RETHINKING REPRESENTATION GEOMETRY IN SUPERVISED CONTRASTIVE LEARNING UNDER LONG-TAILED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised contrastive learning and its balanced variants have demonstrated their effectiveness for representation learning under long-tailed distributions. However, although prior analyses suggest that their optimal representations may exhibit a regular simplex geometry, whether such a geometry is indeed exhibited remains insufficiently understood. In this work, we present a theoretical framework for these objectives that characterizes whether a regular simplex is formed at the optimum. By applying this to representative objectives, we show that these objectives fail to form a regular simplex at the optimum because they remain dependent on the number of samples per class. Motivated by this analysis, we introduce structural balancing that can be applied to contrastive objectives, and propose a fully balanced contrastive objective that provably forms a regular simplex geometry at its optimum. Both theoretical and empirical results demonstrate that the proposed method yields a more balanced representation space and improves performance on long-tailed recognition and transfer learning tasks. The code will be released.

1 INTRODUCTION

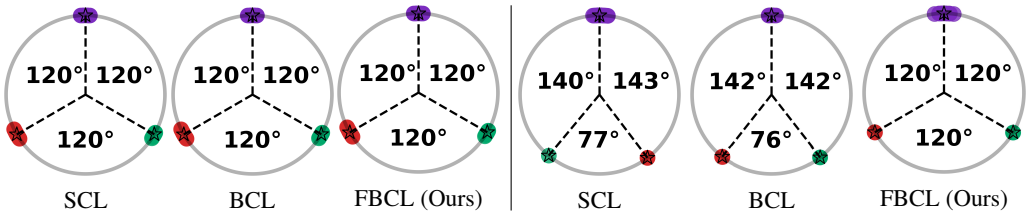
Deep learning has achieved remarkable success over the past decade, yet much of this progress relies on highly curated, balanced datasets (Krizhevsky, 2009; Deng et al., 2009), while real-world data often exhibit long-tailed distributions, in which a small number of head classes are represented by abundant samples while a large number of tail classes contain only a few samples (Van Horn et al., 2018). Models trained under such long-tailed distributions tend to learn skewed decision boundaries, which in turn degrades generalization performance on tail classes (Cao et al., 2019; Ren et al., 2020).

Supervised contrastive objectives have shown promising results in learning representations from long-tailed distributions (Kang et al., 2021; Wang et al., 2021; Cui et al., 2021; Li et al., 2022b; Zhu et al., 2022; Cui et al., 2023; Hou et al., 2023; Du et al., 2024; De Alvis et al., 2025; Yi et al., 2025). As in balanced distributions (Graf et al., 2021; Yi et al., 2026), characterizing an optimum of a loss is important for understanding the representation geometry the loss favors. However, the representation spaces induced at the minimum of these objectives remain insufficiently understood under long-tailed distributions. Among existing methods, balanced contrastive learning (BCL) (Zhu et al., 2022) showed that the loss is minimized if and only if the class centers form a regular simplex, suggesting that the ideal geometry remains optimal even under long-tailed distributions.

In this paper, we establish a theoretical framework for determining whether a regular simplex is a global minimizer of a supervised contrastive objective under long-tailed distributions. We first reduce the analysis from the feature space to the class-wise center space by deriving a lower bound of the loss, whose equality condition corresponds to class collapse. We then show that a center configuration that does not minimize the lower bound cannot minimize the original loss. Finally, we derive a condition that the lower bound must satisfy for a regular simplex to be its minimizer. We apply this framework to two widely used objectives for long-tailed recognition, SCL (Khosla et al., 2020) and BCL, and prove that the regular simplex is not a global minimizer of either loss. Fig. 1 shows this behavior with ImageNet samples.

We further investigate why this result holds by examining how class imbalance affects the loss. From a balancing perspective, SCL and BCL adjust the loss only within the contrastive loss defined for each anchor, the sample for which the contrastive loss is computed. Consequently, as the losses for individual anchors are aggregated into the overall loss, the absence of balancing this aggregation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) 1:1:1 (b) 1000:10:1

Figure 1: Visualizations of embedding spaces learned by three contrastive objectives (SCL, BCL, and FBCL) on a subset of ImageNet: three random classes sampled under (a) a balanced setting and (b) an imbalanced setting with class cardinality ratios of 1:1:1 and 1000:10:1, respectively. Angles are measured between class centers. Under (a), all three methods form a regular simplex; under (b), only FBCL forms it, whereas SCL and BCL fail.

leaves the loss dependent on the number of samples per class. This dependence is inherited by the lower bound, preventing the regular simplex from being a minimizer of the objective. Motivated by this observation, we present structural balancing, a method applicable to supervised contrastive objectives that balances the aggregation across anchors and removes the remaining dependence on the number of samples per class. Based on structural balancing, we propose Fully Balanced Contrastive Learning (FBCL), whose global minimizer is a regular simplex.

We also analyze the effect of structural balancing on the representation space by examining how the loss penalizes the intra-class variance of each class as the loss is minimized. Under SCL and BCL, the coefficient of the intra-class variance penalty is approximately proportional to class frequency, whereas under FBCL it is approximately independent of it. This suggests more balanced contraction across classes, which is further supported by our empirical analysis of intra-class variance. We further evaluate FBCL with other contrastive objectives and apply structural balancing to other long-tailed recognition methods. Results show that FBCL and structural balancing can be effectively combined with other methods and improve their performance.

Our main contributions are summarized as follows:

- We establish a framework for determining whether a regular simplex is a global minimizer of a supervised contrastive objective under long-tailed distributions. Applying this framework to SCL and BCL, we prove that the regular simplex is a global minimizer of neither loss.
- We observe that this failure occurs as a result of the remaining dependence on the number of samples per class in the loss. Based on this observation, we present structural balancing and propose Fully Balanced Contrastive Learning (FBCL), whose loss removes this dependence and admits the regular simplex as its global minimizer.
- We empirically evaluate FBCL on long-tailed recognition and transfer learning, and further examine structural balancing in combination with other methods. The results show consistent improvements in classification accuracy and transferability across these settings.

2 Related Work

Long-Tailed Recognition. Numerous approaches have been proposed to address long-tailed recognition. Re-sampling (He & Garcia, 2009; Buda et al., 2018; Shi et al., 2023) and re-weighting classification losses (Huang et al., 2016; Ren et al., 2018; Cui et al., 2019) were earlier methods in which the former balances the number of training samples and the latter assigns different weights to each training sample. Zhou et al. (2020) and Kang et al. (2020) propose a two-branch network and a two-stage learning framework, respectively, and show that such designs often yield better representations. Several works further demonstrate the effectiveness of these frameworks (Wang et al., 2021; Zhong et al., 2021; Li et al., 2022a; 2024a). Logit adjustment methods (Cao et al., 2019; Menon et al., 2021; Kini et al., 2021; Li et al., 2022a; Wang et al., 2023) enlarge the margins for tail classes. Transfer learning methods (Yin et al., 2019; Li et al., 2024a) enrich tail class features by leveraging abundant information from head classes. Recently, fine-tuning foundation models (Tian et al., 2022; Dong et al., 2023; Shi et al., 2024) has proven effective with efficient training schemes, yet the impact of imbalanced data on fine-tuning is not fully understood.

Contrastive Learning under Long-Tailed Distributions. Contrastive learning (He et al., 2020; Chen et al., 2020) has shown significant progress in representation learning, mostly under balanced data distributions. To address long-tailed distributions, Hybrid-SC (Wang et al., 2021) incorporates the SCL loss into the feature learning branch of a two-branch network, leveraging the quality of its learned embeddings to obtain more discriminative representations. However, subsequent studies have shown that SCL itself is biased toward head classes and yields a skewed representation space (Kang et al., 2021; Cui et al., 2021; Zhu et al., 2022; Li et al., 2022b). To mitigate this issue, PaCo (Cui et al., 2021) introduces parametric class-wise learnable centers, KCL (Kang et al., 2021) uses the same number of positive samples for all classes, and BCL (Zhu et al., 2022) balances the negative terms so that each class contributes equally to the repulsive force. TSC (Li et al., 2022b) assigns uniformly distributed class targets to attract the features of each class. Yi et al. (2025) demonstrates that the geometry of the representation obtained by minimizing the SCL loss is skewed under a specific scenario where all classes share the same number of samples except for one majority class. Different from previous works, we provide a framework for analyzing whether a regular simplex is a global minimizer of supervised contrastive learning objectives under long-tailed distributions, based on the form of their lower bounds. Using our framework, we study two representative objectives: the supervised contrastive loss (Khosla et al., 2020) and the balanced contrastive loss (Zhu et al., 2022), both of which continue to serve as widely used bases for analysis in long-tailed recognition (Cui et al., 2021; 2023; Pan et al., 2024; Li et al., 2024b; Wang et al., 2024; Yi et al., 2025; Son et al., 2025; De Alvis et al., 2025; Wan et al., 2026; Chen et al., 2026). We show that the regular simplex is not a global minimizer of either objective which is nontrivial for BCL, which has been shown to admit such geometry at its minimum. Building on our framework, we propose a balancing method for supervised contrastive objectives and a fully balanced contrastive loss, a simple contrastive objective whose global minimizer is provably a regular simplex under long-tailed distributions.

3 PRELIMINARIES

We consider a K-class classification task and learn a mapping from input data to their associated target labels. Let $\mathcal{X}$ be an input space and $X = (x_1, \dots, x_B) \in \mathcal{X}^B$ be a batch of B training samples. Denote the target space by $[K] = \{1, \dots, K\}$. Let $Y = (y_1, \dots, y_B) \in [K]^B$ be the corresponding targets of the set X. An encoder $f : \mathcal{X} \to \mathcal{Z}$ maps an input to the representation space denoted by $\mathcal{Z} \subset \mathbb{R}^h$, where h is the dimension of the representation. Let $Z = (z_1, \dots, z_B) = (f(x_1), \dots, f(x_B))$ denote the set of representations obtained by applying f to X.

In contrastive learning, each representation lies on a unit hypersphere $\mathbb{S}^{h-1} = \{z \in \mathbb{R}^h : \|z\| = 1\}$, where $\|\cdot\|$ denotes the L2 norm. Following prior analyses of the optimal geometry of supervised contrastive objectives, we consider a mini-batch containing samples from all K classes. This assumption serves as an analytical setting for characterizing the geometry induced at the optimum, rather than requiring every practical mini-batch to contain all classes (Graf et al., 2021; Zhu et al., 2022; Yi et al., 2025). In practical training, where this condition is generally infeasible for large K, we follow prior work and adopt the class-complement setting, which additionally introduces class centers as contrastive pairs (Cui et al., 2021; Zhu et al., 2022; Cui et al., 2023; Son et al., 2025; Yi et al., 2025). Accordingly, our theoretical results characterize the optimum under this full-class analytical setting, while our experiments evaluate structural balancing under practical stochastic mini-batch training. Denote the index set of batch samples by $\mathcal{B} = [B] = \{1, \dots, B\}$, and define a k-class subset as $\mathcal{B}_k = \{i : i \in \mathcal{B}, y_i = k\}$ with cardinality $|\mathcal{B}_k| = B_k$ for each class $k \in [K]$. It follows that $B = \sum_{k\in[K]} B_k$. We assume a batch with $B > 4$, $K \geq 3$ with augmentation, and a long-tailed distribution such that $B_1 \geq \cdots \geq B_K \geq 2$ with $B_1 \gg B_K$. We now define contrastive losses used in our analysis.

