} \hat{S}_c$ and $\mathrm{err}_D(h)$ is the probability that h misclassifies a random test point drawn from the task distribution D. Specifically, we denote linear probing (LP) and nearest-class-center classification (NCCC) errors by $\mathrm{err}^{\mathrm{LP}}_{m,D}(f)$ and $\mathrm{err}^{\mathrm{NCC}}_{m,D}(f)$ which refer to $\mathrm{err}_{m,D}(f)$ with $h_{f,\widehat{S}}$ being a linear classifier that minimizes $\mathrm{err}_{\widehat{S}}(h)$ and the nearest-class-center classifier $\arg\min_{c\in[C']} \|f(x) - \mu_f(\widehat{S}_c)\|_2$.

Prior work [46, 81, 82] shows that this error can be bounded in terms of how *clustered* the representations are. Specifically, for each class i, let $\mu_i = \mathbb{E}_{x\sim D_i}[f(x)]$ and $\sigma_i^2 = \mathrm{Var}_{x\sim D_i}\left[\|f(x) - \mu_i\|_2^2\right]$ denote the mean and variance of the class embeddings, and let $V_f(D_i, D_j) = \sigma_i^2/\|\mu_i - \mu_j\|^2$ be the class-distance-normalized variance [46, 81, 82] between classes i and j.

These works (see, e.g., Prop. 7 in [82]) showed that

$$\mathrm{err}^{\mathrm{LP}}_{m,D}(f) \;\le\; \mathrm{err}^{\mathrm{NCC}}_{m,D}(f) \;\lesssim\; (C'-1)(1+\tfrac{1}{m})\mathrm{Avg}_{i\neq j}[V_f(D_i,D_j)]. \tag{3}$$

In supervised learning, representations tend to become tightly clustered, causing CDNV to decrease, and thus reducing few-shot error [29, 46, 83, 84].

However, in the self-supervised case, labels are not available during training. As a result, there is no mechanism to explicitly encourage intra-class similarity, and thus we cannot generally expect that $\mathrm{Avg}_{i\neq j}[V_f(D_i, D_j)]$ will be small.

To better capture settings where label information is implicit, we introduce a refinement of (3) that prioritizes *directional variability*. For each pair i, j, let: $\sigma_{ij}^2 = \mathrm{Var}_{x\sim D_i}\left[\langle f(x) - \mu_i, u_{ij}\rangle\right]$, where $u_{ij} = \frac{\mu_i - \mu_j}{\|\mu_i - \mu_j\|_2}$. Then the *directional* CDNV is defined as: $\tilde{V}_f(D_i, D_j) = \sigma_{ij}^2/\|\mu_i - \mu_j\|_2^2$, that measures the variation along the line between the class centers.

This leads to the following stronger bound:

Proposition 1. *Let $C' \ge 2$ and $m \ge 10$ be integers. Fix a feature map $f : \mathcal{X} \to \mathbb{R}^d$ and class-conditional distributions $D_1, \dots, D_{C'}$ over $\mathcal{X}$. We have:*

$$\mathrm{err}^{\mathrm{LP}}_{m,D}(f) \;\le\; \mathrm{err}^{\mathrm{NCC}}_{m,D}(f) \;\le\; (C'-1)\left[14\mathrm{Avg}_{i\neq j}[\tilde{V}_f(D_i,D_j)] + \tfrac{28.5}{m}\mathrm{Avg}_{i\neq j}[V_f(D_i,D_j)]\right]. \tag{4}$$

Unlike (3), which depends on the full CDNV, Prop. 1 depends primarily on the directional CDNV, (and on $\frac{1}{m}V_f$), which is always smaller than the CDNV ($\tilde{V}_f \le V_f$) and in isotropic distributions $f \circ D_i$ it even scales as $\tilde{V}_f = \frac{1}{d}V_f$. Thus, the new bound can explain low m-shot error even when the regular CDNV is large. In essence, the bound predicts that SSL-trained models are effective for downstream tasks when their directional CDNV is very low and their CDNV remains moderate. Fig. 5 shows that this prediction holds in practice.

Finally, we note that the coefficients (14 and 28.5) in the bound are sub-optimal. In (8) (Appendix C) we give a more general form of the coefficients, parameterized by a variable a. In Cor. 1 (Appendix C) we give an optimized version of the bound using a near-optimal choice of a. In Fig. 4(bottom), we show that the bound in Cor. 1 is fairly tight in practice.

3.3 Characterizing the representations learned with $\mathcal{L}^{\mathrm{NSCL}}$

Prop. 1 establishes that if the directional CDNV is small and the overall CDNV remains bounded, then the linear probing error of f will likewise be small. Since Thm. 1 guarantees that minimizing the DCL loss also drives down the NSCL loss, we can use NSCL as a stand-in for DCL. Accordingly, in this section we analyze representations learned under the NSCL objective and evaluate how well they recover class labels through linear probing.

Because minimizing over all neural networks f is intractable, we follow prior work [78, 79] and adopt the unconstrained features model. Specifically, we treat f as an arbitrary function that given an augmented sample x_i^l selects an embedding z_i^l as a *free* learnable vector in $\mathbb{R}^d$. The following theorem characterizes the global minimizers of the NSCL loss, showing that they satisfy the NC1, NC2, and NC4 properties of NC [29, 30], and thus learn representations similar to those of minimizers of other classification losses, such as cross-entropy [31, 32, 33], MSE [30, 34, 35], and SCL [33].

Theorem 2. *Let $d \geq C - 1$ and let $S = \{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times [C]$ be a balanced labeled dataset with C classes. Suppose f is a global minimizer of the supervised contrastive loss $\mathcal{L}^{\mathrm{NSCL}}(f)$ (over all functions $f : \mathcal{X} \to \mathbb{R}^d$). Then, the representations satisfy the following properties:*

1. ***Augmentation Collapse:*** *For each $i \in [N]$ and for every pair $l_1, l_2 \in [K]$, we have $z_i^{l_1} = z_i^{l_2}$.*
2. ***Within-Class Collapse:*** *For any two samples x_i and x_j with the same label ($y_i = y_j$), their representations coincide: $z_i = z_j$. Namely, each class has a unique class embedding.*
3. ***Simplex Equiangular Tight Frame:*** *Let $\{\mu_1, \dots, \mu_C\}$ denote the set of class-center embeddings. These vectors form a simplex ETF in $\mathbb{R}^d$; specifically, they satisfy $\sum_{c=1}^C \mu_c = 0$, $\|\mu_c\|_2 = \|\mu_{c'}\|_2$ and $\langle \mu_c, \mu_{c'} \rangle = -\frac{\|\mu_c\|_2^2}{C-1}$ for all $c \neq c' \in [C]$.*

Thm. 2 implies that any global minimizer of $\mathcal{L}^{\mathrm{NSCL}}(f)$ yields perfectly clustered representations. As such, the CDNV and the directional CDNV are zero and by Prop. 1, the 1-shot errors vanish: $\mathrm{err}^{\mathrm{LP}}_{m,S}(f) \leq \mathrm{err}^{\mathrm{NCC}}_{m,S}(f) = 0$. In CL on the other hand, we cannot achieve such collapse, since it would imply that the representations encode only label information–despite the loss being label-agnostic. Still, the DCL-NSCL duality invites a more nuanced question: **Does minimizing CL still promote weak forms of clustering?** While zero CDNV is tied to a specific labeling, the few-shot error $\mathrm{err}^{\mathrm{NCC}}_{m,S}(f)$ and the directional CDNV can be small under many possible labelings. In Sec. 4, we empirically explore the extent to which minimizing the DCL loss induces such a structure.

4 Experiments

4.1 Experimental Setup

Datasets. We experiment with the following datasets - CIFAR10 and CIFAR100 [85], mini-ImageNet [86], and SVHN [87]. For **additional experiments and details**, see Appendix A.

Methods, architectures and optimizers. We trained our models with the SimCLR [5] algorithm. We use a ResNet-50 [88] encoder with a width-multiplier factor of 2. The projection head follows a standard two-layer MLP architecture composed of: `Linear`$(2048 \to 2048)$ $\to$ `ReLU` $\to$ `Linear`$(2048 \to 128)$. For additional experiments with MoCo v2 [6], see Appendix A.

Instead of training our models with the standard InfoNCE loss that is used in SimCLR, we use the DCL loss that avoids positive-negative coupling during training [25]. In order to minimize the loss, we adopt the LARS optimizer [89] which has been shown in [5] to be effective for training with large batch sizes. For LARS, we set the momentum to 0.9 and the weight decay to $1\mathrm{e}^{-6}$. All experiments are carried out with a batch size of $B = 1024$. The base learning rate is scaled with batch size as $0.3 \cdot \lfloor B/256 \rfloor$, following standard practice [5]. We employ a warm-up phase [90] for the first 10 epochs, followed by a cosine learning rate schedule without restarts [91] for the remaining epochs. All models were trained on a single node with two 94 GB NVIDIA H100 GPUs.

Evaluation metrics. In several experiments we monitor $\mathcal{L}^{\mathrm{DCL}}$ (see (1)) and $\mathcal{L}^{\mathrm{NSCL}}$ (see 3.1). To calculate each one of these loss functions, we replace the sum over samples and their K augmentations in the denominator with a sum over a random batch of $B = 1024$ random samples and one random augmentation for each sample as in SimCLR (see Appendix B.1 for details).

To evaluate the quality of learned representations, we use two methods: the Nearest Class-Center Classification (NCCC) accuracy [83], and linear probing accuracy [92, 93, 94]. Suppose we have a feature map f, a classification task with C classes, and a dataset $S = \cup_{i=1}^C S_i$ (either training or test data). We estimate $\mathrm{err}^{\mathrm{LP}}_{m,S}(f)$ and $\mathrm{err}^{\mathrm{NCC}}_{m,S}(f)$ with $h_{f,\widehat{S}}$ being an NCC classifier or a linear classifier.

To train the linear classifier, we use cross-entropy minimization for 500 epochs, with batch size $\min(|\widehat{S}|, 256)$, learning rate $3\mathrm{e}^{-4}$, weight decay $5\mathrm{e}^{-4}$, and momentum 0.9. The expectation over the selection of $\widehat{S} = \cup_{i=1}^{C} \widehat{S}_i$ is estimated by averaging the error over 5 selections of $\widehat{S}$ from S.

We also estimate the bound in Prop. 1. As mentioned above, the coefficients in (4) are sub-optimal, so we instead use the refined bound from Cor. 1 (see Appendix C). To assess its ability to predict downstream performance, we compute the bound on both the training and test data by treating the per-class train/test subsets S_i as the distributions D_i. The CDNV and the directional CDNV are calculated exactly as described in Sec. 3.2.

4.2 Experimental Results

Validating Thm. 1 during training. We train models using SimCLR to minimize the DCL loss. We evaluate both the DCL and NSCL losses on both training and test sets. Additionally, we evaluate the proposed upper bound, given by $\mathcal{L}^{\mathrm{NSCL}}(f) + \log(1 + \frac{n_{\max}\mathrm{e}^2}{N - n_{\max}})$, where N is the total number of samples and $n_{\max}$ is the maximal number of samples per class (see Thm. 1).

In Fig. 2(top) we observe that the DCL loss consistently upper bounds the NSCL loss, and that the two losses become closer for tasks with a larger number of classes (e.g., CIFAR100 and mini-ImageNet compared to CIFAR10 and SVHN). In addition, when C is large (e.g., $C = 100$), the bound becomes very tight, closely matching the losses. Although we notice a slight divergence between the self-supervised and supervised loss metrics as training progresses, the overall trend continues to align well with the inequality stated in Thm. 1. For similar experiments with MoCo-v2, see Fig. 7.

Dataset	CIFAR10		CIFAR100		mini-ImageNet	
	NCCC	LP	NCCC	LP	NCCC	LP
DCL	$85.3 \pm 2\mathrm{e}^{-1}$	$86.3 \pm 8\mathrm{e}^{-3}$	$57.3 \pm 1\mathrm{e}^{-1}$	$61.7 \pm 5\mathrm{e}^{-2}$	$69.0 \pm 2\mathrm{e}^{-1}$	$72.9 \pm 2\mathrm{e}^{-2}$
NSCL	$95.7 \pm 4\mathrm{e}^{-3}$	$95.6 \pm 4\mathrm{e}^{-3}$	$70.8 \pm 1\mathrm{e}^{-1}$	$73.7 \pm 2\mathrm{e}^{-2}$	$79.8 \pm 1\mathrm{e}^{-1}$	$81.3 \pm 2\mathrm{e}^{-2}$

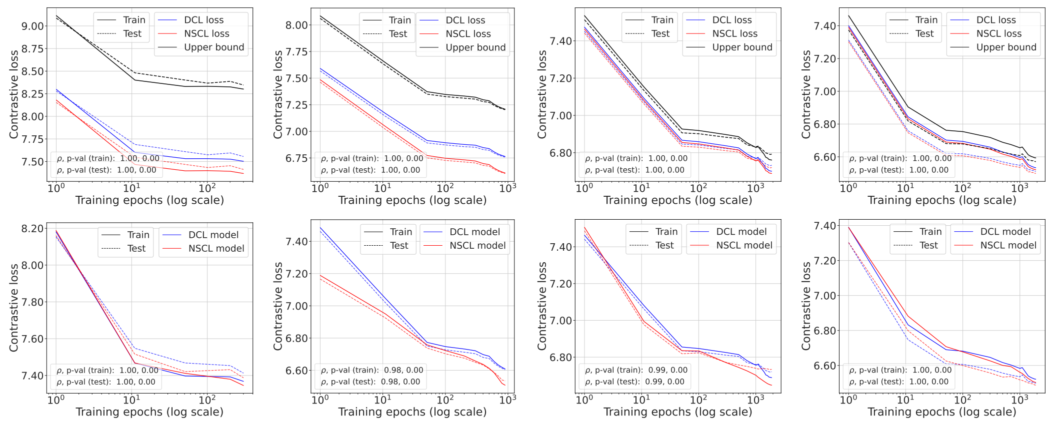
Table 1: NCCC and LP 100-shot test-time accuracy rates (and their standard deviations) of DCL and NSCL-trained models.