Definition 1 (SCL loss (Khosla et al., 2020)). *Let $Z = (z_1, \dots, z_B) \in (\mathbb{S}^{h-1})^B$ be a set of features with corresponding labels $Y = (y_1, \dots, y_B) \in [K]^B$. The SCL loss is defined by*

$$\mathcal{L}_{\mathrm{SCL}}(Z; Y, \mathcal{B}) = \frac{1}{B} \sum_{y\in[K]} \sum_{i\in\mathcal{B}_y} \ell^{(i)}_{\mathrm{SCL}}, \quad \ell^{(i)}_{\mathrm{SCL}} = -\frac{1}{B_y - 1} \sum_{j\in\mathcal{B}_y\setminus\{i\}} \log\left(\frac{\exp(z_i \cdot z_j/\tau)}{\sum_{k\in\mathcal{B}\setminus\{i\}} \exp(z_i \cdot z_k/\tau)}\right). \quad (1)$$

Definition 2 (BCL loss (Zhu et al., 2022)). *Let $Z = (z_1, \dots, z_B) \in (\mathbb{S}^{h-1})^B$ be the features with labels $Y = (y_1, \dots, y_B) \in [K]^B$. The BCL loss is defined by*

$$\mathcal{L}_{\mathrm{BCL}}(Z; Y, \mathcal{B}) = \frac{1}{B} \sum_{y \in [K]} \sum_{i \in \mathcal{B}_y} \ell^{(i)}_{\mathrm{BCL}}, \quad \ell^{(i)}_{\mathrm{BCL}} = -\frac{1}{B_y - 1} \sum_{j \in \mathcal{B}_y \setminus \{i\}} \log \left(\frac{\exp(z_i \cdot z_j / \tau)}{\sum_{k \in [K]} \frac{1}{B_k} \sum_{m \in \mathcal{B}_k} \exp(z_i \cdot z_m / \tau)} \right), \quad (2)$$

where B_k is replaced by $B_k - 1$ when $k = y$.

Following Graf et al. (2021); Zhu et al. (2022); Yi et al. (2025), we set the temperature parameter $\tau = 1$ in the losses for simplicity.

Definition 3 (Regular Simplex). *Let $K, h \in \mathbb{N}$ satisfy $K \leq h + 1$. A K-point configuration $\zeta = (\zeta_1, \dots, \zeta_K) \in (\mathbb{S}^{h-1})^K$ forms a regular simplex if and only if*

(1) $\forall i \in [K], \quad \|\zeta_i\|^2 = 1.$

(2) $\forall i, j \in [K]$ *with* $i \neq j, \quad \zeta_i \cdot \zeta_j = -\frac{1}{K-1}.$

(3) $\sum_{i \in [K]} \zeta_i = 0.$

We regard the regular simplex in Definition 3 as the optimal target geometry on the unit hypersphere: features collapsing to centers minimizes the within-class variance, while conditions (2) and (3) maximize the between-class variance. Thus, the Fisher discriminant ratio (Fisher, 1936) is maximized, leading to an ideal geometry for classification. Also, it has been shown that yielding a regular simplex will improve generalization ability under long-tailed distributions (Liu et al., 2023).

4 Representation Geometry under Long-Tailed Distributions

In this section, we analyze the geometry of the global minimizers of contrastive losses. Section 4.1 develops a framework that derives lower bounds for supervised contrastive objectives and uses them to examine regular simplex optimality. By showing that non-optimality at the lower bound implies non-optimality at the original loss, we prove that a regular simplex is not optimal for either of two representative losses, SCL or BCL. Section 4.2 introduces structural balancing to overcome this failure and proposes Fully Balanced Contrastive Learning (FBCL), whose global minimizer is provably a regular simplex. Section 4.3 analyzes the effect of structural balancing on intra-class variance penalty.

4.1 Geometric Analysis of Supervised Contrastive Objectives

Throughout this section, we refer to the original loss $\mathcal{L}(Z; Y, \mathcal{B})$ defined over the feature configuration Z as the *feature-level objective*. We derive a lower bound of $\mathcal{L}$, denoted by $\mathcal{LB}$, a function over the class centers $\bar{Z}$. We refer to $\mathcal{LB}(\bar{Z}; K)$ as the *class-level lower bound*. We derive the class-level lower bounds for contrastive objectives using Lemma 3 and Lemma 4 in the Appendix. Equality holds when features within each class collapse to their class center. Denote this condition by

(C1) **Class-collapse**: $\forall y \in [K], \forall i \in \mathcal{B}_y : \; z_i = \bar{z}_y$.

Proposition 1 (Lower bound of the SCL loss). *Let $Z = (z_1, \dots, z_B) \in (\mathbb{S}^{h-1})^B$ be a B-point configuration with labels $Y = (y_1, \dots, y_B) \in [K]^B$. It holds that*

$$\mathcal{L}_{\mathrm{SCL}}(Z; Y, \mathcal{B}) \geq \frac{1}{B} \sum_{y \in [K]} B_y \log \Big(B_y - 1 + e^{-1} \sum_{k \in [K] \setminus \{y\}} B_k \exp(\bar{z}_y \cdot \bar{z}_k) \Big) =: \mathcal{LB}_{\mathrm{SCL}}(\bar{Z}; K) \quad (3)$$

where $\bar{z}_y = \frac{1}{B_y} \sum_{i \in \mathcal{B}_y} z_i$, and $\bar{Z} = (\bar{z}_1, \dots, \bar{z}_K)$.

Proposition 2 (Lower bound of the BCL loss). *Let $Z, \bar{Z}$ be defined as in Proposition 1. It holds that*

$$\mathcal{L}_{\mathrm{BCL}}(Z; Y, \mathcal{B}) \geq \frac{1}{B} \sum_{y \in [K]} B_y \log \Big(1 + e^{-1} \sum_{k \in [K] \setminus \{y\}} \exp(\bar{z}_y \cdot \bar{z}_k) \Big) =: \mathcal{LB}_{\mathrm{BCL}}(\bar{Z}; K). \quad (4)$$

Equality in the lower bounds in Eq. (3) and Eq. (4) is attained if and only if (C1) is satisfied.

The proofs of Propositions mainly follow derivations from Graf et al. (2021), and are provided in Appendix C.4.

The class collapse condition alone does not specify the geometry of class centers. We introduce two Lemmas to characterize the geometry of the minimizers of a loss. Rather than analyzing the original loss, we analyze its lower bound. Lemma 1 shows that if a center configuration is not optimal for the lower bound, then the corresponding feature configuration cannot be optimal for the original loss. Lemma 2 then derives a condition that the lower bound must satisfy for a regular simplex to be a minimizer.

Lemma 1. *Let $\mathcal{L}$ be a feature-level contrastive loss, and let $\mathcal{LB}$ be a class-level lower bound of $\mathcal{L}$. Let $\zeta = (\zeta_1, \dots, \zeta_K) \in (\mathbb{S}^{h-1})^K$ be a K-point center configuration, and let $Z^* = (z_1^*, \dots, z_B^*) \in (\mathbb{S}^{h-1})^B$ be a B-point configuration which collapses to ζ, i.e., $z_i^* = \zeta_{y_i}$ for all $i \in [B]$. Then,*

$$\zeta \notin \operatorname*{arg\,min}_{\bar{Z} \in (\mathbb{S}^{h-1})^K} \mathcal{LB}(\bar{Z}; K) \quad \Longrightarrow \quad Z^* \notin \operatorname*{arg\,min}_{Z \in (\mathbb{S}^{h-1})^B} \mathcal{L}(Z; Y, \mathcal{B}). \tag{5}$$

Lemma 2. *Let $\bar{Z}, \bar{Z}^*$ be the center and the regular simplex configuration, respectively. $\mathcal{LB}(\bar{Z}; K)$ can be written with coefficients β, a_y, b_y, c_y for $y \in [K]$,*

$$\mathcal{LB}(\bar{Z}; K) = \frac{1}{\beta} \sum_{y \in [K]} a_y \log \left(b_y + e^{-1} \sum_{m \in [K] \setminus \{y\}} c_m \exp(\bar{z}_y \cdot \bar{z}_m) \right). \tag{6}$$

For $\bar{Z}^*$ to be a minimizer of $\mathcal{LB}(\bar{Z}; K)$, the following condition must be satisfied: For every $y \in [K]$, with $d = -1/(K-1)$,

$$\frac{a_y c_k}{b_y + e^{d-1} \sum_{m \in [K] \setminus \{y\}} c_m} + \frac{a_k c_y}{b_k + e^{d-1} \sum_{m \in [K] \setminus \{k\}} c_m} = \textit{Constant}, \quad \textit{for all } k \in [K] \setminus \{y\}. \tag{7}$$

The proofs are provided in Appendix C.1, C.2 respectively. These results form the basis of our framework: for a supervised contrastive objective, we derive a class-level lower bound $\mathcal{LB}$, examine whether the regular simplex condition is satisfied, and when it is not, transfer this non-optimality to the original loss $\mathcal{L}$. We apply this framework to SCL and BCL.

Theorem 1. *Let $\bar{Z}^* = (\bar{z}_1^*, \dots, \bar{z}_K^*)$ be a regular simplex, and let $Z^* = (z_1^*, \dots, z_B^*)$ collapse class-wise to $\bar{Z}^*$. Then, the regular simplex $\bar{Z}^*$ is not a minimizer of $\mathcal{LB}_{\mathrm{SCL}}(\bar{Z}; K)$ by Lemma 2 and thus, by Lemma 1, Z^* is not a minimizer of $\mathcal{L}_{\mathrm{SCL}}$. That is,*

$$Z^* \notin \operatorname*{arg\,min}_{Z \in (\mathbb{S}^{h-1})^B} \mathcal{L}_{\mathrm{SCL}}(Z; Y, \mathcal{B}). \tag{8}$$

Theorem 1 shows that the regular simplex is not a global minimizer of the SCL loss. The full statement of the theorem and detailed proof are provided in Appendix C.3.1.

Theorem 2. *Let $\bar{Z}^*$ and Z^* be defined as in Theorem 1. Then, $\bar{Z}^*$ is not a minimizer of $\mathcal{LB}_{\mathrm{BCL}}(\bar{Z}; K)$ by Lemma 2 and thus, by Lemma 1, Z^* is not a minimizer of $\mathcal{L}_{\mathrm{BCL}}$. That is,*

$$Z^* \notin \operatorname*{arg\,min}_{Z \in (\mathbb{S}^{h-1})^B} \mathcal{L}_{\mathrm{BCL}}(Z; Y, \mathcal{B}). \tag{9}$$

Theorem 2 shows that the same result holds for BCL: although the BCL loss balances the contributions of negative samples within each anchor-wise loss, $\mathcal{LB}_{\mathrm{BCL}}(\bar{Z}; K)$ still retains B_y, preventing Eq. (7) from being satisfied. The full statement and proof are provided in Appendix C.3.2.

4.2 Fully Balanced Contrastive Learning via Structural Balancing

Theorems 1 and 2 show that the regular simplex is not a global minimizer of the representative losses SCL and BCL. We identify the reason for this failure and introduce a simple way to correct it. Both losses balance the contribution of each class only within the anchor-wise loss $\ell^{(i)}$: SCL loss normalizes over the positive set in $\ell^{(i)}_{\mathrm{SCL}}$ (Eq. (1)), and BCL loss further normalizes contributions of negative samples of each class in the denominator of $\ell^{(i)}_{\mathrm{BCL}}$ (Eq. (2)).