While this result shows that for large C, the NSCL loss decreases alongside the DCL loss, it does not imply that minimizing the DCL loss leads to the same solution as directly minimizing the NSCL loss. To investigate this, we conducted an experiment in which we trained two models to minimize the DCL and NSCL losses (respectively) and compared their resulting NSCL values. As shown in Fig. 2(bottom), the gap between the NSCL losses of the two models is fairly small. This indicates that, in practice, optimizing the DCL loss leads to representations that are fairly clustered in comparison with those obtained by explicitly optimizing the NSCL loss.

Validating Thm. 1 with varying C. In the next experiment, we analyze how the gap $\mathcal{L}^{\mathrm{DCL}}(f) - \mathcal{L}^{\mathrm{NSCL}}(f)$ scales with the number of classes C. Specifically, we empirically validate that this gap is upper-bounded by $\log(1 + \frac{n_{\max}\mathrm{e}^2}{N - n_{\max}}) = \log(1 + \frac{\mathrm{e}^2}{C-1})$ (see Thm. 1), that the gap becomes tighter for large C, and that it is highly correlated with the actual gap between the losses.

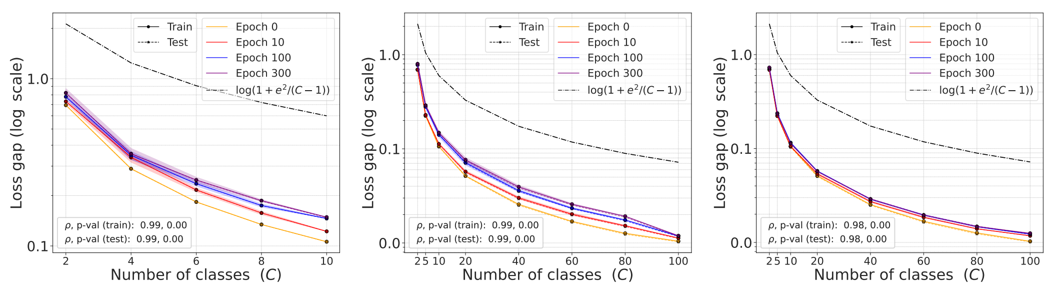
For each value of C, we randomly sample C classes from the full dataset, train a self-supervised model from scratch on data from these classes only, and compute $\mathcal{L}^{\mathrm{DCL}}(f) - \mathcal{L}^{\mathrm{NSCL}}(f)$ at various training epochs. To account for variability in class selection, we report the averaged value of this quantity over five independent random selections of C classes along with error bars. Fig. 3 presents the empirical results, showing that the gap $\mathcal{L}^{\mathrm{DCL}}(f) - \mathcal{L}^{\mathrm{NSCL}}(f)$ slightly increases over the course of training but remains consistently bounded by $\log(1 + \frac{\mathrm{e}^2}{C-1})$ at all epochs. Moreover, the magnitude of this gap decreases with C and is highly correlated with our bound.

Comparing downstream performance. To evaluate the quality of the learned representations, we measure the few-shot downstream performance of models trained with DCL and NSCL. Specifically, we report the NCCC error and linear probing error (see Sec. 4.1). Fig. 4 (top) shows train and test performance across all classes as a function of the number of shots per class (m). As can be seen, although the NCCC and linear probing errors of the DCL-trained models are, as expected, higher than those of the NSCL-trained model, they remain fairly low (see Table 1 for a numeric comparison). This indicates that, despite not being explicitly optimized to align with class labels, the representations learned by DCL still exhibit strong clustering behavior.



(a) SVHN **(b)** CIFAR10 **(c)** CIFAR100 **(d)** mini-ImageNet

Figure 2: **(Top)** We train the model to minimize the DCL loss, tracking the DCL loss, the NSCL loss, and the bound NSCL+$\log(1+\frac{n_{\max}e^2}{N-n_{\max}})$ on both the training and test sets throughout training. *All three quantities are highly correlated.* **(Bottom)** We compare the NSCL loss of two models: one trained with the DCL loss and the other with the NSCL loss. *The resulting NSCL losses are comparable, regardless of the training objective.* In both the top and bottom plots, correlations are computed between the DCL and NSCL losses on the train and test data.

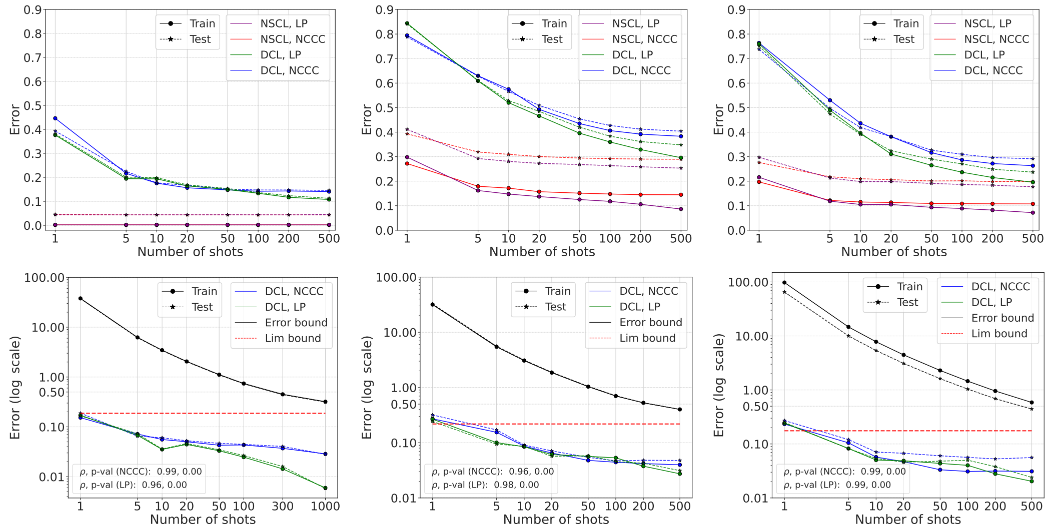


(a) CIFAR10 **(b)** CIFAR100 **(c)** mini-ImageNet

Figure 3: *The gap between the DCL and NSCL losses shrinks as the number of classes C grows.* Models were trained to minimize the DCL loss, and at several training epochs we plot the empirical difference $\mathcal{L}^{\mathrm{DCL}} - \mathcal{L}^{\mathrm{NSCL}}$ alongside the bound $\log(1+\frac{e^2}{C-1})$ as a function of C. We also report correlation between the loss gap at epoch 300 and the bound.

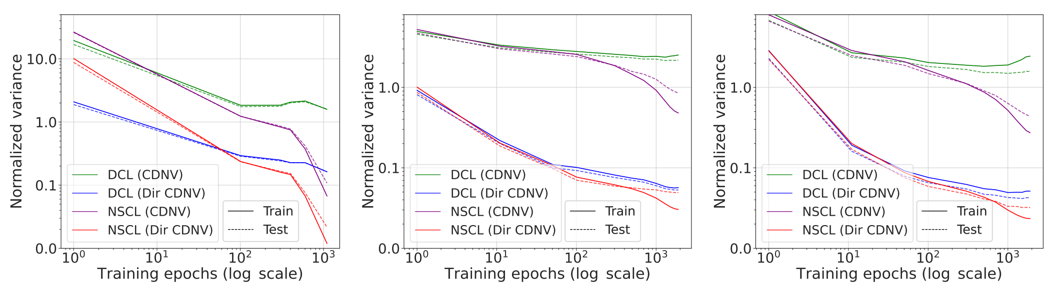
Fig. 4(bottom) compares the bound from Cor. 1 with the NCCC and linear-probe errors of DCL-trained models on 2-way downstream tasks (averaged over 10 tasks), for both train and test data. Unlike Prop. 7 in [82], and despite the constants in our bound, it is fairly tight as m increases; for instance, at $m = 500$, it indicates that the few-shot errors are below 0.45.At $m = 10^6$, the test-time bound (red horizontal line) predicts an error of at most 0.25. Fig. 5 shows that DCL training only modestly lowers CDNV but reduces directional CDNV by about an order of magnitude during training. Because Prop. 1 links few-shot error primarily to the directional CDNV, this sharp drop explains why DCL-trained models can recover labels (see Fig. 4) with limited supervision. In contrast, NSCL reduces both variances substantially, which accounts for its lower m-shot error.

UMAP visualizations of clustering behaviors. We visualize the clustering behavior of DCL and NSCL-trained representations learned on mini-ImageNet using 2D UMAP [95] projections. We randomly selected 5 classes from mini-ImageNet and sampled 200 images per class, projecting their corresponding embeddings into 2D space using UMAP. As shown in Fig. 1, training with the NSCL loss results in tight and clearly separable clusters, as predicted in Thm. 2. Although DCL is label-agnostic, as shown in Thm. 1 and Fig. 2, DCL training implicitly minimizes the NSCL loss, which explains the visible clustering behavior seen in Fig. 1.



(a) CIFAR10 **(b)** CIFAR100 **(c)** mini-ImageNet

Figure 4: **(Top)** We compare the m-shot error (Linear Probing (LP) and Nearest Class-Centered Classifier (NCCC)) for C-way (all-class) classification. *Although the NCCC and LP errors of the DCL-trained models are, as expected, higher than those of the NSCL-trained model, they remain fairly low.* **(Bottom)** We compare the two-way m-shot NCCC and LP errors with the bound in Cor. 1. The errors are bounded by our bound, which decreases with m. The dashed red *'Lim bound'* line specifies the bound at $m = 10^6$, providing a tight test-time error estimate for large m. We also report correlation between the error bound and the errors $\mathrm{err}^{\mathrm{LP}}_{m,D}(f)$ and $\mathrm{err}^{\mathrm{NCCC}}_{m,D}(f)$ (on the test data).



(a) CIFAR10 **(b)** CIFAR100 **(c)** mini-ImageNet

Figure 5: *DCL training yields a moderate reduction in CDNV and a significant reduction in directional CDNV, whereas NSCL training significantly decreases both.* We plot the CDNV and directional CDNV for both train and test data for DCL and NSCL-trained models over 2k epochs.

5 Discussion, Limitations, and Future Work

CL is at the forefront of modern pre-training techniques. Our work takes several steps toward a better understanding of CL by: establishing a duality between CL and NSCL, analyzing the minimizers of the NSCL loss, and linking the downstream error of CL-trained models to directional CDNV.

Nevertheless, our work has several important limitations that remain open for future investigation. While we show that the DCL and NSCL losses are close in value, this does not directly imply that their respective minimizers are close in parameter space or yield similar representations. It would be interesting to explore this connection both theoretically and empirically—for example, by monitoring the distance between two models trained using DCL and NSCL on the same batches. Another limitation is that the bound in Prop. 1 (and Cor. 1) involves large constants and scales linearly with C', which may limit its practical utility. Tightening this bound by relaxing these dependencies would enhance its predictive power. Finally, extending these ideas to multimodal models such as CLIP [96], or to LLMs trained with analogous loss functions, presents an exciting direction for future research.

References

[1] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

[2] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.

[7] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649, October 2021.

[8] Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding, 2020.

[9] Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021.

[10] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations, 2019.

[11] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*, page 776–794, Berlin, Heidelberg, 2020. Springer-Verlag.

[12] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6827–6839. Curran Associates, Inc., 2020.

[13] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

[14] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

[15] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[16] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[17] Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. Cert: Contrastive self-supervised learning for language understanding, 2020.

[18] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[19] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[20] Yixuan Su, Fangyu Liu, Zaiqiao Meng, Tian Lan, Lei Shu, Ehsan Shareghi, and Nigel Collier. TaCL: Improving BERT pre-training with token-aware contrastive learning. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2497–2507, Seattle, United States, July 2022. Association for Computational Linguistics.

[21] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.

[22] Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. Pairwise supervised contrastive learning of sentence representations. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5786–5798, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[23] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online, August 2021. Association for Computational Linguistics.

[24] Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. ConSERT: A contrastive framework for self-supervised sentence representation transfer. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online, August 2021. Association for Computational Linguistics.

[25] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, page 668–684, Berlin, Heidelberg, 2022. Springer-Verlag.

[26] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[28] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010.

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

[30] X.Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2022.

[31] Jianfeng Lu and Stefan Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022. Special Issue on Harmonic Analysis and Machine Learning.

[32] Hien Dang, Tho Tran, Tan Nguyen, and Nhat Ho. Neural collapse for cross-entropy class-imbalanced learning with unconstrained relu features model. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[33] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3821–3830. PMLR, 18–24 Jul 2021.

[34] Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse, 2022.

[35] Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features, 2022.

[36] Pranjal Awasthi, Nishanth Dikkala, and Pritish Kamath. Do more negative samples necessarily hurt in contrastive learning? In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1101–1116. PMLR, 17–23 Jul 2022.

[37] Jaidev Gill, Vala Vakilian, and Christos Thrampoulidis. Engineering the neural collapse geometry of supervised-contrastive loss, 2023.

[38] 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.

[39] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For?*, 2022.

[40] Ido Ben-Shaul, Ravid Shwartz-Ziv, Tomer Galanti, Shai Dekel, and Yann LeCun. Reverse engineering self-supervised learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[41] Xi Weng, Jianing An, Xudong Ma, Binhang Qi, Jie Luo, Xi Yang, Jin Song Dong, and Lei Huang. Clustering properties of self-supervised learning, 2025.

[42] Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, and Nikunj Saunshi. A theoretical analysis of contrastive unsupervised representation learning, 2019.

[43] Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19250–19286. PMLR, 17–23 Jul 2022.

[44] Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021.

[45] Advait Parulekar, Liam Collins, Karthikeyan Shanmugam, Aryan Mokhtari, and Sanjay Shakkottai. Infonce loss provably learns cluster-preserving representations. In Gergely Neu and Lorenzo Rosasco, editors, *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 1914–1961. PMLR, 12–15 Jul 2023.

[46] Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2022.

[47] David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 875–884. PMLR, 26–28 Aug 2020.

[48] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020.

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

[50] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2495–2504, 2021.

[51] Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845, 2021.

[52] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.

[53] Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra. Investigating the role of negatives in contrastive representation learning. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 7187–7209. PMLR, 28–30 Mar 2022.

[54] Kento Nozawa and Issei Sato. Understanding negative samples in instance discriminative self-supervised representation learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 5784–5797. Curran Associates, Inc., 2021.

[55] Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 5000–5011. Curran Associates, Inc., 2021.

[56] Jeff Z. HaoChen, Colin Wei, Ananya Kumar, and Tengyu Ma. Beyond separability: analyzing the linear transferability of contrastive representations to related subpopulations. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[57] Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z. Haochen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 19847–19878. PMLR, 17–23 Jul 2022.

[58] Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *International Conference on Learning Representations*, 2022.

[59] Han Bao, Yoshihiro Nagano, and Kento Nozawa. On the surrogate gap between contrastive and supervised losses. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1585–1606. PMLR, 17–23 Jul 2022.

[60] Jeff Z. HaoChen and Tengyu Ma. A theoretical study of inductive biases in contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[61] Randall Balestriero and Yann LeCun. The birth of self supervised learning: A supervised theory. In *NeurIPS 2024 Workshop: Self-Supervised Learning - Theory and Practice*, 2024.

[62] Yuandong Tian. Understanding deep contrastive learning via coordinate-wise optimization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 19511–19522. Curran Associates, Inc., 2022.

[63] Wenlong Ji, Zhun Deng, Ryumei Nakada, James Zou, and Linjun Zhang. The power of contrast for feature learning: a theoretical analysis. *J. Mach. Learn. Res.*, 24(1), January 2023.

[64] Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 11112–11122. PMLR, 18–24 Jul 2021.

[65] Yuandong Tian. Understanding the role of nonlinearity in training dynamics of contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[66] Shlomo Libo Feigin, Maximilian Fleissner, and Debarghya Ghoshdastidar. A theoretical characterization of optimal data augmentations in self-supervised learning, 2025.

[67] Kartik Gupta, Thalaiyasingam Ajanthan, Anton van den Hengel, and Stephen Gould. Understanding and improving the role of projection head in self-supervised learning, 2022.

[68] Yu Gui, Cong Ma, and Yiqiao Zhong. Unraveling projection heads in contrastive learning: Insights from expansion and shrinkage, 2023.

[69] Yihao Xue, Eric Gan, Jiayi Ni, Siddharth Joshi, and Baharan Mirzasoleiman. Investigating the benefits of projection head for representation learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[70] Zhuo Ouyang, Kaiwen Hu, Qi Zhang, Yifei Wang, and Yisen Wang. Projection head is secretly an information bottleneck. In *The Thirteenth International Conference on Learning Representations*, 2025.

[71] Noga Alon, Dmitrii Avdiukhin, Dor Elboim, Orr Fischer, and Grigory Yaroslavtsev. Optimal sample complexity of contrastive learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[72] Zhuoning Yuan, Yuexin Wu, Zi-Hao Qiu, Xianzhi Du, Lijun Zhang, Denny Zhou, and Tianbao Yang. Provable stochastic optimization for global contrastive learning: Small batch does not harm performance. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 25760–25782. PMLR, 2022.

[73] Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2021.

[74] Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[75] Jason D Lee, Qi Lei, Nikunj Saunshi, and JIACHENG ZHUO. Predicting what you already know helps: Provable self-supervised learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 309–323. Curran Associates, Inc., 2021.

[76] Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann LeCun. On the duality between contrastive and non-contrastive self-supervised learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[77] Ravid Shwartz-Ziv, Randall Balestriero, Kenji Kawaguchi, Tim G. J. Rudner, and Yann LeCun. An information theory perspective on variance-invariance-covariance regularization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[78] Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features, 2020.

[79] 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.

[80] Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 31697–31710. Curran Associates, Inc., 2022.

[81] Tomer Galanti, András György, and Marcus Hutter. Improved generalization bounds for transfer learning via neural collapse. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022.

[82] Tomer Galanti, András György, and Marcus Hutter. Generalization bounds for few-shot transfer learning with pretrained classifiers, 2023.

[83] Tomer Galanti, Liane Galanti, and Ido Ben-Shaul. Comparative generalization bounds for deep neural networks. *Transactions on Machine Learning Research*, 2023.

[84] Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

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

[86] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[87] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

[88] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[89] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. *arXiv preprint arXiv:1708.03888*, 2017.

[90] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[91] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

[92] Arne Köhn. What's in an embedding? analyzing word embeddings through multilingual evaluation. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2067–2073, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[93] Abhijeet Gupta, Gemma Boleda, Marco Baroni, and Sebastian Padó. Distributional vectors encode referential attributes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 12–21, 2015.

[94] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

[95] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*, February 2018.

[96] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[97] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[98] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract outlines both the theoretical and empirical contributions, as well as the main assumptions made in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide a limitations section in the main text.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The paper provides the full set of assumptions for each theoretical result, along with complete and correct proofs in the appendix. All key logical steps in the derivations are clearly presented.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe the experiments and their setup in detail in Sec. 4.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide well-documented code in the supplemental material to reproduce our experimental results. We use publicly available datasets in our work.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Yes, we provide full details required to reproduce our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We include error bars when appropriate and feasible, given our computational budget. For instance, in Tab. 1 and Fig. 3, where the experiments were averaged over different sample selections or class subsets, we report the results using means and standard deviations or error bars. Due to limited computational resources, we could not afford to repeat the more computationally intensive experiments. Instead, we prioritized running them on multiple datasets.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes, we provide information about our compute resources, and memory.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This is a theoretical paper on self-supervised learning. While it analyzes algorithms that significantly impact widely used applications, the contributions of this work have no direct societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: As a theoretical paper, it does not pose risks of this kind.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use publicly available datasets and code with appropriate attribution and in accordance with their respective licenses. Specifically:

    - **Datasets:** CIFAR-10 and CIFAR100 [85], mini-ImageNet [86] and SVHN [87] that are explained in Appendix A.
    - **Frameworks:** Our models are implemented in PyTorch [97], which is released under the BSD 3 License.
    - **Codebases:** We build on official or open-source implementations of SimCLR and MoCo, distributed under Apache 2.0 and CC-BY-NC 4.0 Licenses, respectively.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not create any new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

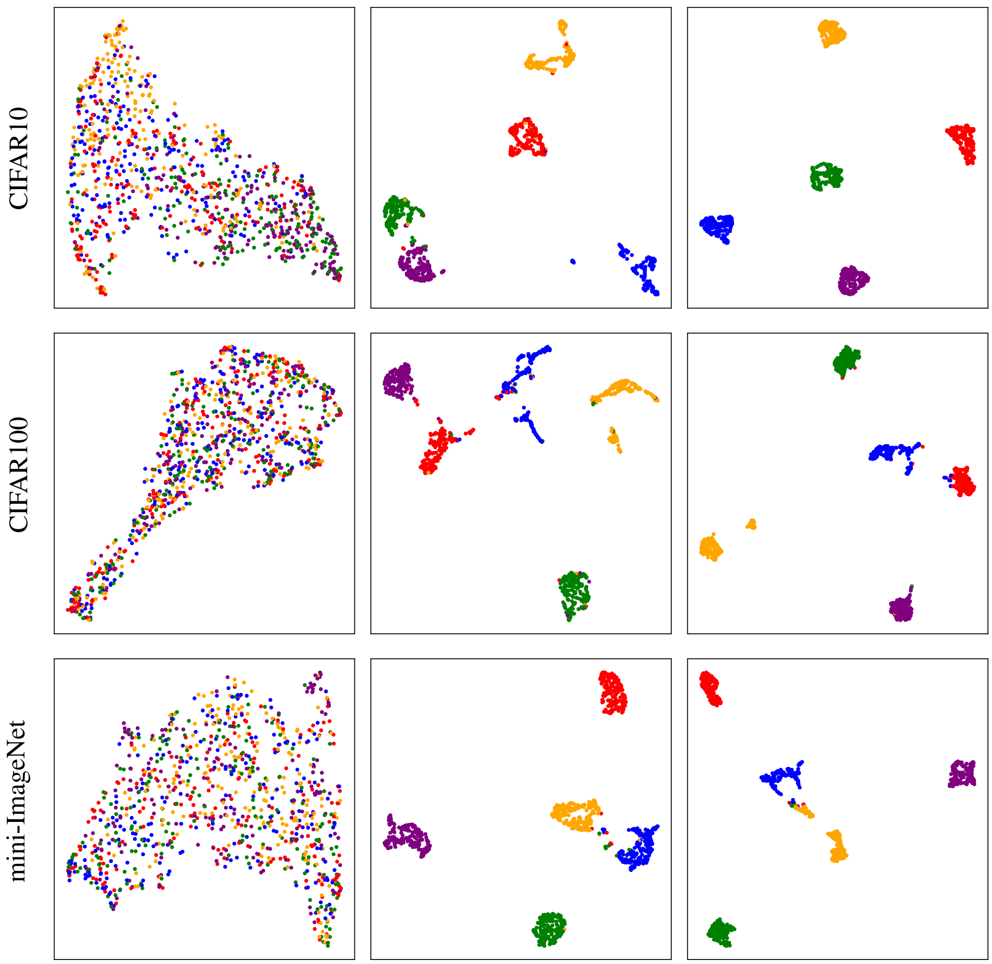
Justification: We use LLMs for editing (e.g., grammar, spelling, word choice).

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Additional Experiments

Datasets. We experiment with the following standard vision classification datasets - CIFAR10 and CIFAR100 [85], mini-ImageNet [86], and SVHN [87]. CIFAR10 and CIFAR100 both consist of 50000 training images and 10000 validation images with 10 classes and 100 classes, respectively, uniformly distributed across the dataset, i.e., CIFAR10 has 5000 samples per class and CIFAR100 has 500 samples per class. mini-ImageNet also has 5000 test images on top of 50000 train and 10000 validation images, with 100 of 1000 classes from ImageNet-1K [98] (at the original resolution). SVHN consists of digit classification data with 10 classes from real-world images, organized into "train" (73,257 samples), "test" (26,032 samples), and "extra" (531,131 samples) splits. For scalability verification, we combine the "train" and "extra" splits during training for validating Thm. 1 (shown in Fig. 2).



(a) Initialization **(b)** DCL **(c)** NSCL

Figure 6: UMAP visualizations of representations from models trained with different objectives (Random init, DCL, NSCL) on CIFAR10, CIFAR100, and mini-ImageNet. Each point corresponds to an image embedding colored by class. Better clustering and separation are evident as we move from Random to NSCL.

Data augmentations. We use the same augmentations as in SimCLR [5]. For experiments on mini-ImageNet and SVHN, we use the following pipeline: random resized cropping to 224×224, random horizontal flipping, color jittering (brightness, contrast, saturation: 0.8; hue: 0.2), random grayscale conversion ($p = 0.2$), and Gaussian blur (applied with probability 0.1 using a 3×3 kernel and $\sigma = 1.5$). For CIFAR datasets, we adopt a similar pipeline with appropriately scaled parameters.