This anchor-wise balancing alone is insufficient: the anchor-wise losses are summed into an overall loss, which can be decomposed into two summations: $\sum_{i \in \mathcal{B}} = \sum_{y \in [K]} \sum_{i \in \mathcal{B}_y}$. This summation weights the contribution of each class by its cardinality B_y. This can be seen by rewriting the BCL loss as the following:

$$\mathcal{L}_{\mathrm{BCL}}(Z; Y, \mathcal{B}) = \frac{1}{B} \sum_{i \in \mathcal{B}} \ell^{(i)}_{\mathrm{BCL}} = \frac{1}{B} \sum_{y \in [K]} \sum_{i \in \mathcal{B}_y} \ell^{(i)}_{\mathrm{BCL}} = \sum_{y \in [K]} \frac{B_y}{B} \left(\frac{1}{B_y} \sum_{i \in \mathcal{B}_y} \ell^{(i)}_{\mathrm{BCL}} \right). \tag{10}$$

Although the anchor-wise loss $\ell^{(i)}_{\mathrm{BCL}}$ is balanced, summing it over all samples assigns each class a weight of B_y/B, as shown in Eq. (10). Consequently, $\mathcal{LB}_{\mathrm{BCL}}(\bar{Z}; K)$ inherits the dependence on B_y. Therefore, to remove this dependence, the summation of the anchors must be balanced.

We propose structural balancing, which extends balancing in contrastive objectives from the anchor-wise loss to the aggregation of these losses into the overall objective. While SCL and BCL balance within each anchor-wise loss, structural balancing balances the contributions of anchors across classes when forming the overall contrastive objective which differs in the motivation and usage from the conventional re-weighting methods. Details are provided in the Appendix D.7. We incorporate structural balancing into the BCL loss, leading to the following FBCL objective.

Definition 4 (FBCL loss). *Let $Z = (z_1, \ldots, z_B) \in (\mathbb{S}^{h-1})^B$ be features with labels $Y = (y_1, \ldots, y_B) \in [K]^B$. The FBCL loss is defined by*

$$\mathcal{L}_{\mathrm{FBCL}}(Z; Y, \mathcal{B}) = \frac{1}{K} \sum_{y \in [K]} \frac{1}{B_y} \sum_{i \in \mathcal{B}_y} \ell^{(i)}_{\mathrm{FBCL}}, \quad \ell^{(i)}_{\mathrm{FBCL}} = -\frac{1}{B_y - 1} \sum_{j \in \mathcal{B}_y \setminus \{i\}} \log \left(\frac{\exp(z_i \cdot z_j/\tau)}{\sum_{k \in [K]} \frac{1}{B_k} \sum_{m \in \mathcal{B}_k} \exp(z_i \cdot z_m/\tau)} \right), \tag{11}$$

where $\ell^{(i)}_{\mathrm{FBCL}} = \ell^{(i)}_{\mathrm{BCL}}$ and B_k is replaced by $B_k - 1$ when $k = y$.

In Eq. (11), the factor $1/K$ corresponds to $1/B$ in the SCL and BCL losses, matching the scale. We derive the class-level lower bound of FBCL to verify whether structural balancing removes the dependence on class cardinality at the lower bound.

Proposition 3 (Lower bound of FBCL). *Let $Z, \bar{Z}$ be defined as in Proposition 1. It holds that*

$$\mathcal{L}_{\mathrm{FBCL}}(Z; Y, \mathcal{B}) \geq \frac{1}{K} \sum_{y \in [K]} \log \left(1 + e^{-1} \sum_{k \in [K] \setminus \{y\}} \exp(\bar{z}_y \cdot \bar{z}_k) \right) =: \mathcal{LB}_{\mathrm{FBCL}}(\bar{Z}; K) \tag{12}$$

The equality in Eq. (12) is attained if and only if (C1) is satisfied.

The proof is provided in Appendix C.4. Proposition 3 shows that $\mathcal{LB}_{\mathrm{FBCL}}(\bar{Z}; K)$ assigns the same weight, $1/K$, to every class and contains no coefficient of class cardinalities, removing the class dependence. As in Section 4.1, we examine the configuration of class centers that minimizes $\mathcal{LB}_{\mathrm{FBCL}}$.

Theorem 3. *Let $\bar{Z}^*$ and Z^* be defined as in Theorem 1. We show that $\bar{Z}^*$ is a minimizer of $\mathcal{LB}_{\mathrm{FBCL}}$ and $\mathcal{L}_{\mathrm{FBCL}}$. That is,*

$$Z^* \in \operatorname*{arg\,min}_{Z \in (\mathbb{S}^{h-1})^B} \mathcal{L}_{\mathrm{FBCL}}(Z; Y, \mathcal{B}). \tag{13}$$

In particular, the minimum of the FBCL loss is given by

$$\mathcal{L}_{\mathrm{FBCL}}(Z; Y, \mathcal{B}) \geq \log \left(1 + (K-1) \exp\left(-\tfrac{K}{K-1} \right) \right), \tag{14}$$

where equality is attained if and only if the representation is in the form of Z^*, a regular simplex.

The full statement of the theorem and its proof are provided in Appendix C.3.3. Theorem 3 characterizes the geometric configuration of Z at the minimum of $\mathcal{L}_{\mathrm{FBCL}}$. It shows that $\mathcal{L}_{\mathrm{FBCL}}$ is minimized if and only if the representations collapse to class centers forming a regular simplex.

4.3 Analysis of intra-class variance penalty

We further analyze the effect of structural balancing on intra-class variance. Following Wang & Isola (2020), we decompose each loss as $\mathcal{L} = \mathcal{L}^{\mathrm{alignment}} + \mathcal{L}^{\mathrm{uniformity}}$, where the alignment term encourages similarity between positive pairs. Full decompositions are provided in Appendix C.5. Since SCL and BCL differ only in their denominator, they share the same alignment term. Let $V_y = \frac{1}{B_y} \sum_{i \in \mathcal{B}_y} \|z_i - \bar{z}_y\|^2$ denote the intra-class variance of class y. Then,

$$\mathcal{L}^{\mathrm{alignment}}_{\mathrm{SCL}} = \mathcal{L}^{\mathrm{alignment}}_{\mathrm{BCL}} = -\frac{1}{B} \sum_{y \in [K]} \sum_{i \in \mathcal{B}_y} \frac{1}{B_y - 1} \sum_{j \in \mathcal{B}_y \setminus \{i\}} (z_i \cdot z_j) = \sum_{y \in [K]} \frac{B_y^2}{B(B_y - 1)} V_y - 1. \tag{15}$$

The coefficient of V_y satisfies $\frac{B_y^2}{B(B_y - 1)} \approx \frac{B_y}{B}$, so the variance penalty is approximately proportional to class cardinality, resulting in stronger contraction for head classes. After structural balancing, the

alignment term becomes

$$\mathcal{L}_{\text{FBCL}}^{\text{alignment}} = -\frac{1}{K}\sum_{y\in[K]}\frac{1}{B_y}\sum_{i\in\mathcal{B}_y}\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}(z_i\cdot z_j) = \sum_{y\in[K]}\frac{B_y}{K(B_y-1)}V_y - 1. \quad (16)$$

Here, $\frac{B_y}{K(B_y-1)} \approx \frac{1}{K}$, which is approximately independent of class cardinality. Thus, within the alignment term, structural balancing makes the penalty on intra-class variance nearly uniform across classes.

5 Experiments

We conduct experiments on standard long-tailed datasets. We also conduct experiments on diverse downstream datasets to evaluate not only classification performance but also the transferability of representations learned under class imbalance, an aspect that has not been widely reported.

5.1 Datasets

Long-tailed recognition We use ImageNet-LT (Liu et al., 2019), a long-tailed subset of ImageNet (Deng et al., 2009) containing 115.8K images from 1,000 classes (cardinalities ranging from 5 to 1,280), and iNaturalist (Van Horn et al., 2018), a large-scale imbalanced dataset of 437.5K images from 8,142 classes. We report accuracy on many-shot (>100 images), medium-shot (20 $\sim$ 100), few-shot (<20), and all classes.

Transfer learning To evaluate the transferability of the learned representations, we run experiments on the following datasets: CIFAR-10/100 (Krizhevsky, 2009), MIT67 (Quattoni & Torralba, 2009), DTD (Cimpoi et al., 2014), Oxford-IIIT Pets (Parkhi et al., 2012), Caltech101 (Fei-Fei et al., 2004), Flowers (Nilsback & Zisserman, 2008), CUB200 (Wah et al., 2011), Food (Bossard et al., 2014), Stanford Cars (Krause et al., 2013) and FGVC Aircraft (Maji et al., 2013). We report results under both linear evaluation and fine-tuning.

5.2 Dual-branch Training Framework

We use the dual-branch framework, which is commonly used to learn better representations under long-tailed distributions (Kang et al., 2020; Wang et al., 2021; Zhu et al., 2022; Suh & Seo, 2023; Du et al., 2024; Yi et al., 2025). The model consists of a base encoder $\phi : \mathcal{X} \to \mathcal{H}$, a prediction head $\psi : \mathcal{H} \to \mathbb{R}^K$, and a contrastive head $g : \mathcal{H} \to \mathcal{Z}$. Given an input (x, y), the model produces $h = \phi(x)$, class logits $p = \psi(h)$, and a normalized contrastive feature $z = g(h)$. The prediction head is optimized using a logit-adjusted cross-entropy loss $\mathcal{L}_{LC}$ (Menon et al., 2021):

$$\mathcal{L}_{LC}(x, y) = -\log\left(\frac{\exp(p_y + \delta_y)}{\sum_{y'\in[K]}\exp(p_{y'} + \delta_{y'})}\right), \quad (17)$$

where π_y is the empirical class prior of class y, and $\delta_y = \log \pi_y$ is the logit adjustment term. The contrastive head is optimized using a contrastive loss, $\mathcal{L}_{CL}$. Overall, the training objective is:

$$\mathcal{L} = \lambda\mathcal{L}_{LC} + \mu\mathcal{L}_{CL} \quad (18)$$

where λ and μ are hyperparameters controlling the contributions of the two objectives.

5.3 Implementation Details

Table 1: Comparison of classification accuracy across contrastive objectives on ImageNet-LT.

Method	ImageNet-LT			
	Many	Medium	Few	All
SimCLR (Chen et al., 2020)	65.0	52.6	36.2	55.2 ± 0.1
SCL (Khosla et al., 2020)	65.8	53.3	36.1	55.8 ± 0.4
BCL (Zhu et al., 2022)	65.8	53.7	36.2	56.0 ± 0.1
FBCL (Ours)	**66.4**	**54.1**	**37.5**	**56.6** ± 0.1

Our implementation follows Zhu et al. (2022). We use ResNet-50 (He et al., 2016) as the model backbone, and use RandAug (Cubuk et al., 2020) for data augmentation. To compare FBCL with different contrastive objectives within the same training framework, we report the results with $\mathcal{L}_{CL}$ in Eq. (18) set to SimCLR, SCL, and BCL losses. For a fair comparison, the dimension of the contrastive head is 1024 for both ImageNet-LT and iNaturalist, and 128 for CIFAR-100-LT, following (Zhu et al., 2022). Detailed hyperparameter settings are provided in Appendix D.2. For transfer learning, we first train the backbone on ImageNet-LT and then perform linear evaluation and full fine-tuning on each dataset. Details on transfer learning are provided in Appendix D.3.

Table 2: Comparison of long-tailed recognition methods and the effect of structural balancing on ImageNet-LT and iNaturalist. * denotes results reproduced with 90 epochs on ImageNet-LT and 100 epochs on iNaturalist for a fair comparison. **Bold** indicates the best results. Standard deviations are reported in Appendix D.5.