The crop size is adjusted to 32×32, and the color jitter parameters are scaled to saturation 0.4, and hue 0.1.

UMAP visualizations of clustering behaviors across datasets. In Fig. 1, we show a visual comparison of the clusters formed by the learned representations of DCL and NSCL-based models during training for mini-ImageNet. In Fig. 6, we demonstrate that the behavior generalizes across different datasets, $i.e.$, CIFAR10, CIFAR100, and mini-ImageNet. We extract these embeddings using models trained for 2k epochs on their respective datasets.

A.1 Extension to MoCo

To verify the generality of our empirical findings in the main text, we repeat some primary experiments using the Momentum Contrast method, specifically MoCo v2 [6]. This section summarizes empirical results of MoCo, mirroring our analyses for SimCLR which we presented earlier.

MoCo setup. We use the same architecture as with SimCLR. To train our model, we use SGD as an optimizer. We set momentum to 0.9 and the weight decay to $1\mathrm{e}^{-4}$. All experiments are carried out with a batch size of $B = 256$. The base learning rate is set to 0.03 [27]. Similar to our SimCLR training strategy, we employ a warm-up phase [90] for the first 10 epochs, followed by a cosine learning rate schedule [91].

A.1.1 Results

Validating Thm. 1 during training. We train MoCo using the DCL loss for 2k epochs, and evaluate both the DCL and NSCL losses on training and test datasets, along with the theoretical upper bound $\mathcal{L}^{\mathrm{NSCL}}(f) + \log(1 + \frac{n_{max}\mathrm{e}^2}{N - n_{max}})$. We repeat the experiments illustrated in Fig. 2(top). In Fig. 7(top), we notice similar trends for MoCo as we have shown for SimCLR. The DCL loss consistently upper bounds the NSCL loss, with a narrowing gap with increase in number of classes C. The empirical $\mathcal{L}^{\mathrm{CL}}$ and $\mathcal{L}^{\mathrm{NSCL}}$ values remain tightly bounded by our proposed bound throughout training. We further validate that minimizing the DCL loss implicitly leads to low NSCL loss by repeating the experiments in Fig. 2(bottom). As shown in Fig. 7(bottom), $\mathcal{L}^{\mathrm{NSCL}}$ for both DCL and NSCL-trained models remain close. This confirms our earlier conclusion that optimizing DCL implicitly reduces the NSCL loss.

Comparing downstream performance. We evaluate the quality of the learned representations from MoCo models via few-shot error analysis as earlier shown in Fig. 4. Specifically, we report the NCCC error on both train and test datasets. In Fig. 8(top) for MoCo, we evaluate NCCC error on the complete dataset as a function of number of shots per class (m) and show a comparison between DCL and NSCL-based models. In Fig. 8(bottom), we perform 2-way classification (averaged over 10 tasks) and compare the NCCC error with the bound from Cor. 1. We notice that the theoretical bound remains tight for increasing m. For example, at large m (e.g., $m = 10^6$) the test-time error bound drops down to 0.39 for CIFAR100.

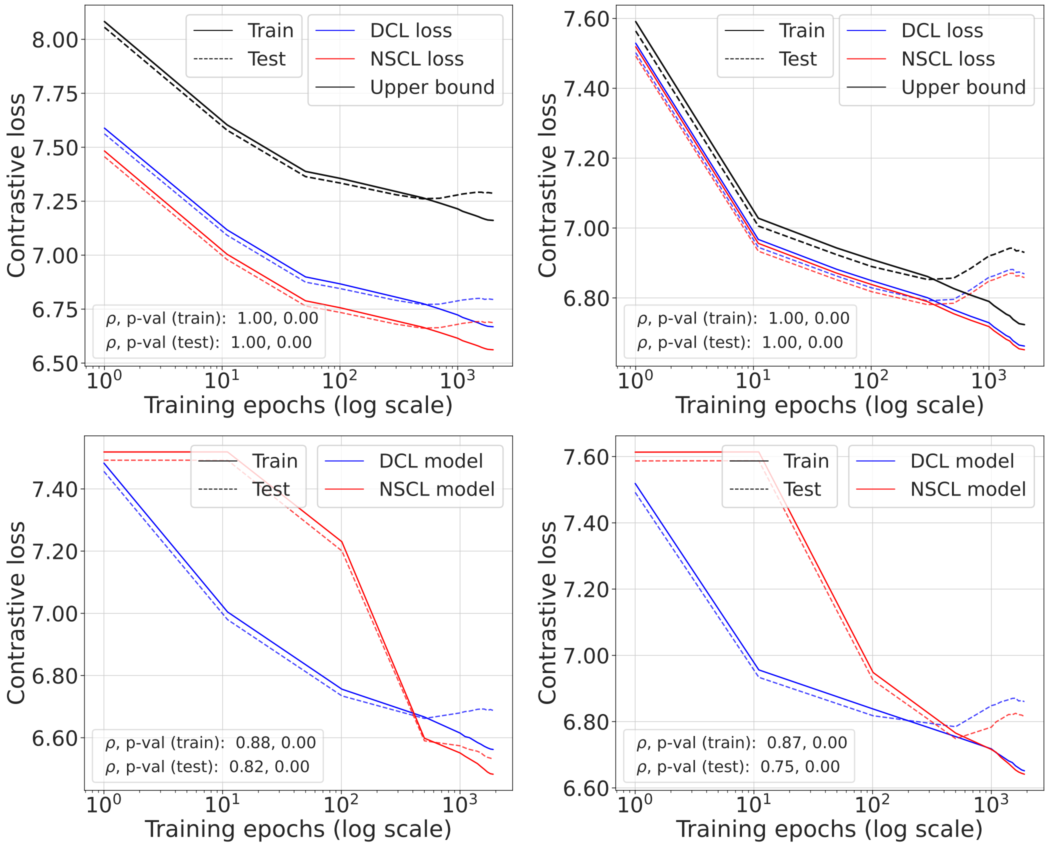
Similarly to our previous analysis, we also evaluate CDNV and directional CDNV in Fig. 9 for both DCL and NSCL-based MoCo models. We observe that while CDNV initially decreases for both models, the NSCL-trained models achieve substantially lower CDNV values as training progresses, whereas DCL-trained models maintain higher CDNV values. For directional CDNV, both the models achieve substantial reductions (an order of magnitude). For CIFAR10, directional CDNV for NSCL drops an additional order of magnitude compared to DCL.

B Proof of Thm. 1

Theorem 1. *Let $S = \{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times [C]$ be a labeled dataset with C classes, each containing at most $n_{\max}$ distinct samples. Let $f : \mathcal{X} \to \mathbb{R}^d$ be any function. Then, we have*

$$\mathcal{L}^{\mathrm{NSCL}}(f) \;\le\; \mathcal{L}^{\mathrm{DCL}}(f) \;\le\; \mathcal{L}^{\mathrm{NSCL}}(f) + \log\left(1 + \tfrac{n_{\max}\mathrm{e}^2}{N - n_{\max}}\right) \;\le\; \mathcal{L}^{\mathrm{NSCL}}(f) + \tfrac{n_{\max}\mathrm{e}^2}{N - n_{\max}},$$

where e *denotes Euler's constant. For a balanced classification problem, $\frac{n_{\max}}{N - n_{\max}} = \frac{1}{C-1}$.*



(a) CIFAR10 **(b)** CIFAR100

Figure 7: **The DCL and NSCL losses, along with our bound, for MoCo.** Same as Fig. 2 with MoCo training in place of SimCLR.

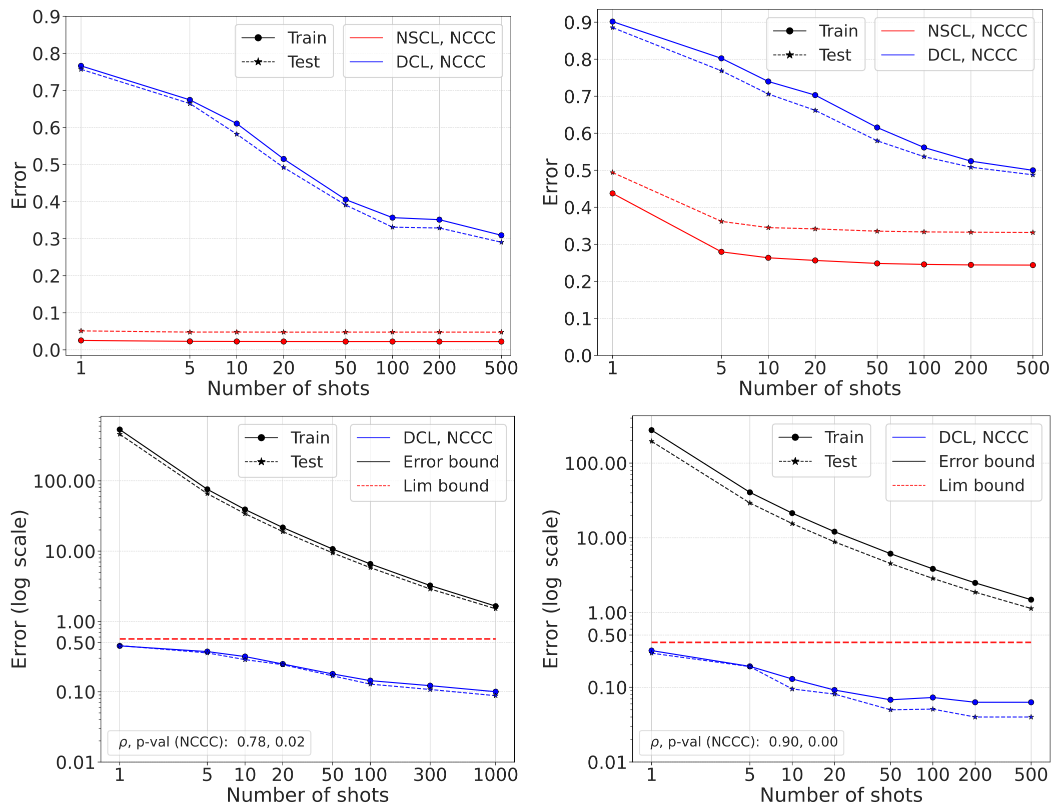
Proof. First, we show that $\mathcal{L}^{\mathrm{NSCL}}(f) \leq \mathcal{L}^{\mathrm{DCL}}(f)$. For each anchor (i, l_1) (with $i \in \{1, \dots, N\}$ and $l_1 \in \{1, \dots, K\}$), define

$$Z_{\mathrm{neg}}^{i,l_1} = \sum_{l_3=1}^{K} \sum_{\substack{j=1 \\ y_j \neq y_i}}^{N} \exp(\mathrm{sim}(z_i^{l_1}, z_j^{l_3})), \quad Z_{\mathrm{pos}}^{i,l_1} = \sum_{l_3=1}^{K} \sum_{\substack{j=1 \\ y_j = y_i}}^{N} \exp(\mathrm{sim}(z_i^{l_1}, z_j^{l_3}))$$

$$\text{and } Z_{\mathrm{pos}\setminus\mathrm{self}}^{i,l_1} = \sum_{l_3=1}^{K} \sum_{\substack{j=1 \\ y_j = y_i, j \neq i}}^{N} \exp(\mathrm{sim}(z_i^{l_1}, z_j^{l_3})).$$

By the definitions of the decoupled contrastive loss $\mathcal{L}^{\mathrm{DCL}}(f)$ and the negatives-only supervised contrastive loss $\mathcal{L}^{\mathrm{NSCL}}(f)$, we obtain

$$\begin{aligned}
&\mathcal{L}^{\mathrm{DCL}}(f) - \mathcal{L}^{\mathrm{NSCL}}(f) \\
&= \frac{1}{K^2 N} \sum_{i=1}^{N} \sum_{l_1=1}^{K} \sum_{l_2=1}^{K} \left[-\log\left(\frac{\exp(\mathrm{sim}(z_i^{l_1}, z_i^{l_2}))}{Z_{\mathrm{neg}}^{i,l_1} + Z_{\mathrm{pos}\setminus\mathrm{self}}^{i,l_1}} \right) + \log\left(\frac{\exp(\mathrm{sim}(z_i^{l_1}, z_i^{l_2}))}{Z_{\mathrm{neg}}^{i,l_1}} \right) \right] \\
&= \frac{1}{K^2 N} \sum_{i=1}^{N} \sum_{l_1=1}^{K} \sum_{l_2=1}^{K} \log\left(\frac{Z_{\mathrm{neg}}^{i,l_1} + Z_{\mathrm{pos}\setminus\mathrm{self}}^{i,l_1}}{Z_{\mathrm{neg}}^{i,l_1}} \right) \\
&= \frac{1}{N} \sum_{i=1}^{N} \log\left(1 + \frac{Z_{\mathrm{pos}\setminus\mathrm{self}}^{i,l_1}}{Z_{\mathrm{neg}}^{i,l_1}} \right).
\end{aligned} \tag{5}$$



(a) CIFAR10 **(b)** CIFAR100

Figure 8: **(Top)** m-shot error (Nearest Class-Centered Classifier (NCCC)) for C-way (all-class) classification. **(Bottom)** Two-way m-shot NCCC errors with the bound in Cor. 1. Same as Fig. 4 with MoCo training in place of SimCLR.