Method	ImageNet-LT				iNaturalist			
	Many	Medium	Few	All	Many	Medium	Few	All
OLTR (Liu et al., 2019)	43.2	35.1	18.5	35.6	59.0	64.1	64.9	63.9
τ-norm (Kang et al., 2020)	56.6	44.2	27.4	46.7	65.6	65.3	65.9	65.6
cRT (Kang et al., 2020)	58.8	44.0	26.1	47.3	69.0	71.1	70.2	70.6
PaCo* (Cui et al., 2021)	55.3	50.2	26.7	48.9	60.2	67.7	68.5	67.2
GPaCo* (Cui et al., 2023)	59.4	53.4	30.5	52.6	65.6	72.2	72.0	71.4
LA (Menon et al., 2021)	62.8	49.0	31.8	52.0	68.4	69.4	69.2	69.2
DisAlign (Zhang et al., 2021)	61.3	52.2	31.4	52.9	61.6	70.8	69.9	69.5
Hybrid-PSC (Wang et al., 2021)	-	-	-	-	-	-	-	68.1
MiSLAS (Zhong et al., 2021)	61.7	51.3	35.8	52.7	**73.3**	72.4	70.4	71.6
KCL (Kang et al., 2021)	61.8	49.4	30.9	51.5	-	-	-	68.6
TSC (Li et al., 2022b)	63.5	49.7	30.4	52.4	72.6	70.6	67.8	69.7
SBCL (Hou et al., 2023)	63.8	51.3	31.2	53.4	**73.3**	71.9	68.6	70.8
Focal-SAM (Li et al., 2025)	63.9	52.2	34.4	54.3	68.4	72.0	72.5	71.8
BCL* (Zhu et al., 2022)	65.8	53.7	36.2	56.0	70.0	72.0	71.1	71.3
BCL+SB	66.4	54.1	37.5	56.6	70.4	72.8	71.9	72.2
DBM+BCL* (Son et al., 2025)	64.4	54.3	41.4	56.5	66.3	71.8	73.3	71.8
DBM+BCL+SB	64.9	54.5	**42.3**	56.9	66.5	72.4	**73.8**	72.3
ProCo* (Du et al., 2024)	66.5	54.8	38.1	57.1	70.4	73.1	71.9	72.3
ProCo+SB	**66.8**	**55.2**	38.8	**57.4**	70.7	**73.4**	72.3	**72.7**

Table 3: Comparison of transferability across 11 natural image classification datasets. All backbones are pretrained on ImageNet-LT with 90 epochs of pretraining for a fair comparison. † denotes results reproduced using the training framework in Eq. (18), replacing the contrastive objective. **Bold** and underline indicate the best and second-best results, respectively.

Method	CIFAR10	CIFAR100	MIT67	DTD	Pets	Caltech101	Flowers	CUB200	Food	Cars	Aircraft	Mean
Linear evaluation												
Cross-Entropy	86.8	65.4	60.5	63.5	80.1	83.1	84.6	49.2	60.5	35.2	37.4	64.2
SimCLR (Chen et al., 2020)	83.0	61.2	57.8	64.4	57.2	69.9	78.7	24.6	58.2	28.4	36.5	56.4
PaCo (Cui et al., 2021)	86.2	63.8	60.9	63.3	82.3	83.5	83.5	50.4	60.8	37.1	39.2	64.6
GPaCo (Cui et al., 2023)	**88.0**	67.0	62.8	65.5	84.5	85.8	83.6	50.9	61.2	38.9	40.8	66.3
SimCLR† (Chen et al., 2020)	86.8	67.3	64.6	66.0	84.9	87.1	88.6	**55.6**	63.2	45.4	43.8	68.5
SCL† (Khosla et al., 2020)	87.2	67.2	**66.5**	65.6	84.1	88.1	88.3	54.8	**63.5**	43.4	45.3	68.6
BCL† (Zhu et al., 2022)	86.8	66.5	65.7	66.2	**85.9**	88.1	**88.9**	54.9	62.8	45.1	45.0	68.7
FBCL† (Ours)	87.4	**67.5**	66.1	**66.4**	85.7	**88.3**	88.4	55.5	62.9	**46.0**	**45.5**	**69.1**
Fine-tuning												
Cross-Entropy	91.7	80.2	66.0	59.9	80.2	81.6	83.7	57.7	84.6	80.0	74.2	75.5
SimCLR (Chen et al., 2020)	96.1	78.4	67.0	65.2	81.3	76.6	90.3	67.2	85.7	88.3	79.9	79.6
PaCo (Cui et al., 2021)	93.4	78.8	71.6	67.6	87.1	88.0	93.9	73.2	**86.1**	**90.7**	80.1	82.7
GPaCo (Cui et al., 2023)	**96.8**	**82.6**	**73.4**	69.9	86.9	90.0	95.1	75.0	85.9	90.2	81.9	84.3
SimCLR† (Chen et al., 2020)	96.3	81.1	69.3	70.3	87.3	88.4	95.1	73.1	85.5	90.1	81.5	83.5
SCL† (Khosla et al., 2020)	96.0	82.0	71.0	69.4	87.0	88.7	94.7	73.1	85.1	90.2	80.8	83.5
BCL† (Zhu et al., 2022)	96.0	82.3	69.0	70.3	88.4	89.6	94.9	**75.3**	85.2	89.5	81.6	83.8
FBCL† (Ours)	96.3	**82.6**	72.4	**71.0**	**88.5**	**90.5**	**95.2**	74.6	85.6	90.2	**82.8**	**84.5**

5.4 Results

Table 2 reports the top-1 accuracy of long-tailed recognition methods and methods combined with structural balancing on ImageNet-LT and iNaturalist. Structural balancing consistently improves over the base methods, showing its applicability and that it can be readily incorporated into other methods. Table 3 reports the transferability results, with FBCL achieving the highest average performance on both evaluations. Tables 1 and 2 show that FBCL achieves the highest results among contrastive methods based on SCL or BCL (Wang et al., 2021; Cui et al., 2021; 2023; Kang et al., 2021; Li et al., 2022b; Zhu et al., 2022; Hou et al., 2023). Results with other methods, datasets and standard deviations are provided in Appendix D.5.

5.5 Ablation Study

To examine the individual role of each balancing method in the FBCL loss, we conduct an ablation study on ImageNet-LT. As described in Section 4.2, FBCL contains three balancing components:

Table 4: Comparison of intra-class variance and Top-1 accuracy. (a) Results on synthetic data. (b) Ablation study of positive balancing (PB), negative balancing (NB), and structural balancing (SB) on ImageNet-LT. Details are provided in Appendix D.5.

(a) Synthetic data

Method	Intra-class variance			Acc.
	Many	Medium	Few	
CE	0.459	0.533	0.721	73.8 ± 0.16
SCL	0.272	0.346	0.575	74.1 ± 2.74
BCL	0.250	0.385	0.749	74.1 ± 2.36
FBCL (Ours)	0.409	0.453	0.438	75.5 ± 3.10

(b) ImageNet-LT

Method	PB	NB	SB	Intra-class variance			Acc.
				Many	Medium	Few	
PB (SCL)	✓	✗	✗	0.189	0.298	0.357	55.78 ± 0.41
PB+NB (BCL)	✓	✓	✗	0.167	0.293	0.362	56.03 ± 0.07
PB+SB	✓	✗	✓	0.368	0.502	0.509	55.92 ± 0.19
NB+SB	✗	✓	✓	0.597	0.623	0.602	55.35 ± 0.17
PB+NB+SB (Ours)	✓	✓	✓	0.196	0.281	0.339	56.64 ± 0.12

positive balancing (PB), negative balancing (NB), and structural balancing (SB). Please refer to Appendix C.6 for each balancing component. From Table 4b, we first observe that SB reduces the disparity of intra-class variance across classes. Without SB, the intra-class variance gap between many and few classes is 0.168 for SCL and 0.195 for BCL, while adding SB reduces these gaps to 0.141 for PB+SB and 0.143 for FBCL. The results of NB+SB further show that reducing the variance gap alone is not sufficient, and that lowering the intra-class variance within each class group is also important. Comparing NB+SB with FBCL isolates the role of PB: adding PB reduces intra-class variance across all groups while improving overall accuracy. Likewise, comparing PB+SB with FBCL shows that removing NB decreases accuracy, indicating the contribution of NB. Together, these results show that PB, NB, and SB are complementary.

6 Discussion

Our theoretical analysis follows the standard setting adopted in prior geometric analyses of contrastive learning, where each batch contains all classes (Graf et al., 2021; Zhu et al., 2022; Yi et al., 2025). This setting is used to characterize the geometry of the objective at its optimum, and such analyses still provide a principled indication of the geometric tendency induced by the objective under practical mini-batch training. Within this analytical setting, our contribution is to show that the regular simplex guarantee does not hold for previous methods under long-tailed distributions, and to establish such a guarantee for FBCL. The dimensional condition is satisfied on ImageNet-LT but not on iNaturalist. The attainability of such geometric optima in large-scale training is beyond the scope of this work and remains underexplored even for contrastive objectives under large-scale balanced distributions (Graf et al., 2021); how these geometric tendencies manifest under partial class coverage in mini-batches or when $K > h+1$ provides an interesting direction for future study. Nevertheless, structural balancing itself is not restricted to these analytical conditions and consistently improves the methods with which it is combined in practical large-scale settings, including iNaturalist. Thus, our theory characterizes how balancing changes the objective and its optimum under the same analytical regime used in prior works, while the experiments show that the resulting mechanism remains effective under practical stochastic training.

7 Conclusion

In this paper, we studied the representation geometry induced by supervised contrastive objectives under long-tailed distributions and developed a framework for analyzing regular simplex optimality. We showed that the regular simplex is not a global minimizer of SCL or BCL under class imbalance due to the remaining dependence on class cardinality in the aggregation of anchor-wise losses. To remove this dependence, we introduced structural balancing and proposed Fully Balanced Contrastive Learning (FBCL), which admits a regular simplex as a global optimum. We further showed that structural balancing makes the alignment term penalty on intra-class variance approximately independent of class frequency, and empirically observed more balanced intra-class variances across classes. Experiments on ImageNet-LT, iNaturalist, and transfer tasks demonstrate improvements in long-tailed recognition and representation transfer, as well as the effectiveness of structural balancing when combined with other methods. Our work thus provides a principled way to analyze and improve supervised contrastive learning under long-tailed distributions.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs, assisting with translation, implementing methods, providing feedback on research experiments, and interpreting results. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, cleaning and reformatting datasets. Supporting qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for editing software codes, summarizing existing literature, editing a research paper to improve readability, and identifying relevant literature. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the assumptions and complete proofs of our theoretical results in the appendix. For the experimental results, implementation details, training configurations, hyperparameters, and dataset settings are provided in the appendix. We also provide the source code required to reproduce our experiments as supplementary material.

REFERENCES

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014. 7

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 2018. 2

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019. 1, 2

Shenghan Chen, Yiming Liu, Yanzhen Wang, Yujia Wang, and Xiankai Lu. Reframing long-tailed learning via loss landscape geometry. In *CVPR*, 2026. 3

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3, 7, 8, 29, 30

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 7

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, 2020. 7

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *ICCV*, 2021. 1, 3, 8

Jiequan Cui, Zhisheng Zhong, Zhuotao Tian, Shu Liu, Bei Yu, and Jiaya Jia. Generalized parametric contrastive learning. *TPAMI*, 2023. 1, 3, 8

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 2, 21, 32

Charika De Alvis, Dishanika Denipitiyage, and Suranga Seneviratne. Long-tail learning with rebalanced contrastive loss. *Neurocomputing*, 2025. 1, 3, 20, 32

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 7

Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. LPT: long-tailed prompt tuning for image classification. In *ICLR*, 2023. 2

Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. Probabilistic contrastive learning for long-tailed visual recognition. *TPAMI*, 2024. 1, 7, 8, 31

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVPR Workshop*, 2004. 7

Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936. 4

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In *ICML*, 2021. 1, 3, 4, 9, 14

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 2009. 2

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

Chengkai Hou, Jieyu Zhang, Haonan Wang, and Tianyi Zhou. Subclass-balancing contrastive learning for long-tailed recognition. In *ICCV*, 2023. 1, 8

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, 2016. 2

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020. 2, 7, 8, 32

Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *ICLR*, 2021. 1, 3, 8

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020. 1, 3, 7, 8, 30

Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. In *NeurIPS*, 2021. 2

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 *ICLR*, 2024. 19

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshop*, 2013. 7

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report*, 2009. 1, 7

Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data. In *ICLR*, 2023. 28

Mengke Li, Yiu-ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *CVPR*, 2022a. 2

Mengke Li, HU Zhikai, Yang Lu, Weichao Lan, Yiu-ming Cheung, and Hui Huang. Feature fusion from head to tail for long-tailed visual recognition. In *AAAI*, 2024a. 2

Sicong Li, Qianqian Xu, Zhiyong Yang, Zitai Wang, Linchao Zhang, Xiaochun Cao, and Qingming Huang. Focal-sam: Focal sharpness-aware minimization for long-tailed classification. In *ICML*, 2025. 8

Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *CVPR*, 2022b. 1, 3, 8

Weiqi Li, Fan Lyu, Fanhua Shang, Liang Wan, and Wei Feng. Long-tailed learning as multi-objective optimization. In *AAAI*, 2024b. 3

Xuantong Liu, Jianfeng Zhang, Tianyang Hu, He Cao, Yuan Yao, and Lujia Pan. Inducing neural collapse in deep long-tailed learning. In *ICML*, 2023. 4

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019. 7, 8

Subhransu Maji, Juho Kannala, Esa Rahtu, and Matthew B Blaschko. Fine-grained visual classification of aircraft. Technical report, 2013. 7

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021. 2, 7, 8

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 7

Jeongheon Oh and Kibok Lee. On the effectiveness of supervision in asymmetric non-contrastive learning. In *ICML*, 2024. 29

Haolin Pan, Yong Guo, Mianjie Yu, and Jian Chen. Enhanced long-tailed recognition with contrastive cutmix augmentation. *TIP*, 2024. 3

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 7

Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *CVPR*, 2009. 7

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *NeurIPS*, 2020. 1

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, 2018. 2

Sumin Roh, Harim Kim, Ho Yun Lee, and Il Yong Chun. Equilibrium contrastive learning for imbalanced image classification. *TETCI*, 2026. 21

Jiang-Xin Shi, Tong Wei, Yuke Xiang, and Yu-Feng Li. How re-sampling helps for long-tail learning? In *NeurIPS*, 2023. 2

Jiang-Xin Shi, Tong Wei, Zhi Zhou, Jie-Jing Shao, Xin-Yan Han, and Yufeng Li. Long-tail learning with foundation model: Heavy fine-tuning hurts. In *ICML*, 2024. 2

Minseok Son, Inyong Koo, Jinyoung Park, and Changick Kim. Difficulty-aware balancing margin loss for long-tailed recognition. In *AAAI*, 2025. 3, 8, 31

Min-Kook Suh and Seung-Woo Seo. Long-tailed recognition by mutual information maximization between latent features and ground-truth labels. In *ICML*, 2023. 7

Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *ECCV*, 2022. 2

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 1, 7

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011. 7

Wenhai Wan, Teng Zhang, Shao-Yuan Li, Xinrui Wang, Qiang-Sheng Hua, and Songcan Chen. Shortcut-resistant cam distillation for long-tailed recognition. In *ICML*, 2026. 3

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *CVPR*, 2021. 28

Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *CVPR*, 2021. 1, 2, 3, 7, 8

Pengkun Wang, Zhe Zhao, HaiBin Wen, Fanfu Wang, Binwu Wang, Qingfu Zhang, and Yang Wang. Llm-autoda: Large language model-driven automatic data augmentation for long-tailed problems. *NeurIPS*, 2024. 3, 31

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020. 6

Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. A unified generalization analysis of re-weighting and logit-adjustment for imbalanced learning. In *NeurIPS*, 2023. 2

Lingjie Yi, Jiachen Yao, Weimin Lyu, Haibin Ling, Raphael Douady, and Chao Chen. Geometry of long-tailed representation learning: Rebalancing features for skewed distributions. In *ICLR*, 2025. 1, 3, 4, 7, 9, 14, 19, 20

Lingjie Yi, Raphael Douady, and Chao Chen. The convergent representation of contrastive vision-language models: Geometry, modality gap and shared space alignment. In *ICML*, 2026. 1

Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *CVPR*, 2019. 2

Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, 2021. 8

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *CVPR*, 2021. 2, 8

Zhisheng Zhong, Jiequan Cui, Zeming Li, Eric Lo, Jian Sun, and Jiaya Jia. Rebalanced siamese contrastive mining for long-tailed recognition. *arXiv preprint arXiv:2203.11506*, 2022. 20, 32

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, 2020. 2, 32

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *CVPR*, 2022. 1, 3, 4, 7, 8, 9, 14, 28, 29, 30

A Appendix

B Preliminaries

We summarize and add details of the preliminaries introduced in Section 3.

- We adopt the setting that a mini-batch contains all classes, as is commonly assumed when analyzing the optimal geometry (Graf et al., 2021; Zhu et al., 2022; Yi et al., 2025).
- Let $\mathcal{B}$ be an index set of an arbitrary batch.
- K: number of classes, assume $K \geq 3$.
- $\mathcal{B} = \dot{\bigcup}_{y\in[K]} \mathcal{B}_y$. And thus, $|\mathcal{B}| = \sum_{y\in[K]} |\mathcal{B}_y|$.
- Let $|\mathcal{B}| = B > 4$, $|\mathcal{B}_y| = B_y$ for any $y \in [K]$. As we augment samples for contrastive pairs, we assume $|B_y| > 2$.
- Long-tailed distribution: $B_1 \geq B_2 \geq \cdots \geq B_K$ with $B_1 \gg B_K$.
- Let $\mathbb{S}^{h-1} = \{z \in \mathbb{R}^h : \|z\|^2 = 1\}$ denote the unit sphere.
- Following Graf et al. (2021),Zhu et al. (2022), and Yi et al. (2025), the temperature parameter is fixed to $\tau = 1$.

C Theoretical framework and proofs

We establish our framework in three steps that can be used for any contrastive objective. Then, we apply our framework to representative objectives under a long-tailed distribution.

For a given supervised contrastive objective $\mathcal{L}$, Lemmas 3 and 4 derive $\mathcal{LB}$ of $\mathcal{L}$, a lower bound whose equality condition corresponds to class-collapse (C1).

Results of Lemmas 1 and 2 correspond to consistency and characterization, respectively. Concretely, we analyze the loss through its lower bound to obtain data-agnostic results since it depends only on the number of classes. To ensure that the results derived from the lower bound are consistent with those from the original loss, that is, that the geometry of the representations is consistent under both analyses, we establish Lemma 1. Therefore, if a particular geometry configuration does not minimize the lower bound, it also does not minimize the original loss. If it does, we further derive a lower bound of $\mathcal{LB}$ whose equality condition characterizes the optimal geometry. Lemma 2 allows us to analyze a lower bound through its coefficients and characterize whether a regular simplex configuration minimizes it.

With these Lemmas, our framework is as follows:

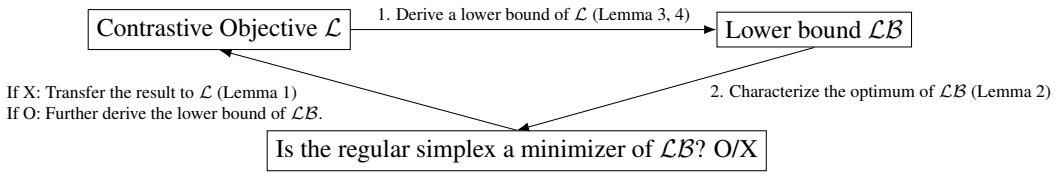


Figure 2: Schematic overview of our framework.

Leveraging this framework, we investigate the optimality of SCL, BCL, and FBCL in Theorems 1, 2, and 3.

C.1 Reduction from feature-level to class-level

The following lemma establishes the consistency between the two minimization problems when a geometry fails to minimize the lower bound: a class-collapsed feature configuration fails to be a global minimizer of $\mathcal{L}$ if its center configuration fails to be a minimizer of $\mathcal{LB}$. This consistency allows us to conduct the analysis at the lower bound and determine when a regular simplex fails to be a global minimizer.

Lemma 1. *Let $\mathcal{L}$ be a feature-level contrastive loss, and let $\mathcal{LB}$ be a class-level lower bound of $\mathcal{L}$. Let $\zeta = (\zeta_1, \ldots, \zeta_K) \in (\mathbb{S}^{h-1})^K$ be a K-point center configuration, and let $Z^* = (z_1^*, \ldots, z_B^*) \in$*

$(\mathbb{S}^{h-1})^B$ *be a* B*-point configuration which collapses to* ζ*, i.e.,* $z_i^* = \zeta_{y_i}$ *for all* $i \in [B]$*. Then,*

$$\zeta \notin \operatorname*{arg\,min}_{\bar{Z}\in(\mathbb{S}^{h-1})^K} \mathcal{LB}(\bar{Z}; K) \quad \Longrightarrow \quad Z^* \notin \operatorname*{arg\,min}_{Z\in(\mathbb{S}^{h-1})^B} \mathcal{L}(Z; Y, \mathcal{B}). \tag{19}$$

Proof. For brevity, we omit the dependence on Y and $\mathcal{B}$ and write $\mathcal{L}(Z)$ in place of $\mathcal{L}(Z; Y, \mathcal{B})$. Suppose that

$$\zeta \notin \operatorname*{arg\,min}_{\bar{Z}\in(\mathbb{S}^{h-1})^K} \mathcal{LB}(\bar{Z}). \tag{20}$$

Then, there exists $\tilde{\zeta} = (\tilde{\zeta}_1, \ldots, \tilde{\zeta}_K) \in (\mathbb{S}^{h-1})^K$ such that

$$\mathcal{LB}(\tilde{\zeta}) < \mathcal{LB}(\zeta). \tag{21}$$

Define the class-collapsed feature configuration $\tilde{Z} = (\tilde{z}_1, \ldots, \tilde{z}_B) \in (\mathbb{S}^{h-1})^B$ by

$$\tilde{z}_i = \tilde{\zeta}_{y_i} \quad \text{for each } i \in [B]. \tag{22}$$

Since $\tilde{Z}$ collapses to $\tilde{\zeta}$ and Z^* collapses to ζ, the equality of the lower bound under class collapse gives

$$\mathcal{L}(\tilde{Z}) = \mathcal{LB}(\tilde{\zeta}), \qquad \mathcal{L}(Z^*) = \mathcal{LB}(\zeta). \tag{23}$$

Therefore,

$$\mathcal{L}(\tilde{Z}) = \mathcal{LB}(\tilde{\zeta}) < \mathcal{LB}(\zeta) = \mathcal{L}(Z^*). \tag{24}$$

Hence, Z^* cannot be a global minimizer of $\mathcal{L}$, and thus

$$Z^* \notin \operatorname*{arg\,min}_{Z\in(\mathbb{S}^{h-1})^B} \mathcal{L}(Z; Y, \mathcal{B}). \tag{25}$$

□

Moreover, **we give another version** of Lemma 1, which we do not use in this work, **but which may be useful for future studies**. This Lemma gives an if and only if condition of the optimality between the class-level lower bound and the original loss. The only added assumption is the dimension condition, $K < h$.