Since $\log(1+x) \geq 0$ for all $x \geq 0$, it follows that $\mathcal{L}^{\mathrm{NSCL}}(f) \leq \mathcal{L}^{\mathrm{DCL}}(f)$. A similar argument shows that $\mathcal{L}^{\mathrm{DCL}}(f) \leq \mathcal{L}^{\mathrm{CL}}(f)$.

Next, by using the same reasoning as in (5) but replacing $Z^{i,l_1}_{\mathrm{pos}\setminus\mathrm{self}}$ with Z^{i,l_1}_{pos}, we have

$$\mathcal{L}^{\mathrm{CL}}(f) - \mathcal{L}^{\mathrm{NSCL}}(f) \;=\; \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + \frac{Z^{i,l_1}_{\mathrm{pos}}}{Z^{i,l_1}_{\mathrm{neg}}}\right).$$

Next, we bound the ratio $\frac{Z^{i,l_1}_{\mathrm{pos}}}{Z^{i,l_1}_{\mathrm{neg}}}$. For a fixed anchor (i, l_1), note that there are at most $Kn_{\max}$ positive terms (since each class contains at most $n_{\max}$ samples) and at least $K(N - n_{\max})$ negative terms. Moreover, because $\mathrm{sim}(z, z') \in [-1, 1]$, every term satisfies

$$\exp(-1) \;\leq\; \exp(\mathrm{sim}(z_i^{l_1}, z_j^{l_3})) \;\leq\; \exp(1).$$

Thus, we obtain

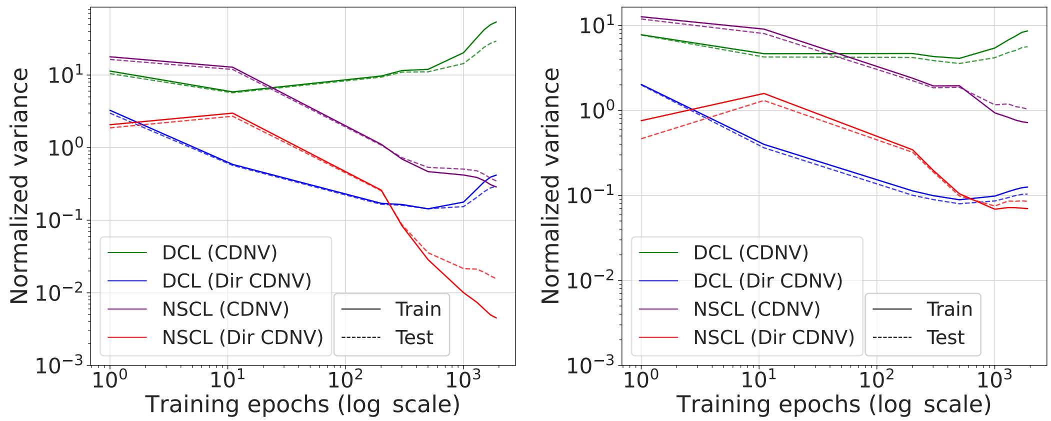
$$Z^{i,l_1}_{\mathrm{pos}} \;\leq\; Kn_{\max}\exp(1) \quad \text{and} \quad Z^{i,l_1}_{\mathrm{neg}} \;\geq\; K(N - n_{\max})\exp(-1).$$

Hence,

$$\frac{Z^{i,l_1}_{\mathrm{pos}}}{Z^{i,l_1}_{\mathrm{neg}}} \;\leq\; \frac{Kn_{\max}\exp(1)}{K(N - n_{\max})\exp(-1)} \;=\; \frac{n_{\max}\mathrm{e}^2}{N - n_{\max}}.$$

It follows that for each anchor,

$$\log\left(1 + \frac{Z^{i,l_1}_{\mathrm{pos}}}{Z^{i,l_1}_{\mathrm{neg}}}\right) \;\leq\; \log\left(1 + \frac{n_{\max}\mathrm{e}^2}{N - n_{\max}}\right).$$



(a) CIFAR10 **(b)** CIFAR100

Figure 9: CDNV and directional CDNV for both train and test data for DCL and NSCL-trained models over 2k epochs, for MoCo. Same as Fig. 5 with MoCo training in place of SimCLR.

Since this bound is uniform in i, l_1, and l_2, we conclude that

$$\mathcal{L}^{\mathrm{CL}}(f) - \mathcal{L}^{\mathrm{NSCL}}(f) \;\leq\; \log\left(1 + \frac{n_{\max}\mathrm{e}^2}{N - n_{\max}}\right) \;\leq\; \frac{n_{\max}\mathrm{e}^2}{N - n_{\max}},$$

where the last inequality follows from $\log(1+x) \leq x$ for all $x \geq 0$. $\square$

B.1 Batch-Based Contrastive Losses

In practice, contrastive learning objectives are implemented using batches of samples rather than full-dataset sums [5]. To formalize this, let $B \in \mathbb{N}$ denote a chosen batch size. We assume access to a dataset $S = \{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times [C]$, where $x_i \in \mathcal{X}$ are inputs (e.g., images) and $y_i \in [C]$ are their class labels. For each sample x_i, let $x_i' \sim \alpha(x_i)$ be an independently generated augmentation from a distribution of augmentations $\alpha(x_i)$. Given a batch $\mathcal{B} = \{(x_{j_t}, x'_{j_t}, y_{j_t})\}_{t=1}^B$ sampled with replacement from S, we define a per-sample contrastive loss for an anchor (x_i, y_i) as:

$$\ell_{i,\mathcal{B}}(f) \;=\; -\log\left(\frac{\exp(\mathrm{sim}(f(x_i), f(x_i')))}{\sum_{j\in\mathcal{B}}\left[\exp(\mathrm{sim}(f(x_i), f(x_j))) + \exp(\mathrm{sim}(f(x_i), f(x_j')))\right]}\right). \tag{6}$$

The self-supervised contrastive loss is then defined as the expectation over randomly chosen anchors, their augmentations and batches:

$$\mathcal{L}_B^{\mathrm{CL}}(f) \;=\; \mathbb{E}_{x_i, x_i'}\mathbb{E}_{\mathcal{B}}\left[\ell_{i,\mathcal{B}}(f)\right]. \tag{7}$$

In contrast, the negatives-only supervised contrastive loss restricts the negatives in each batch to only include examples from different classes. Specifically, for each anchor (x_i, y_i), define $\mathcal{B}_i' = \{(x_{j_t}, x'_{j_t}, y_{j_t})\}_{t=1}^B$ as a batch of size B, drawn (with replacement) from $S \setminus \{(x_j, y_j) : y_j = y_i\}$. Then, the loss is:

$$\mathcal{L}_B^{\mathrm{NSCL}}(f) \;=\; \mathbb{E}_{x_i, x_i'}\mathbb{E}_{\mathcal{B}_i}\left[\ell_{i,\mathcal{B}_i'}(f)\right].$$

With these notations we are ready to describe a bound on the gap between these two loss functions.

Theorem 3. *Let $S = \{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times [C]$ be a labeled dataset with C classes, each containing n distinct samples $\big(N = Cn\big)$. Let $B \in \mathbb{N}$ be the batch size, $\epsilon > 0$ be a positive number and $\bar{B} = \lceil B(1 - \frac{1}{C} - \epsilon)\rceil$. Let $f : \mathcal{X} \to \mathbb{R}^d$ be any function. Then, the difference between the self-supervised contrastive loss $\mathcal{L}_B^{\mathrm{CL}}(f)$ and the decoupled supervised contrastive loss $\mathcal{L}_{\bar{B}}^{\mathrm{NSCL}}(f)$ satisfies*

$$-2(\log(2B)+2)\exp(-2B\epsilon^2) \;\leq\; \mathcal{L}_B^{\mathrm{CL}}(f) - \mathcal{L}_{\bar{B}}^{\mathrm{NSCL}}(f) \;\leq\; \frac{\mathrm{e}^2(1+\epsilon C)}{C(1-\epsilon)-1} + 2(\log(2B)+2)\exp(-2B\epsilon^2).$$

Proof. Fix an arbitrary anchor sample $(x_i, y_i) \in S$ and let x_i' be an augmentation of x_i. Denote

$$z_i \;=\; f(x_i) \quad \text{and} \quad z_i' \;=\; f(x_i')$$

the corresponding embeddings. Next, let $\mathcal{B} = \{(x_{j_t}, x_{j_t}', y_{j_t})\}_{t=1}^{B}$ be a random batch of B samples (with replacement) from S along with their augmentations. For the anchor (x_i, y_i), define the per-sample contrastive loss by

$$\ell_{i,\mathcal{B}}(f) \;=\; -\log\left(\frac{\exp\left(\mathrm{sim}(z_i, z_i')\right)}{\sum_{j\in\mathcal{B}}\left[\exp\left(\mathrm{sim}(z_i, z_j)\right) + \exp\left(\mathrm{sim}(z_i, z_j')\right)\right]}\right).$$

Thus, the overall self-supervised contrastive loss is

$$\mathcal{L}_B^{\mathrm{CL}}(f) \;=\; \mathbb{E}_i\, \mathbb{E}_{\mathcal{B}}\left[\ell_{i,\mathcal{B}}(f)\right].$$

For each anchor i, consider instead a batch

$$\mathcal{B}_i' \;=\; \{(x_{j_t}, x_{j_t}', y_{j_t})\}_{t=1}^{\bar{B}}$$

of $\bar{B} = \lceil B(1 - \frac{1}{C} - \epsilon)\rceil$ samples drawn (with replacement) from $S \setminus \{(x_j, y_j) : y_j = y_i\}$, i.e., only from the negatives relative to i. The decoupled supervised contrastive loss is defined as

$$\mathcal{L}_{\bar{B}}^{\mathrm{NSCL}}(f) \;=\; \mathbb{E}_i\, \mathbb{E}_{\mathcal{B}_i'}\left[\ell_{i,\mathcal{B}_i'}(f)\right].$$

For each anchor i, define the event

$$A_i \;=\; \left\{\mathcal{B} : \#\{j \in \mathcal{B} : y_j \neq y_i\} \geq \bar{B}\right\}.$$

When A_i holds, the distribution of a random subset $\mathcal{B}_i''$ of $\bar{B}$ negatives from $\mathcal{B}$ coincides with that of $\mathcal{B}_i'$. By the law of total expectation,

$$\begin{aligned}\mathcal{L}_B^{\mathrm{CL}}(f) - \mathcal{L}_{\bar{B}}^{\mathrm{NSCL}}(f) \;=&\; \mathbb{E}_i\left[\mathbb{P}[A_i]\, \mathbb{E}_{\mathcal{B},\mathcal{B}_i''|A_i}\left[\ell_{i,\mathcal{B}}(f) - \ell_{i,\mathcal{B}_i''}(f)\right]\right] \\ &+ \mathbb{E}_i\left[\mathbb{P}[\bar{A}_i]\,\left(\mathbb{E}_{\mathcal{B}|\bar{A}_i}\left[\ell_{i,\mathcal{B}}(f)\right] - \mathbb{E}_{\mathcal{B}_i'}\left[\ell_{i,\mathcal{B}_i'}(f)\right]\right)\right].\end{aligned}$$

To bound the difference on the event A_i, define

$$Z_{i,\mathcal{B}} \;=\; \sum_{j\in\mathcal{B}}\left[\exp\left(\mathrm{sim}(z_i, z_j)\right) + \exp\left(\mathrm{sim}(z_i, z_j')\right)\right].$$

Then, one may write

$$\ell_{i,\mathcal{B}}(f) - \ell_{i,\mathcal{B}_i''}(f) \;=\; \log\left(\frac{Z_{i,\mathcal{B}}}{Z_{i,\mathcal{B}_i''}}\right) \;=\; \log\left(1 + \frac{Z_{i,\mathcal{B}\setminus\mathcal{B}_i''}}{Z_{i,\mathcal{B}_i''}}\right).$$

Since $\log(1+u) \leq u$ for all $u \geq 0$, it follows that

$$\ell_{i,\mathcal{B}}(f) - \ell_{i,\mathcal{B}_i''}(f) \;\leq\; \frac{Z_{i,\mathcal{B}\setminus\mathcal{B}_i''}}{Z_{i,\mathcal{B}_i''}}.$$

Observe that $Z_{i,\mathcal{B}\setminus\mathcal{B}_i''}$ is a sum of $B - \bar{B}$ terms and, since $\mathrm{sim}(\cdot,\cdot) \in [-1, 1]$, each term is bounded above by e. Similarly, $Z_{i,\mathcal{B}_i''}$ is a sum of $\bar{B}$ terms, each bounded below by e^{-1}. Therefore,

$$\frac{Z_{i,\mathcal{B}\setminus\mathcal{B}_i''}}{Z_{i,\mathcal{B}_i''}} \;\leq\; \frac{\mathrm{e}(B-\bar{B})}{\mathrm{e}^{-1}\bar{B}} \;=\; \frac{\mathrm{e}^2(B-\bar{B})}{\bar{B}} \;\leq\; \frac{\mathrm{e}^2(B - B(1-\frac{1}{C}-\epsilon))}{B\left(1-\frac{1}{C}-\epsilon\right)} \;=\; \frac{\mathrm{e}^2(\frac{1}{C}+\epsilon)}{1-\frac{1}{C}-\epsilon} \;=\; \frac{\mathrm{e}^2(1+\epsilon C)}{C(1-\epsilon)-1}.$$