Lemma 1. *Let* $\mathcal{L}$ *be a feature-level contrastive loss, and let* $\mathcal{LB}$ *be a class-level lower bound of* $\mathcal{L}$*. Let* $\zeta = (\zeta_1, \ldots, \zeta_K) \in (\mathbb{S}^{h-1})^K$ *be a* K*-point center configuration, and let* $Z^* = (z_1^*, \ldots, z_B^*) \in (\mathbb{S}^{h-1})^B$ *be a* B*-point configuration which collapses to* ζ*, i.e.,* $z_i^* = \zeta_{y_i}$ *for all* $i \in [B]$*. Then, with* $K < h$*,*

$$\zeta \in \operatorname*{arg\,min}_{\bar{Z}\in(\mathbb{S}^{h-1})^K} \mathcal{LB}(\bar{Z}; K) \iff Z^* \in \operatorname*{arg\,min}_{Z\in(\mathbb{S}^{h-1})^B} \mathcal{L}(Z; Y, \mathcal{B}). \tag{26}$$

Proof. For brevity, we omit the dependence on Y and $\mathcal{B}$ and write $\mathcal{L}(Z)$ in place of $\mathcal{L}(Z; Y, \mathcal{B})$. We use the following three properties of the class-level lower bound.

First, for every feature configuration $Z \in (\mathbb{S}^{h-1})^B$, the induced class-level configuration $C(Z) = \left(\frac{1}{B_1}\sum_{i\in\mathcal{B}_1} z_i, \ldots, \frac{1}{B_K}\sum_{i\in\mathcal{B}_K} z_i\right) \in (\mathbb{B}^h)^K$ (where $\mathbb{B}$ is the unit ball) satisfies

$$\mathcal{L}(Z) \geq \mathcal{LB}(C(Z)). \tag{27}$$

Second, if Z collapses to a class-level configuration $\bar{Z} \in (\mathbb{S}^{h-1})^K$ (i.e., $z_i = \bar{z}_{y_i}$ for all i), then equality holds in the lower bound, that is,

$$\mathcal{L}(Z) = \mathcal{LB}(\bar{Z}). \tag{28}$$

Third, the infimum of $\mathcal{LB}$ over the unit ball is attained on the sphere:

$$\inf_{\bar{Z}\in(\mathbb{B}^h)^K} \mathcal{LB}(\bar{Z}) \;=\; \inf_{\bar{Z}\in(\mathbb{S}^{h-1})^K} \mathcal{LB}(\bar{Z}). \tag{29}$$

We justify Eq. (29) as follows. Observe that each class-level lower bound $\mathcal{LB}$ considered in this paper (Eq. (44)) depends on $\bar{Z}$ only through the off-diagonal inner products $\{\bar{z}_y \cdot \bar{z}_k\}_{y\neq k}$; the diagonal entries $\|\bar{z}_y\|^2$ do not appear in $\mathcal{LB}$.

Fix any $\bar{Z} \in (\mathbb{B}^h)^K$ and let $G \in \mathbb{R}^{K\times K}$ denote its Gram matrix, $G_{yk} = \bar{z}_y \cdot \bar{z}_k$. Then G is positive semi-definite (PSD) and $G_{yy} = \|\bar{z}_y\|^2 \leq 1$ for every y. Define

$$\tilde{G} \;\coloneqq\; G + \operatorname{diag}\big(1 - G_{11},\, \ldots,\, 1 - G_{KK}\big). \tag{30}$$

Since $1 - G_{yy} \geq 0$, the added diagonal matrix is PSD, so $\tilde{G}$ is a sum of two PSD matrices and hence PSD. By construction,

$$\tilde{G}_{yy} = 1 \quad \text{for all } y, \qquad \tilde{G}_{yk} = G_{yk} \quad \text{for all } y \neq k. \tag{31}$$

Because $\tilde{G}$ is a $K \times K$ PSD matrix, $\operatorname{rank}(\tilde{G}) \leq K$. Since $K < h$, the Cholesky decomposition of $\tilde{G}$ yields vectors $\tilde{z}_1, \ldots, \tilde{z}_K \in \mathbb{R}^h$ with $\tilde{z}_y \cdot \tilde{z}_k = \tilde{G}_{yk}$. The diagonal condition $\tilde{G}_{yy} = 1$ gives $\|\tilde{z}_y\| = 1$, so $\tilde{Z} \coloneqq (\tilde{z}_1, \ldots, \tilde{z}_K) \in (\mathbb{S}^{h-1})^K$. Moreover, the off-diagonal inner products of $\tilde{Z}$ coincide with those of $\bar{Z}$, and since $\mathcal{LB}$ depends only on these,

$$\mathcal{LB}(\tilde{Z}) \;=\; \mathcal{LB}(\bar{Z}). \tag{32}$$

This shows that for every $\bar{Z} \in (\mathbb{B}^h)^K$, there exists $\tilde{Z} \in (\mathbb{S}^{h-1})^K$ with the same value of $\mathcal{LB}$. Consequently,

$$\inf_{\bar{Z}\in(\mathbb{S}^{h-1})^K} \mathcal{LB}(\bar{Z}) \;\leq\; \inf_{\bar{Z}\in(\mathbb{B}^h)^K} \mathcal{LB}(\bar{Z}). \tag{33}$$

The reverse inequality is immediate from $\mathbb{S}^{h-1} \subset \mathbb{B}^h$, which proves Eq. (29).

Now, we use the properties above to prove the Lemma. We first prove the ($\Longleftarrow$) direction. Assume, for contradiction, that $\zeta \notin \arg\min_{\bar{Z}\in(\mathbb{S}^{h-1})^K} \mathcal{LB}(\bar{Z})$. Then there exists $\tilde{\zeta} = (\tilde{\zeta}_1, \ldots, \tilde{\zeta}_K) \in (\mathbb{S}^{h-1})^K$ such that

$$\mathcal{LB}(\tilde{\zeta}) < \mathcal{LB}(\zeta). \tag{34}$$

Define the class-collapsed feature configuration $\tilde{Z} = (\tilde{z}_1, \ldots, \tilde{z}_B) \in (\mathbb{S}^{h-1})^B$ by

$$\tilde{z}_i = \tilde{\zeta}_{y_i} \quad \text{for each } i \in [B]. \tag{35}$$

Since $\tilde{Z}$ collapses to $\tilde{\zeta}$ and Z^* collapses to ζ, the equality property gives

$$\mathcal{L}(\tilde{Z}) = \mathcal{LB}(\tilde{\zeta}), \qquad \mathcal{L}(Z^*) = \mathcal{LB}(\zeta). \tag{36}$$

Therefore,

$$\mathcal{L}(\tilde{Z}) = \mathcal{LB}(\tilde{\zeta}) < \mathcal{LB}(\zeta) = \mathcal{L}(Z^*), \tag{37}$$

which contradicts the assumption that Z^* is a global minimizer of $\mathcal{L}$. Hence,

$$\zeta \in \operatorname*{arg\,min}_{\bar{Z}\in(\mathbb{S}^{h-1})^K} \mathcal{LB}(\bar{Z}). \tag{38}$$

It remains to prove the ($\Longrightarrow$) direction. Suppose that $\zeta \in \arg\min_{\bar{Z}\in(\mathbb{S}^{h-1})^K} \mathcal{LB}(\bar{Z})$, and let Z^* be the class-collapsed feature configuration satisfying $z_i^* = \zeta_{y_i}$ for each $i \in [B]$. By the sphere reduction property (29), ζ is also a global minimizer of $\mathcal{LB}$ over the unit ball:

$$\zeta \in \operatorname*{arg\,min}_{\bar{Z}\in(\mathbb{B}^h)^K} \mathcal{LB}(\bar{Z}). \tag{39}$$

Let $Z \in (\mathbb{S}^{h-1})^B$ be an arbitrary feature configuration. By the lower bound property,

$$\mathcal{L}(Z) \geq \mathcal{LB}(C(Z)). \tag{40}$$

Since $C(Z) \in (\mathbb{B}^h)^K$, (39) yields

$$\mathcal{LB}(C(Z)) \geq \mathcal{LB}(\zeta). \tag{41}$$

Moreover, because Z^* collapses to ζ, the equality property gives

$$\mathcal{L}(Z^*) = \mathcal{LB}(\zeta). \tag{42}$$

Combining these relations,

$$\mathcal{L}(Z) \;\geq\; \mathcal{LB}(C(Z)) \;\geq\; \mathcal{LB}(\zeta) \;=\; \mathcal{L}(Z^*). \tag{43}$$

Since this holds for every $Z \in (\mathbb{S}^{h-1})^B$, Z^* is a global minimizer of $\mathcal{L}$. □

C.2 A CONDITION OF THE LOWER BOUND FOR A REGULAR SIMPLEX OPTIMALITY

Here, we give a detailed version of Lemma 2. The class-level lower bounds share a common functional form,

$$\mathcal{LB}(\bar{Z}; K) = \frac{1}{\beta} \sum_{y \in [K]} a_y \log\left(\Gamma_y(\bar{Z})\right), \tag{44}$$

$$\text{where} \quad \Gamma_y(\bar{Z}) = b_y + e^{-1} \sum_{k \in [K]\setminus\{y\}} c_k \exp(\bar{z}_y \cdot \bar{z}_k), \tag{45}$$

with $\beta = \sum_{y \in [K]} a_y$. The coefficients $\{a_y\}, \{b_y\}, \{c_y\}$ arise from applying Jensen's inequality during the derivation of the lower bound, and β is a normalization factor. The following Lemma states a single condition under which a regular simplex satisfies the first-order Lagrangian condition for any objective of the form in Eq. (44).

Lemma 2. *(Necessary condition for regular simplex optimality) Let $\bar{Z}^*$ be a K-point configuration forming a regular simplex on $\mathbb{S}^{h-1}$. $\bar{Z}^*$ satisfies the first-order condition*

$$\nabla_{\bar{z}_i} \mathcal{LB}(\bar{Z}^*) = \lambda_i \nabla_{\bar{z}_i} \left(\|\bar{z}_i\|^2 - 1\right) \quad \textit{for all } i \in [K], \tag{46}$$

if and only if, for every $i \in [K]$,

$$\frac{a_i c_k}{\Gamma_i} + \frac{a_k c_i}{\Gamma_k} = \textit{Constant}, \quad \textit{for all } k \in [K]\setminus\{i\}, \tag{47}$$

where $\Gamma_y = b_y + e^{d-1} \sum_{k \in [K]\setminus\{y\}} c_k$ and $d = \bar{z}_i^* \cdot \bar{z}_j^*$ for all $i \neq j$.

When such λ_i exists, it is given by

$$\lambda_i = -\frac{e^{d-1}}{2\beta} \left(\frac{a_i c_k}{\Gamma_i} + \frac{a_k c_i}{\Gamma_k} \right). \tag{48}$$

Proof. Denote $\Gamma_y(\bar{Z}) = \Gamma_y$ for simplicity. Consider the constrained optimization problem of minimizing $\mathcal{LB}(\bar{Z}; K)$ subject to $G_i(\bar{Z}) = \|\bar{z}_i\|^2 - 1 = 0$ for all $i \in [K]$. Then $\bar{Z}^*$ satisfies the first-order condition of the associated Lagrangian if and only if for each $i \in [K]$ there exists $\lambda_i \in \mathbb{R}$ such that

$$\nabla_{\bar{z}_i} \mathcal{LB}(\bar{Z}^*) = \lambda_i \nabla_{\bar{z}_i} G_i(\bar{Z}^*). \tag{49}$$

Fix an arbitrary class $i \in [K]$. We first derive the left-hand side of Eq. (49). The gradient of Γ_y is

$$\nabla_{\bar{z}_i} \Gamma_y = \begin{cases} e^{-1} \displaystyle\sum_{k \in [K]\setminus\{y\}} c_k \exp(\bar{z}_y \cdot \bar{z}_k) \bar{z}_k, & \text{if } y = i, \quad (50a) \\ e^{-1} c_i \exp(\bar{z}_y \cdot \bar{z}_i) \bar{z}_y, & \text{if } y \in [K]\setminus\{i\}. \quad (50b) \end{cases}$$

Hence,

$$\nabla_{\bar{z}_i} \mathcal{LB}(\bar{Z}; K) = \nabla_{\bar{z}_i} \left(\frac{1}{\beta} \sum_{y \in [K]} a_y \log(\Gamma_y) \right) \tag{51}$$

$$= \frac{1}{\beta} \left(a_i \nabla_{\bar{z}_i} \log(\Gamma_i) + \sum_{y \in [K]\setminus\{i\}} a_y \nabla_{\bar{z}_i} \log(\Gamma_y) \right) \tag{52}$$

$$\overset{\text{Eq. (50a), (50b)}}{=} \frac{1}{\beta} \left(a_i \frac{e^{-1} \sum_{k \in [K]\setminus\{i\}} c_k \exp(\bar{z}_i \cdot \bar{z}_k) \bar{z}_k}{\Gamma_i} + \sum_{y \in [K]\setminus\{i\}} a_y \frac{e^{-1} c_i \exp(\bar{z}_y \cdot \bar{z}_i) \bar{z}_y}{\Gamma_y} \right) \tag{53}$$

$$= \frac{e^{-1}}{\beta} \sum_{k \in [K]\setminus\{i\}} \left(\frac{a_i c_k}{\Gamma_i} + \frac{a_k c_i}{\Gamma_k} \right) \exp(\bar{z}_i \cdot \bar{z}_k) \bar{z}_k. \tag{54}$$

Next, we derive the right-hand side:

$$\lambda_i \nabla_{\bar{z}_i} G_i(\bar{Z}) = \lambda_i \nabla_{\bar{z}_i} \left(\|\bar{z}_i\|^2 - 1\right) = 2\lambda_i \bar{z}_i. \tag{55}$$

Therefore, the Lagrangian equation (49) becomes

$$\frac{e^{-1}}{\beta}\sum_{k\in[K]\setminus\{i\}}\left(\frac{a_i c_k}{\Gamma_i}+\frac{a_k c_i}{\Gamma_k}\right)\exp(\bar{z}_i\cdot\bar{z}_k)\bar{z}_k = 2\lambda_i\bar{z}_i. \tag{56}$$

We now evaluate this condition at $\bar{Z}=\bar{Z}^*$ and denote $\Gamma_y(\bar{Z}^*)=\Gamma_y$. Since $\bar{Z}^*$ is a regular simplex, there exists $d\in\mathbb{R}$ such that $\bar{z}_i^*\cdot\bar{z}_j^*=d$ for all $i\neq j$. Thus,

$$\frac{e^{d-1}}{\beta}\sum_{k\in[K]\setminus\{i\}}\left(\frac{a_i c_k}{\Gamma_i}+\frac{a_k c_i}{\Gamma_k}\right)\bar{z}_k^* = 2\lambda_i\bar{z}_i^*, \tag{57}$$

$$\text{where}\quad \Gamma_y = b_y + e^{d-1}\sum_{k\in[K]\setminus\{y\}} c_k. \tag{58}$$

By the regular simplex property $\bar{z}_i^*=-\sum_{k\in[K]\setminus\{i\}}\bar{z}_k^*$, we can merge the right-hand side into the left-hand side and obtain

$$\sum_{k\in[K]\setminus\{i\}}\left(\frac{a_i c_k}{\Gamma_i}+\frac{a_k c_i}{\Gamma_k}+\frac{2\beta}{e^{d-1}}\lambda_i\right)\bar{z}_k^* = 0. \tag{59}$$

Since any subset of $K-1$ vertices from a regular simplex is linearly independent, the coefficients in Eq. (59) must all vanish. Such a λ_i exists if and only if

$$\frac{a_i c_k}{\Gamma_i}+\frac{a_k c_i}{\Gamma_k} = \text{Constant},\quad \text{for all } k\in[K]\setminus\{i\}, \tag{60}$$

which proves the result. The expression for λ_i in Eq. (48) follows immediately. □

C.3 Applying our framework to supervised contrastive objectives

We now apply the framework to supervised contrastive objectives in the theorems. For each loss, the procedure is the same: 1. Derive the lower bound using Lemmas 3 and 4. 2. From the lower bound, we identify the coefficients $\{a_y\},\{b_y\},\{c_y\},\beta$; substitute them into the condition of Lemma 2; and check whether the resulting equality holds across classes. 3. Transfer the result back to the original loss by Lemma 1.

For SCL and BCL, dependence on the class cardinalities $\{B_y\}$ prevents the condition of Lemma 2 from being satisfied, so the regular simplex fails to be even a stationary point of the lower bound and therefore fails to be a global minimizer. For FBCL, the condition holds since structural balancing eliminates dependence on $\{B_y\}$. We further show that this stationary point is indeed a global minimum by deriving subsequent lower bounds whose equality conditions characterize a regular simplex.

C.3.1 SCL does not yield a regular simplex at the optimum under long-tailed distributions

Theorem 1 states that under long-tailed distributions, the SCL loss does not attain a regular simplex as the geometry of its global minimum. We prove this by showing that the regular simplex configuration $\bar{Z}^*$ does not satisfy the first-order Lagrangian condition for $\mathcal{LB}_{\mathrm{SCL}}$, and applying Lemma 1.

Theorem 1. *Let $\bar{Z}^*=(\bar{z}_1^*,\dots,\bar{z}_K^*)$ be a K-point configuration forming a regular simplex on $\mathbb{S}^{h-1}$, and $Z^*=(z_1^*,\dots,z_B^*)$ be a configuration which collapses to $\bar{Z}^*$. Consider the constrained optimization problem of minimizing $\mathcal{LB}_{\mathrm{SCL}}$ defined in Proposition 1 subject to $G_i(\bar{Z})=\|\bar{z}_i\|^2-1=0$ for all $i\in[K]$. Then,*

$$\exists\, i\in[K] \text{ such that } \nexists\,\lambda_i\in\mathbb{R} \text{ for which } \nabla_{\bar{z}_i}\mathcal{LB}_{\mathrm{SCL}}(\bar{Z}^*)=\lambda_i\nabla_{\bar{z}_i}G_i(\bar{Z}^*). \tag{61}$$

Consequently, $\bar{Z}^*$ fails to be a stationary point of $\mathcal{LB}_{\mathrm{SCL}}$, and therefore fails to be a global minimizer. By Lemma 1,

$$Z^*\notin\operatorname*{arg\,min}_{Z\in(\mathbb{S}^{h-1})^B}\mathcal{L}_{\mathrm{SCL}}(Z;Y,\mathcal{B}). \tag{62}$$

Proof of Theorem 1.

Proof. By Proposition 1, $\mathcal{L}_{\mathrm{SCL}}$ is bounded below by

$$\mathcal{L}_{\mathrm{SCL}}(Z;Y,\mathcal{B}) \overset{\text{Prop. 1}}{\geq} \mathcal{LB}_{\mathrm{SCL}}(\bar{Z};K) = \frac{1}{B}\sum_{y\in[K]} B_y \log\left(\Gamma_y\right), \tag{63}$$

$$\text{where} \quad \Gamma_y = B_y - 1 + e^{-1}\sum_{k\in[K]\setminus\{y\}} B_k \exp(\bar{z}_y\cdot\bar{z}_k). \tag{64}$$

Substituting $a_y = B_y$, $b_y = B_y - 1$, $c_y = B_y$, and $\beta = B$ into Eq. (44) yields $\mathcal{LB}_{\mathrm{SCL}}(\bar{Z};K)$. Thus the condition Eq. (47) in Lemma 2 becomes

$$\frac{B_iB_k}{\Gamma_i} + \frac{B_kB_i}{\Gamma_k} = \text{Constant}, \quad \text{for all } k\in[K]\setminus\{i\}, \tag{65}$$

$$\text{where} \quad \Gamma_y = B_y - 1 + e^{d-1}\sum_{k\in[K]\setminus\{y\}} B_k.$$

Since B_i remains fixed in (65), it suffices to examine the dependence on k, determined by

$$\frac{B_k}{\Gamma_i} + \frac{B_k}{\Gamma_k} = \text{Constant}, \quad \text{for all } k\in[K]\setminus\{i\}. \tag{66}$$

Given $K\geq 3$, there exist distinct indices $i\neq r,s\in[K]$ such that $B_r > B_s$. We prove

$$\frac{B_r}{\Gamma_i} + \frac{B_r}{\Gamma_r} \neq \frac{B_s}{\Gamma_i} + \frac{B_s}{\Gamma_s}$$

by showing that $\frac{B_y}{\Gamma_y}$ is a monotonically increasing function of B_y. Indeed,

$$\Gamma_y = B_y - 1 + e^{d-1}\sum_{k\in[K]\setminus\{y\}} B_k \tag{67}$$

$$= B_y - 1 + (B - B_y)e^{d-1} \tag{68}$$

$$= (1-e^{d-1})B_y + (Be^{d-1} - 1). \tag{69}$$

Note that $d = \bar{z}_i^*\cdot\bar{z}_j^* = -1/(K-1)$ for any $i\neq j\in[K]$ by condition (2). Hence $1-e^{d-1} > 0$. Also, given $B > 4$ and $K\geq 3$, we have $Be^{d-1} - 1 > 0$. Therefore, $\frac{B_y}{\Gamma_y}$ is increasing in B_y, and since Γ_i also remains fixed, with $B_r > B_s$,

$$\frac{B_r}{\Gamma_i} + \frac{B_r}{\Gamma_r} > \frac{B_s}{\Gamma_i} + \frac{B_s}{\Gamma_s}.$$

Consequently, condition (65) cannot be satisfied. By Lemma 2, $\bar{Z}^*$ does not satisfy the Lagrangian first-order condition for $\mathcal{LB}_{\mathrm{SCL}}$, i.e.,

$$\bar{Z}^* \notin \operatorname*{arg\,min}_{\bar{Z}\in(\mathbb{S}^{h-1})^K} \mathcal{LB}_{\mathrm{SCL}}(\bar{Z};K). \tag{70}$$

By Lemma 1,

$$Z^* \notin \operatorname*{arg\,min}_{Z\in(\mathbb{S}^{h-1})^B} \mathcal{L}_{\mathrm{SCL}}(Z;Y,\mathcal{B}), \tag{71}$$

i.e., the SCL loss fails to form a regular simplex when attaining its minimum. □

Remark. Yi et al. (2025) and Kini et al. (2024) analyzed the representation geometry of SCL under class imbalance assuming a controlled distribution or the non-negativity of features, but they did not show that SCL fails to form a regular simplex without such assumptions, as Theorem 1 does. Moreover, neither work proposes an objective that provably induces a regular simplex under class imbalance.

C.3.2 BCL DOES NOT YIELD A REGULAR SIMPLEX AT THE OPTIMUM UNDER LONG-TAILED DISTRIBUTIONS

Theorem 2 states that under any long-tailed distribution, the BCL loss does not yield a regular simplex as the geometry of its global minimum.

Why BCL failed to induce a regular simplex. The original work of BCL derives a final lower bound that induces regular simplex conditions. However, our framework reveals that the final lower bound is not tight; this is because the application of Jensen's inequality to $\sum_{i\in\mathcal{B}}$ is not tight (it should be used sequentially to $\sum_{i\in\mathcal{B}_y}$ and then to $\sum_{y\in[K]}$). We show this by showing that the regular simplex configuration already fails to minimize the class-level lower bound, shown below.