Thus, we deduce

$$\mathbb{E}_{\mathcal{B},\mathcal{B}_i''|A_i}\left[\ell_{i,\mathcal{B}}(f) - \ell_{i,\mathcal{B}_i''}(f)\right] \;\leq\; \frac{\mathrm{e}^2(1+\epsilon C)}{C(1-\epsilon)-1}.$$

On the complement $\bar{A}_i$ the batch $\mathcal{B}$ contains fewer than $\bar{B}$ negatives. In this case one may bound the losses uniformly. In fact, using

$$\log\Big(\sum_{j=1}^{B}\exp(\alpha_j)\Big) \;\leq\; \max\{\alpha_1, \ldots, \alpha_B\} + \log(B)$$

and the fact that $\mathrm{sim}(\cdot,\cdot) \in [-1,1]$, we obtain

$$|\ell_{i,\mathcal{B}}(f)| \;\leq\; |\mathrm{sim}(z_i, z_i')| + \log\left(\sum_{j\in\mathcal{B}} \exp(\mathrm{sim}(z_i, z_j)) + \exp(\mathrm{sim}(z_i, z_j'))\right) \;\leq\; \log(2B) + 2,$$

and similarly,

$$\left|\ell_{i,\mathcal{B}_i'}(f)\right| \;\leq\; \log(2\bar{B}) + 2 \;\leq\; \log(2B) + 2.$$

Define the indicator variable $Y_j = \mathbf{1}[y_j \neq y_i]$ so that $\mathbb{E}[Y_j] = 1 - \frac{1}{C}$. By Hoeffding's inequality,

$$\mathbb{P}\left[\sum_{j=1}^{B} Y_j \leq \lceil B(1-\tfrac{1}{C}) - B\epsilon\rceil\right] \;=\; \mathbb{P}\left[\sum_{j=1}^{B} Y_j \leq B(1-\tfrac{1}{C}) - B\epsilon\right] \;\leq\; \exp(-2B\epsilon^2).$$

Hence, $\mathbb{P}[\bar{A}_i] \leq \exp\left(-B\epsilon^2\right)$ and the contribution of the event $\bar{A}_i$ is bounded by

$$\begin{aligned}
&\mathbb{E}_i\left[\mathbb{P}[\bar{A}_i]\cdot\left(\mathbb{E}_{\mathcal{B}|\bar{A}_i}\left[\ell_{i,\mathcal{B}}(f)\right] - \mathbb{E}_{\mathcal{B}_i''}\left[\ell_{i,\mathcal{B}_i''}(f)\right]\right)\right]\\
&\leq\; \mathbb{E}_i\left[\mathbb{P}[\bar{A}_i]\cdot\left(\mathbb{E}_{\mathcal{B}|\bar{A}_i}\left[|\ell_{i,\mathcal{B}}(f)|\right] + \mathbb{E}_{\mathcal{B}_i''}\left[|\ell_{i,\mathcal{B}_i''}(f)|\right]\right)\right]\\
&\leq\; 2\left(\log(2B)+2\right)\exp\left(-B\epsilon^2\right).
\end{aligned}$$

Combining the bounds on A_i and $\bar{A}_i$, we conclude that

$$\mathcal{L}_B^{\mathrm{CL}}(f) - \mathcal{L}_{\bar{B}}^{\mathrm{NSCL}}(f) \;\leq\; \frac{\mathrm{e}^2(1+\epsilon C)}{C(1-\epsilon)-1} + 2\left(\log(2B)+2\right)\exp\left(-2B\epsilon^2\right).$$

Finally, we note that under A_i, $\ell_{i,\mathcal{B}}(f) \geq \ell_{i,\mathcal{B}_i''}(f)$. In addition, similar to the above:

$$\mathbb{E}_i\left[\mathbb{P}[\bar{A}_i]\cdot\left(\mathbb{E}_{\mathcal{B}|\bar{A}_i}\left[\ell_{i,\mathcal{B}}(f)\right] - \mathbb{E}_{\mathcal{B}_i''}\left[\ell_{i,\mathcal{B}_i''}(f)\right]\right)\right] \;\geq\; -2\left(\log(2B)+2\right)\exp\left(-B\epsilon^2\right).$$

yields the corresponding lower bound:

$$\mathcal{L}_B^{\mathrm{CL}}(f) - \mathcal{L}_{\bar{B}}^{\mathrm{NSCL}}(f) \;\geq\; -2\left(\log(2B)+2\right)\exp\left(-2B\epsilon^2\right).$$

This completes the proof. □

C Proof of Prop. 1

Proposition 1. *Let $C' \geq 2$ and $m \geq 10$ be integers. Fix a feature map $f : \mathcal{X} \to \mathbb{R}^d$ and class-conditional distributions $D_1, \dots, D_{C'}$ over $\mathcal{X}$. We have:*

$$\mathrm{err}_{m,D}^{\mathrm{LP}}(f) \;\leq\; \mathrm{err}_{m,D}^{\mathrm{NCC}}(f) \;\leq\; (C'-1)\left[14\mathrm{Avg}_{i\neq j}[\tilde{V}_f(D_i,D_j)] + \tfrac{28.5}{m}\mathrm{Avg}_{i\neq j}[V_f(D_i,D_j)]\right]. \tag{4}$$

Proof. The proof is a technical modification of the proof of Prop. 7 in [82]. Fix a pair of classes $i \neq j$. Draw an independent dataset $\widehat{S}_i = \{x_{i,1},\dots,x_{i,m}\} \sim D_i^m$ and $\widehat{S}_j = \{x_{j,1},\dots,x_{j,m}\} \sim D_j^m$. Let $\widehat{\mu}_i = \frac{1}{m}\sum_{s=1}^{m} f(x_{i,s})$.

For a test point $x_i \sim D_i$ with embedding $z_i = f(x_i)$, the NCC rule predicts class j instead of i if and only if $\|z_i - \widehat{\mu}_j\|_2 \leq \|z_i - \widehat{\mu}_i\|_2$. By the triangle inequality and union bound:

$$\begin{aligned}
&\Pr\left[\|z_i - \widehat{\mu}_j\| \leq \|z_i - \widehat{\mu}_i\|\right]\\
&\leq\; \Pr\left[\|z_i - \mu_j\| \leq \|z_i - \mu_i\| + \|\widehat{\mu}_i - \mu_i\| + \|\widehat{\mu}_j - \mu_j\|\right]\\
&\leq\; \Pr\left[\|z_i - \mu_j\| \leq \|z_i - \mu_i\| + 2\Delta_{ij} \;\vee\; \|\widehat{\mu}_i - \mu_i\| \geq \Delta_{ij} \;\vee\; \|\widehat{\mu}_j - \mu_j\| \geq \Delta_{ij}\right]\\
&\leq\; \Pr\left[\|z_i - \mu_j\| \leq \|z_i - \mu_i\| + 2\Delta_{ij}\right] + \Pr\left[\|\widehat{\mu}_i - \mu_i\| \geq \Delta_{ij}\right] + \Pr\left[\|\widehat{\mu}_j - \mu_j\| \geq \Delta_{ij}\right].
\end{aligned}$$

By Markov's inequality,

$$\Pr\left[\|\widehat{\mu}_i - \mu_i\| \geq \Delta_{ij}\right] \;\leq\; \frac{\mathrm{Var}(\widehat{\mu}_i)}{\Delta_{ij}^2} \;=\; \frac{\mathrm{Var}(z_i)}{\Delta_{ij}^2 m^2},$$

and similarly

$$\Pr\left[\|\widehat{\mu}_j - \mu_j\| \geq \Delta_{ij}\right] \;\leq\; \frac{\mathrm{Var}(z_j)}{\Delta_{ij}^2 m^2}.$$

Assume that $\|z_i - \mu_j\| \leq \|z_i - \mu_i\| + 2\Delta_{ij}$ and $\|z_i - \mu_i\| \leq \alpha_{ij}$. Then,

$$\begin{aligned}
\|z_i - \mu_i\|^2 + 4\alpha_{ij}\Delta_{ij} + 4\Delta_{ij}^2 \;&\geq\; \|z_i - \mu_i\|^2 + 2\|z_i - \mu_i\| \cdot 2\Delta_{ij} + 4\Delta_{ij}^2 \\
&=\; (\|z_i - \mu_i\| + 2\Delta_{ij})^2 \\
&\geq\; \|z_i - \mu_j\|^2 \\
&=\; \|(z_i - \mu_i) - (\mu_j - \mu_i)\|^2 \\
&=\; \|z_i - \mu_i\|^2 - 2(z_i - \mu_i)^\top(\mu_j - \mu_i) + d_{ij}^2.
\end{aligned}$$

Therefore, we have $2(z_i - \mu_i)^\top(\mu_j - \mu_i) \;\geq\; d_{ij}^2 - 4\alpha_{ij}\Delta_{ij} - 4\Delta_{ij}^2$, which can also be written as follows

$$(z_i - \mu_i)^\top u_{ij} \;\geq\; \beta_{ij} \;:=\; \frac{1}{2}d_{ij} - \frac{2(\alpha_{ij}\Delta_{ij} + \Delta_{ij}^2)}{d_{ij}}.$$

Next we upper bound the terms $\Pr[(z_i - \mu_i)^\top u_{ij} \geq \beta_{ij}]$ and $\Pr[\|z_i - \mu_i\| \geq \alpha_{ij}]$. Namely, by Markov's inequality,

$$\Pr[(z_i - \mu_i)^\top u_{ij} \geq \beta_{ij}] \;\leq\; \frac{\mathrm{Var}((z_i - \mu_i)^\top u_{ij})}{\beta_{ij}^2},$$

and

$$\Pr[\|z_i - \mu_i\| \geq \alpha_{ij}] \;\leq\; \frac{\mathrm{Var}(z_i)}{\alpha_{ij}^2}.$$

Therefore, by combining everything together, we have:

$$\Pr\left[\|z_i - \widehat{\mu}_j\| \leq \|z_i - \widehat{\mu}_i\|\right] \;\leq\; \frac{\mathrm{Var}((z_i - \mu_i)^\top u_{ij})}{\beta_{ij}^2} + \frac{\mathrm{Var}(z_i)}{\alpha_{ij}^2} + \frac{\mathrm{Var}(z_i) + \mathrm{Var}(z_j)}{\Delta_{ij}^2 m^2}.$$

We choose $\alpha_{ij} = \frac{d_{ij}\sqrt{m}}{\gamma_1}$ and $\Delta_{ij} = \frac{d_{ij}}{\gamma_2\sqrt{m}}$, then $\beta_{ij} = d_{ij}\left(\frac{1}{2} - \frac{2}{\gamma_1\gamma_2} - \frac{2}{\gamma_2^2 m}\right)^2$. We have:

$$\begin{aligned}
&\Pr\left[\|z_i - \widehat{\mu}_j\| \leq \|z_i - \widehat{\mu}_i\|\right] \\
&\leq\; \left(\tfrac{1}{2} - \tfrac{2}{\gamma_1\gamma_2} - \tfrac{2}{\gamma_2^2 m}\right)^{-2} \frac{\mathrm{Var}((z_i - \mu_i)^\top u_{ij})}{d_{ij}^2} + \frac{\gamma_1^2 \mathrm{Var}(z_i)}{m d_{ij}^2} + \frac{\gamma_2^2(\mathrm{Var}(z_i) + \mathrm{Var}(z_j))}{m d_{ij}^2} \\
&\leq\; \left(\tfrac{1}{2} - \tfrac{2}{\gamma_1\gamma_2} - \tfrac{2}{\gamma_2^2 m}\right)^{-2} \tilde{V}_f(D_i, D_j) + \frac{\gamma_1^2 V_f(D_j, D_i) + 2\gamma_2^2 V_f(D_i, D_j)}{m}
\end{aligned}$$

For the true class i a mistake can occur in favor of any $j \neq i$. A union bound therefore gives

$$\begin{aligned}
\mathrm{err}_{m,D}^{\mathrm{NCC}}(f) \;&=\; \frac{1}{C'}\sum_{i=1}^{C'} \Pr\big[\text{misclassify } x \sim D_i\big] \\
&\leq\; (C'-1)\left[\left(\tfrac{1}{2} - \tfrac{2}{\gamma_1\gamma_2} - \tfrac{2}{\gamma_2^2 m}\right)^{-2} \mathrm{Avg}_{i\neq j}[\tilde{V}_f(D_i, D_j)] + \tfrac{\gamma_1^2 + 2\gamma_2^2}{m}\mathrm{Avg}_{i\neq j}[V_f(D_i, D_j)]\right].
\end{aligned}$$

Since the bound holds uniformly for all $\gamma_1, \gamma_2 > 0$ such that $\frac{1}{2} - \frac{2}{\gamma_1\gamma_2} - \frac{2}{\gamma_2^2 m} > 0$, we can pick γ_1, γ_2 such that their product is $a \geq 5$ and minimize $\gamma_1^2 + 2\gamma_2^2$. In this case, $\gamma_1 = 2^{1/4}\sqrt{a}$, $\gamma_2 = 2^{-1/4}\sqrt{a}$ and $\gamma_1^2 + 2\gamma_2^2 = 2\sqrt{2}a$. Let $\tilde{V}_f = \mathrm{Avg}_{i\neq j}[\tilde{V}_f(D_i, D_j)]$ and $V_f = \mathrm{Avg}_{i\neq j}[V_f(D_i, D_j)]$. The bound then becomes:

$$\mathrm{err}_{m,D}^{\mathrm{NCC}}(f) \;\leq\; (C'-1)\inf_{a\geq 5}\left[\left(\tfrac{1}{2} - \tfrac{2}{a} - \tfrac{2^{3/2}}{am}\right)^{-2}\tilde{V}_f + \tfrac{2\sqrt{2}a}{m}V_f\right]. \tag{8}$$

For $a = 10$ we obtain that $\left(\frac{1}{2} - \frac{2}{a} - \frac{2^{3/2}}{am}\right)^{-2} < 14$ and $2\sqrt{2}a < 28.5$ which gives us the coefficients above.

Finally, $\mathrm{err}_{m,D}^{\mathrm{LP}} \leq \mathrm{err}_{m,D}^{\mathrm{NCC}}$ because the optimal linear probe cannot perform worse than the NCC classifier, since the NCC classifier is a special case of a linear probe. $\square$

Corollary 1. *Let $C' \geq 2$ and $m \geq 10$ be integers. Fix a feature map $f : \mathcal{X} \to \mathbb{R}^d$ and class-conditional distributions $D_1, \ldots, D_{C'}$ over $\mathcal{X}$. Denote $\tilde{V}_f = \mathrm{Avg}_{i \neq j}[\tilde{V}_f(D_i, D_j)]$ and $V_f = \mathrm{Avg}_{i\neq j}[V_f(D_i, D_j)]$. We have:*

$$\mathrm{err}^{\mathrm{NCC}}_{m,D}(f) \;\leq\; (C'-1)\frac{\sqrt{2}V_f}{2Am}(y^* + 2A)\big(y^* + 4A\big),$$

where

$$A \;=\; 2 + \frac{2^{3/2}}{m} \quad \text{and} \quad F \;=\; \frac{\tilde{V}_f A m}{\sqrt{2}V_f}$$

and

$$\begin{aligned} A^2 \geq \frac{8F}{27} \quad &\Longrightarrow \quad y^* \;=\; \sqrt[3]{4F\Big(2A + \sqrt{A^2 - \tfrac{8F}{27}}\Big)} + \sqrt[3]{4F\Big(2A - \sqrt{A^2 - \tfrac{8F}{27}}\Big)} \\ A^2 < \frac{8F}{27} \quad &\Longrightarrow \quad y^* \;=\; 4\sqrt{\frac{2F}{3}}\cos\left(\tfrac{1}{3}\arccos\Big(3A\sqrt{\frac{3}{8F}}\Big)\right) \end{aligned}$$

Proof. In order to find the $a \geq 5$ that minimizes the RHS, we consider $\tilde{V}_f, V_f, m > 0$ be fixed and consider the objective

$$E(a) \;=\; \left(\frac{1}{2} - \frac{2}{a} - \frac{2^{3/2}}{am}\right)^{-2}\tilde{V}_f + \frac{2\sqrt{2}V_f}{m}a.$$

Define an auxiliary function

$$c(a) \;:=\; \frac{1}{2} - \frac{A}{a} = \frac{a-2A}{2a}.$$

Then

$$E(a) = c(a)^{-2}\tilde{V}_f + Ba, \qquad B \;:=\; \frac{2\sqrt{2}V_f}{m} \;>\; 0,$$

Next, we take the derivative of the function with respect to a:

$$\frac{\mathrm{d}}{\mathrm{d}a}E(a) \;=\; -\frac{2\tilde{V}_f A}{a^2 c(a)^3} + B = \;0 \qquad \Longrightarrow \qquad a^2 c(a)^3 \;=\; \frac{2\tilde{V}_f A}{B}.$$

Insert $c(a) = \dfrac{a-2A}{2a} =: \dfrac{y}{2(y+2A)}$ into the previous equation:

$$y^3 - 8Fy - 16FA \;=\; 0, \qquad F \;:=\; \frac{\tilde{V}_f A m}{\sqrt{2}V_f} > 0.$$

Therefore, we need to solve a depressed cubic $y^3 - 8Fy - 16FA = 0$. We first recall the *generic trigonometric solution* of a depressed cubic and then specialize it to the present parameters.

Generic form. For a depressed cubic

$$y^3 + py + q \;=\; 0,$$

define

$$R \;:=\; 2\sqrt{-\frac{p}{3}}, \qquad \Phi \;:=\; \arccos\Big(\frac{3q}{2p}\sqrt{-\frac{3}{p}}\Big), \qquad 0 \leq \Phi \leq \pi.$$

The three real roots are

$$y_k \;=\; R\cos\Big(\tfrac{\Phi + 2\pi k}{3}\Big), \qquad k = 0, 1, 2.$$

If the *discriminant* $\Delta := \big(\frac{q}{2}\big)^2 + \big(\frac{p}{3}\big)^3$ is non–negative, two of the roots become complex conjugates and Cardano's radicals give the single real one.

Our cubic. In our derivation we have $p = -8F < 0$, $q = -16FA$, $\Delta = 64F^2(A^2 - \frac{8F}{27})$, $R = 4\sqrt{\frac{2F}{3}}$ and $\Phi = \arccos(3A\sqrt{\frac{3}{8F}})$.

Case $\Delta < 0$ **(**$A^2 < \frac{8F}{27}$**).** In this case all three roots are real; the unique positive one—needed because $a = 2A + y$ must exceed $2A$—is

$$y^* \;=\; R\cos\!\Big(\tfrac{\Phi}{3}\Big) \;=\; 4\sqrt{\frac{2F}{3}}\,\cos\!\Big(\tfrac{1}{3}\arccos\!\big(3A\sqrt{\tfrac{3}{8F}}\big)\Big).$$

Case $\Delta \geq 0$ **(**$A^2 \geq \frac{8F}{27}$**).** Only one root is real; Cardano's formula gives it:

$$y^* \;=\; \sqrt[3]{4F\Big(2A + \sqrt{A^2 - \tfrac{8F}{27}}\Big)} + \sqrt[3]{4F\Big(2A - \sqrt{A^2 - \tfrac{8F}{27}}\Big)}.$$

(The two expressions coincide when $\Delta = 0$).

With y^* from either line we set $a^* = 2A + y^*$. The objective therefore attains its minimum at $E(a^*) = \frac{\sqrt{2}V_f}{2Am}\, a^*\big(a^* + 2A\big)$. $\square$

D Proof of Thm. 2

Theorem 2. *Let $d \geq C-1$ and let $S = \{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times [C]$ be a balanced labeled dataset with C classes. Suppose f is a global minimizer of the supervised contrastive loss $\mathcal{L}^{\mathrm{NSCL}}(f)$ (over all functions $f : \mathcal{X} \to \mathbb{R}^d$). Then, the representations satisfy the following properties:*

1. ***Augmentation Collapse:*** *For each $i \in [N]$ and for every pair $l_1, l_2 \in [K]$, we have $z_i^{l_1} = z_i^{l_2}$.*
2. ***Within-Class Collapse:*** *For any two samples x_i and x_j with the same label ($y_i = y_j$), their representations coincide: $z_i = z_j$. Namely, each class has a unique class embedding.*
3. ***Simplex Equiangular Tight Frame:*** *Let $\{\mu_1, \dots, \mu_C\}$ denote the set of class-center embeddings. These vectors form a simplex ETF in $\mathbb{R}^d$; specifically, they satisfy $\sum_{c=1}^C \mu_c = 0$, $\|\mu_c\|_2 = \|\mu_{c'}\|_2$ and $\langle \mu_c, \mu_{c'}\rangle = -\frac{\|\mu_c\|_2^2}{C-1}$ for all $c \neq c' \in [C]$.*

Proof. We divide the proof into three parts. In the first part, we restate and relax the problem. In the second part, we show that the embedding vectors of augmentations of the same sample collapse to the same vector. In the third part, we demonstrate that the embeddings of samples from the same class collapse to the same vectors, and the class vectors form a Simplex ETF.

Restating the problem. Since we focus on the unconstrained features model [78, 79], we can rewrite $\mathcal{L}^{\mathrm{NSCL}}(f)$ in the following way:

$$\mathcal{L}^{\mathrm{NSCL}}(Z) \;=\; -\frac{1}{K^2N}\sum_{l_1,l_2=1}^{K}\sum_{i=1}^{N}\log\!\left(\frac{\exp\!\big(\mathrm{sim}(z_i^{l_1}, z_i^{l_2})\big)}{\sum_{l_3=1}^{K}\sum_{\substack{j=1\\ y_j\neq y_i}}^{N}\exp\!\big(\mathrm{sim}(z_i^{l_1}, z_j^{l_3})\big)}\right),$$

where $Z = (z_i^l)_{i\in[N], l\in[K]}$ denotes a collection of learnable representations, where each z_i^l is the embedding of sample x_i under augmentation l.

Since one may either normalize the vectors externally or inside the cosine similarity and because unit norm constraint is stronger than a norm upper bound constraint, we have:

$$\begin{aligned}
&\min_Z\left\{-\frac{1}{K^2N}\sum_{l_1,l_2=1}^{K}\sum_{i=1}^{N}\log\!\left(\frac{\exp\!\big(\mathrm{sim}(z_i^{l_1}, z_i^{l_2})\big)}{\sum_{l_3=1}^{K}\sum_{j:y_j\neq y_i}^{N}\exp\!\big(\mathrm{sim}(z_i^{l_1}, z_j^{l_3})\big)}\right)\right\}\\
&=\min_Z\left\{-\frac{1}{K^2N}\sum_{l_1,l_2=1}^{K}\sum_{i=1}^{N}\log\!\left(\frac{\exp\!\big(\langle z_i^{l_1}, z_i^{l_2}\rangle\big)}{\sum_{l_3=1}^{K}\sum_{j:y_j\neq y_i}^{N}\exp\!\big(\langle z_i^{l_1}, z_j^{l_3}\rangle\big)}\right)\ \text{s.t.}\ \forall i\in[N], l\in[K]:\ \|z_i^l\|_2 = 1\right\}\\
&\geq\min_Z\left\{\underbrace{-\frac{1}{K^2N}\sum_{l_1,l_2=1}^{K}\sum_{i=1}^{N}\log\!\left(\frac{\exp\!\big(\langle z_i^{l_1}, z_i^{l_2}\rangle\big)}{\sum_{l_3=1}^{K}\sum_{j:y_j\neq y_i}^{N}\exp\!\big(\langle z_i^{l_1}, z_j^{l_3}\rangle\big)}\right)}_{=:\mathcal{Q}(Z)}\ \text{s.t.}\ \forall i\in[N], l\in[K]:\ \|z_i^l\|_2 \leq 1\right\}.
\end{aligned}$$

We will show that the global minimum of the last optimization problem is obtained when we have augmentation collapse, within-class collapse, and a simplex ETF behavior. In particular, this will give us the result that the solution to the latter problems is achieved in the same way.

Augmentation Collapse. We begin by proving that for every $i_0 \in [N]$, the vectors $z_{i_0}^1, \dots, z_{i_0}^K$ are identical at any global minimum. Fix some index $i_0 \in [N]$ and suppose for contradiction that there exist two distinct augmentations $l_1^* \neq l_2^*$ with $z_{i_0}^{l_1^*} \neq z_{i_0}^{l_2^*}$. Define the averaged vector $\tilde{z}_{i_0} = \frac{1}{K}\sum_l z_{i_0}^l$ and let $\tilde{Z}$ be the collection obtained by replacing $z_{i_0}^l$ by $\tilde{z}_{i_0}$ for all $l \in [K]$. We will show that $\mathcal{Q}(Z) > \mathcal{Q}(\tilde{Z})$.

Consider the loss function:

$$\mathcal{Q}(Z) = -\frac{1}{K^2 N}\sum_{i=1}^{N}\sum_{l_1,l_2=1}^{K}\log\left(\frac{\exp\big(\langle z_i^{l_1}, z_i^{l_2}\rangle\big)}{\sum_{l_3=1}^{K}\sum_{j:\, y_j\neq y_i}\exp\big(\langle z_i^{l_1}, z_j^{l_3}\rangle\big)}\right).$$

Now, fix the index i_0 and split the sum over i into the contribution from i_0 and the contributions from all other indices $i \neq i_0$. In addition, let $\delta_{i,j}^{l_1,l_2} = \exp\big(\langle z_i^{l_1}, z_j^{l_2}\rangle\big)$. With this separation, we obtain

$$\begin{aligned}
\mathcal{Q}(Z) \;&=\; -\frac{1}{K^2N}\sum_{l_1,l_2=1}^{K}\sum_{i=1}^{N}\log\left(\frac{\exp\big(\langle z_i^{l_1}, z_i^{l_2}\rangle\big)}{\sum_{l_3=1}^{K}\sum_{j:y_j\neq y_i}\exp\big(\langle z_i^{l_1}, z_j^{l_3}\rangle\big)}\right)\\
&=\; \frac{1}{N}\left[-\frac{1}{K^2}\sum_{l_1,l_2=1}^{K}\langle z_{i_0}^{l_1}, z_{i_0}^{l_2}\rangle - \frac{1}{K^2}\sum_{l_1,l_2=1}^{K}\sum_{i\neq i_0} -\langle z_i^{l_1}, z_i^{l_2}\rangle\right]\\
&\quad + \frac{1}{K^2N}\sum_{l_1,l_2=1}^{K}\sum_{i\neq i_0}^{N}\log\left(\sum_{l_3=1}^{K}\sum_{j:y_j\neq y_i}\exp\big(\langle z_i^{l_1}, z_j^{l_3}\rangle\big)\right)\\
&\quad + \frac{1}{KN}\sum_{l_1=1}^{K}\log\left(\sum_{l_3=1}^{K}\sum_{j:y_j\neq y_{i_0}}\exp\big(\langle z_{i_0}^{l_1}, z_j^{l_3}\rangle\big)\right).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathcal{Q}(\tilde{Z}) \;&=\frac{1}{N}\left[-\langle \tilde{z}_{i_0}, \tilde{z}_{i_0}\rangle - \frac{1}{K^2}\sum_{l_1,l_2=1}^{K}\sum_{i\neq i_0} -\langle z_i^{l_1}, z_i^{l_2}\rangle\right]\\
&\quad + \frac{1}{K^2N}\sum_{l_1,l_2=1}^{K}\sum_{i\neq i_0}^{N}\log\left(\sum_{l_3=1}^{K}\sum_{\substack{j\in[N]\\ y_j\neq y_i\\ j\neq i_0}}\exp\big(\langle z_i^{l_1}, z_j^{l_3}\rangle\big) + K\exp\big(\langle z_i^{l_1}, \tilde{z}_{i_0}\rangle\big)\right)\\
&\quad + \frac{1}{N}\log\left(\sum_{l_3=1}^{K}\sum_{j:y_j\neq y_{i_0}}\exp\big(\langle \tilde{z}_{i_0}, z_j^{l_3}\rangle\big)\right).
\end{aligned}$$

By definition, since $\tilde{z}_{i_0} = \frac{1}{K}\sum_l z_{i_0}^l$, its squared norm is given by

$$\langle \tilde{z}_{i_0}, \tilde{z}_{i_0}\rangle \;=\; \|\tilde{z}_{i_0}\|_2^2 \;=\; \frac{1}{K^2}\sum_{l_1,l_2=1}^{K}\langle z_{i_0}^{l_1}, z_{i_0}^{l_2}\rangle.$$

Since $z_i^{l_1} \neq 0$ for every $i \in [N]$ and $l_1 \in [K]$ (otherwise the cosine similarity would be undefined), the function

$$h_i^{l_1}(x) \;=\; \exp\big(\langle z_i^{l_1}, x\rangle\big)$$

is convex in x. Consequently, by Jensen's inequality, for any subset S and any collection of vectors $\{z_j^{l_3} : j \in S\}$ we have

$$\frac{1}{|S|}\sum_{j\in S}\exp\big(\langle z_i^{l_1}, z_j^{l_3}\rangle\big) \;\geq\; \exp\left(\left\langle z_i^{l_1}, \frac{1}{|S|}\sum_{j\in S} z_j^{l_3}\right\rangle\right),$$

with equality if and only if all vectors $z_j^{l_3}$ in the subset are identical.

Applying this inequality to the relevant subsets yields

$$\log\left(\sum_{l_3=1}^{K}\sum_{j:\,y_j\neq y_i}\exp\big(\langle z_i^{l_1}, z_j^{l_3}\rangle\big)\right) > \log\left(\sum_{l_3=1}^{K}\sum_{\substack{j:\,y_j\neq y_i\\ j\neq i_0}}\exp\big(\langle z_i^{l_1}, z_j^{l_3}\rangle\big) + K\,\exp\big(\langle z_i^{l_1}, \tilde{z}_{i_0}\rangle\big)\right),$$

where the strict inequality follows from the assumption that for some indices l_1^* and l_2^* (with $l_1^* \neq l_2^*$), we have $z_{i_0}^{l_1^*} \neq z_{i_0}^{l_2^*}$.

Next, define the function

$$G(x) \;=\; \log\left(\sum_{l_3=1}^{K}\sum_{j:\,y_j\neq y_{i_0}}\exp\big(\langle x, z_j^{l_3}\rangle\big)\right).$$

It is well known that $G(x)$ is convex in x. Therefore, by Jensen's inequality we obtain

$$\frac{1}{K}\sum_{l_1=1}^{K} G\big(z_{i_0}^{l_1}\big) \;\geq\; G\Big(\frac{1}{K}\sum_{l_1=1}^{K} z_{i_0}^{l_1}\Big),$$

which can be written equivalently as

$$\frac{1}{K}\sum_{l_1=1}^{K}\log\left(\sum_{l_3=1}^{K}\sum_{j:\,y_j\neq y_{i_0}}\exp\big(\langle z_{i_0}^{l_1}, z_j^{l_3}\rangle\big)\right) \;\geq\; \log\left(\sum_{l_3=1}^{K}\sum_{j:\,y_j\neq y_{i_0}}\exp\big(\langle \tilde{z}_{i_0}, z_j^{l_3}\rangle\big)\right).$$

Combining this inequality with the previous expansion, we deduce that $\mathcal{Q}(\tilde{Z}) < \mathcal{Q}(Z)$, which contradicts the assumption that Z is a global minimizer of the loss. Consequently, for each $i_0 \in [N]$, it must be that all vectors $z_{i_0}^1, z_{i_0}^2, \ldots, z_{i_0}^K$ are equal at a global minimum.

Within-class collapse and Simplex ETF. Since the augmentations collapse, with no loss of generality we assume that $K = 1$. We now analyze the embeddings z_i by applying the arithmetic–geometric mean (AGM) inequality to the $n(C-1)$ positive numbers

$$\big\{\exp\big(\langle z_i, z_j\rangle\big) \mid j : y_j \neq y_i\big\}\,,$$

for each fixed $i \in [N]$. The AGM inequality yields

$$\frac{1}{n(C-1)}\sum_{j:y_j\neq y_i}\exp\big(\langle z_i, z_j\rangle\big) \;\geq\; \left(\prod_{j:y_j\neq y_i}\exp\big(\langle z_i, z_j\rangle\big)\right)^{\frac{1}{n(C-1)}}, \tag{9}$$

Taking natural logarithms on both sides gives

$$\log\left(\sum_{j:y_j\neq y_i}\exp\big(\langle z_i, z_j\rangle\big)\right) \geq \log(n(C-1)) + \frac{1}{n(C-1)}\sum_{j:y_j\neq y_i}\langle z_i, z_j\rangle.$$

Averaging this inequality over $i \in [N]$ yields

$$\frac{1}{N}\sum_{i=1}^{N}\log\left(\sum_{j:y_j\neq y_i}\exp\big(\langle z_i, z_j\rangle\big)\right) \;\geq\; \log(n(C-1)) + \frac{1}{Nn(C-1)}\sum_{i=1}^{N}\sum_{j:y_j\neq y_i}\langle z_i, z_j\rangle. \tag{10}$$

Notice that the double sum of inner products can be rewritten as

$$\begin{aligned}\sum_{i=1}^{N}\sum_{j:y_j\neq y_i}\langle z_i, z_j\rangle \;&=\; \sum_{i,j=1}^{N}\langle z_i, z_j\rangle - \sum_{i=1}^{N}\sum_{j:y_j=y_i}\langle z_i, z_j\rangle\\ &=\; \Big\|\sum_{i=1}^{N} z_i\Big\|_2^2 - \sum_{c=1}^{C}\Big\|\sum_{i:y_i=c} z_i\Big\|_2^2\\ &=\; \Big\|n\sum_{c=1}^{C}\mu_c\Big\|_2^2 - \sum_{c=1}^{C}\|n\mu_c\|_2^2\end{aligned}$$

where we define the class-mean embeddings as $\mu_c = \frac{1}{n}\sum_{i:y_i=c} z_i$. Substituting this identity into (10) yields

$$\frac{1}{N}\sum_{i=1}^{N}\log\Bigg(\sum_{j:y_j\neq y_i}\exp\big(\langle z_i, z_j\rangle\big)\Bigg) \;\geq\; \log(n(C-1)) + \frac{1}{C(C-1)}\Big(\Big\|\sum_{c=1}^{C}\mu_c\Big\|_2^2 - \sum_{c=1}^{C}\|\mu_c\|_2^2\Big). \tag{11}$$

By Jensen's inequality,

$$\|\mu_c\|_2^2 \;=\; \Big\|\sum_{i:y_i=c} z_i\Big\|_2^2 \;\leq\; \tfrac{1}{n}\sum_{i:y_i=c}\|z_i\|_2^2 \;\leq\; 1, \tag{12}$$

it follows that

$$\sum_{i=1}^{N}\sum_{j:y_j\neq y_i}\langle z_i, z_j\rangle \;\geq\; n^2\Big\|\sum_{c=1}^{C}\mu_c\Big\|_2^2 - n^2 C \tag{13}$$

Substituting this estimate into (10) gives

$$\frac{1}{N}\sum_{i=1}^{N}\log\Bigg(\sum_{j:y_j\neq y_i}\exp\big(\langle z_i, z_j\rangle\big)\Bigg) \;\geq\; \log(n(C-1)) + \frac{1}{C(C-1)}\Big(\Big\|\sum_{c=1}^{C}\mu_c\Big\|_2^2 - C\Big).$$

Substituting the bound (11) into the expression for $\mathcal{Q}(Z)$ (the loss) results in

$$\mathcal{Q}(Z) \;\geq\; -\frac{1}{N}\sum_{i=1}^{N}\|z_i\|_2^2 + \log(n(C-1)) + \frac{1}{C(C-1)}\Big(\Big\|\sum_{c=1}^{C}\mu_c\Big\|_2^2 - C\Big). \tag{14}$$

Since

$$-\frac{1}{N}\sum_{i=1}^{N}\|z_i\|_2^2 \;\geq\; -1 \quad \text{and} \quad \Big\|\sum_{c=1}^{C}\mu_c\Big\|_2^2 \;\geq\; 0, \tag{15}$$

we obtain the lower bound

$$\mathcal{Q}(Z) \;\geq\; \log\big(n(C-1)\big) - 1 - \frac{1}{C-1}. \tag{16}$$

Equality in (16) is achieved only if all the preceding inequalities hold as equalities. In particular, equality in (15) requires $\sum_{c=1}^{C}\mu_c = 0$ and $\forall i \in [N] : \|z_i\|_2 = 1$. Similarly, equality in (12) (used in the derivation) forces $\Big\|\sum_{i:y_i=c} z_i\Big\|_2^2 = \frac{1}{n}\sum_{i:y_i=c}\|z_i\|_2^2$ which, via Jensen's inequality, implies that $z_i = \mu_{y_i}$ for all $i \in [N]$. Furthermore, equality in (9) is attained if and only if

$$\forall i : \exp\big(\langle z_i, z_j\rangle\big) \text{ is constant w.r.t. } j \text{ such that } y_j \neq y_i.$$

In particular, by applying natural logarithm in both sides, and considering that fact that $z_i = \mu_{y_i}$

$$\forall c \in [C] : \langle \mu_c, \mu_{c'}\rangle \text{ is constant for all } c' \neq c.$$

In particular, $\langle \mu_i, \mu_j\rangle = \langle \mu_r, \mu_j\rangle = \langle \mu_r, \mu_t\rangle$ and we have the more general equation:

$$\langle \mu_c, \mu_{c'}\rangle \text{ is constant for all } c \neq c' \in [C].$$

Together, these conditions exactly characterize a simplex equiangular tight frame (ETF); in particular, they imply that

$$\|\mu_c\|_2 \;=\; 1, \quad \sum_{c=1}^{C}\mu_c \;=\; 0, \quad \text{and} \quad \langle \mu_c, \mu_{c'}\rangle \;=\; -\frac{1}{C-1} \text{ forall } c \neq c' \in [C].$$

A straightforward calculation shows that

$$\mathcal{Q}(Z) \;=\; -1 + \frac{1}{C}\sum_{i=1}^{C}\log\Big((C-1)\exp\Big(-\tfrac{1}{C-1}\Big)\Big) \;=\; \log(C-1) - 1 - \frac{1}{C-1},$$

so that the lower bound in (14) is attained. Consequently, a set of vectors Z is a global minimizer of $\mathcal{Q}(Z)$ if and only if it forms a simplex equiangular tight frame. □