Theorem 2. *Let $\bar{Z}^*, Z^*$ be defined as in Theorem 1. Consider the constrained optimization problem of minimizing $\mathcal{LB}_{\mathrm{BCL}}$ defined in Proposition 2 subject to $G_i(\bar{Z}) = \|\bar{z}_i\|^2 - 1 = 0$ for all $i \in [K]$. Then,*

$$\exists\, i \in [K] \text{ such that } \nexists\, \lambda_i \in \mathbb{R} \text{ for which } \nabla_{\bar{z}_i}\mathcal{LB}_{\mathrm{BCL}}(\bar{Z}^*) = \lambda_i \nabla_{\bar{z}_i} G_i(\bar{Z}^*). \tag{72}$$

Consequently, $\bar{Z}^*$ fails to be a stationary point of $\mathcal{LB}_{\mathrm{BCL}}$, and therefore fails to be a global minimizer. By Lemma 1,

$$Z^* \notin \operatorname*{arg\,min}_{Z\in(\mathbb{S}^{h-1})^B} \mathcal{L}_{\mathrm{BCL}}(Z; Y, \mathcal{B}). \tag{73}$$

Proof of Theorem 2.

Proof. By Proposition 2, $\mathcal{L}_{\mathrm{BCL}}$ is bounded below by

$$\mathcal{L}_{\mathrm{BCL}}(Z; Y, \mathcal{B}) \overset{\text{Prop. 2}}{\geq} \mathcal{LB}_{\mathrm{BCL}}(\bar{Z}; K) = \frac{1}{B}\sum_{y\in[K]} B_y \log\left(\Gamma_y\right), \tag{74}$$

$$\text{where} \quad \Gamma_y = 1 + e^{-1} \sum_{k\in[K]\setminus\{y\}} \exp(\bar{z}_y \cdot \bar{z}_k). \tag{75}$$

Substituting $a_y = B_y$, $b_y = c_y = 1$, and $\beta = B$ into Eq. (44) yields $\mathcal{LB}_{\mathrm{BCL}}(\bar{Z}; K)$. Thus the condition Eq. (47) in Lemma 2 becomes

$$\frac{B_i}{\Gamma_i} + \frac{B_k}{\Gamma_k} = \text{Constant}, \quad \text{for all } k \in [K]\setminus\{i\}, \tag{76}$$
$$\text{where} \quad \Gamma_y = 1 + (K-1)e^{d-1}.$$

It follows that $\Gamma_1 = \Gamma_2 = \cdots = \Gamma_K$. Given $K \geq 3$, there exist distinct indices $r, s \in [K]$ such that $B_r > B_s$. Then,

$$\frac{B_i}{\Gamma_i} + \frac{B_r}{\Gamma_r} > \frac{B_i}{\Gamma_i} + \frac{B_s}{\Gamma_s},$$

so condition Eq. (76) cannot be satisfied. By Lemma 2, $\bar{Z}^*$ does not satisfy the Lagrangian first-order condition for $\mathcal{LB}_{\mathrm{BCL}}$, and thus

$$\bar{Z}^* \notin \operatorname*{arg\,min}_{\bar{Z}\in(\mathbb{S}^{h-1})^K} \mathcal{LB}_{\mathrm{BCL}}(\bar{Z}). \tag{77}$$

By Lemma 1,

$$Z^* \notin \operatorname*{arg\,min}_{Z\in(\mathbb{S}^{h-1})^B} \mathcal{L}_{\mathrm{BCL}}(Z; Y, \mathcal{B}), \tag{78}$$

i.e., the BCL loss fails to form a regular simplex when attaining its minimum. □

Optimality of SCL and BCL. Characterizing the optimality of SCL and BCL under long-tailed distribution is a harder problem which has to deal with numerical relationships across angles between class centers (Yi et al., 2025). Therefore, we rather investigate whether a particular geometry is attained at the minimum of an objective.

C.3.3 FBCL YIELDS A REGULAR SIMPLEX AT THE OPTIMUM UNDER LONG-TAILED DISTRIBUTIONS

Balancing the aggregation across anchors. Our analysis reveals that balancing within each anchor-wise loss is not sufficient to remove the dependence of the objective on class cardinality. In SCL, the positive samples are averaged within each anchor-wise loss, and BCL additionally balances the contributions of negative classes in its denominator. Moreover, Zhong et al. (2022), De Alvis

et al. (2025), and Roh et al. (2026) have studied re-balancing supervised contrastive objectives using the effective number of samples (Cui et al., 2019) or class cardinalities in the numerator or denominator; **these methods still perform balancing within each anchor-wise loss**. However, when these anchor-wise losses are aggregated to form the overall objective, each class contributes in proportion to its number of anchors. This remaining class-dependent weighting is consequently inherited by the class-level lower bound, which prevents the regular simplex from being optimal under long-tailed distributions. Structural balancing (SB) normalizes the aggregation of anchor-wise losses across classes. Although this operation can be viewed as a class-dependent weighting of anchor losses, its role here is determined by the geometric analysis: SB removes the remaining class-cardinality dependence from the class-level lower bound and thereby recovers the regular simplex as an optimum. FBCL combines this structural balancing with the positive and negative balancing already present in BCL.

Theorem 3 states that the regular simplex is a global minimizer of the FBCL loss. The proof proceeds in two steps. We first show that the regular simplex configuration $\bar{Z}^*$ satisfies the first-order Lagrangian condition for $\mathcal{LB}_{\mathrm{FBCL}}$, and is therefore a stationary point. We then show that $\bar{Z}^*$ is in fact a global minimizer of $\mathcal{LB}_{\mathrm{FBCL}}$ by further deriving a lower bound, with equality attained at the regular simplex.

Theorem 3. *Let $\bar{Z}^*, Z^*$ be defined as in Theorem 1. Consider the constrained optimization problem of minimizing $\mathcal{LB}_{\mathrm{FBCL}}$ defined in Proposition 3 subject to $G_i(\bar{Z}) = \|\bar{z}_i\|^2 - 1 = 0$ for all $i \in [K]$. We show that, for each $i \in [K]$, there exists $\lambda_i \in \mathbb{R}$ such that*

$$\nabla_{\bar{z}_i} \mathcal{LB}_{\mathrm{FBCL}}(\bar{Z}^*) = \lambda_i \nabla_{\bar{z}_i} G_i(\bar{Z}^*), \tag{79}$$

where $d = \bar{z}_i^* \cdot \bar{z}_j^*$ for any $i \neq j$. Moreover, we show that $\bar{Z}^*$ is a global minimizer of $\mathcal{LB}_{\mathrm{FBCL}}$:

$$Z^* \in \underset{Z \in (\mathbb{S}^{h-1})^B}{\arg\min}\, \mathcal{L}_{\mathrm{FBCL}}(Z; Y, \mathcal{B}). \tag{80}$$

Specifically, the loss is bounded by the following minimum:

$$\mathcal{L}_{\mathrm{FBCL}}(Z; Y, \mathcal{B}) \geq \log\left(1 + (K-1)\exp\left(-\frac{1}{K-1} - 1\right)\right), \tag{81}$$

where equality is attained if and only if Z collapses to Z^*, a regular simplex.

Proof of Theorem 3.

Proof. The proof proceeds in two stages. We first show that the regular simplex $\bar{Z}^*$ is a stationary point of $\mathcal{LB}_{\mathrm{FBCL}}$ via the unified Lagrangian condition. We then extend this to global optimality by deriving the lower bound of $\mathcal{LB}_{\mathrm{FBCL}}$.

Stage 1: Stationary point. By Proposition 3, $\mathcal{L}_{\mathrm{FBCL}}$ is bounded below by

$$\mathcal{L}_{\mathrm{FBCL}}(Z; Y, \mathcal{B}) \overset{\text{Prop. 3}}{\geq} \mathcal{LB}_{\mathrm{FBCL}}(\bar{Z}; K) = \frac{1}{K} \sum_{y \in [K]} \log(\Gamma_y), \tag{82}$$

$$\text{where} \quad \Gamma_y = 1 + e^{-1} \sum_{k \in [K] \setminus \{y\}} \exp(\bar{z}_y \cdot \bar{z}_k). \tag{83}$$

Substituting $a_y = b_y = c_y = 1$ and $\beta = K$ into Eq. (44) yields $\mathcal{LB}_{\mathrm{FBCL}}(\bar{Z}; K)$. Thus the condition Eq. (47) in Lemma 2 becomes

$$\frac{1}{\Gamma_i} + \frac{1}{\Gamma_k} = \text{Constant}, \quad \text{for all } k \in [K] \setminus \{i\}, \tag{84}$$
$$\text{where} \quad \Gamma_y = 1 + (K-1)e^{d-1}.$$

Since $\Gamma_1 = \Gamma_2 = \cdots = \Gamma_K$, condition Eq. (84) is automatically satisfied. The corresponding multiplier follows from Eq. (48):

$$\lambda_i \overset{\text{Eq. (48)}}{=} -\frac{e^{d-1}}{K + K(K-1)e^{d-1}}. \tag{85}$$

By Lemma 2, the regular simplex configuration $\bar{Z}^*$ satisfies the Lagrangian condition, and thus $\bar{Z}^*$ serves as a stationary point of $\mathcal{LB}_{\mathrm{FBCL}}$.

Stage 2: Global optimality. We further derive the lower bound of $\mathcal{LB}_{\mathrm{FBCL}}$ and eventually show that the regular simplex configuration $\bar{Z}^*$ is a global minimizer of $\mathcal{LB}_{\mathrm{FBCL}}$:

$$\mathcal{LB}_{\mathrm{FBCL}}(\bar{Z};K) = \frac{1}{K}\sum_{y\in[K]}\log\left(1+e^{-1}\sum_{k\in[K]\setminus\{y\}}\exp(\bar{z}_y\cdot\bar{z}_k)\right) \tag{86}$$

$$\geq \frac{1}{K}\sum_{y\in[K]}\log\left(1+e^{-1}(K-1)\exp\left(\frac{1}{K-1}\sum_{k\in[K]\setminus\{y\}}\bar{z}_y\cdot\bar{z}_k\right)\right) \tag{87}$$

$$= \frac{1}{K}\sum_{y\in[K]}\phi\left(\frac{1}{K-1}\sum_{k\in[K]\setminus\{y\}}\bar{z}_y\cdot\bar{z}_k\right) \tag{88}$$

$$\geq \phi\left(\frac{1}{K}\sum_{y\in[K]}\frac{1}{K-1}\sum_{k\in[K]\setminus\{y\}}\bar{z}_y\cdot\bar{z}_k\right) \tag{89}$$

$$= \phi\left(\frac{1}{K(K-1)}\sum_{\substack{i,j\in[K]\\ i\neq j}}\bar{z}_i\cdot\bar{z}_j\right) \tag{90}$$

$$\geq \phi\left(-\frac{1}{K-1}\right) \tag{91}$$

$$= \log\left(1+(K-1)\exp\left(-\frac{1}{K-1}-1\right)\right). \tag{92}$$

We justify each step in turn. The transition (86) $\to$ (87) follows from Jensen's inequality: since $\exp(x)$ is convex and $\log(x)$ is monotonically increasing, the inner average over k can be moved inside the exponential. Defining $\phi(x) = \log(1+(K-1)\exp(x-1))$ gives (88). Since ϕ is itself a convex function, Jensen's inequality applied to the outer average over y yields (88) $\to$ (89). The equality (89) $\to$ (90) simply rewrites the double sum. Finally, the transition (90) $\to$ (91) follows from the inequality

$$\sum_{i\neq j}\bar{z}_i\cdot\bar{z}_j = \left\|\sum_{i=1}^{K}\bar{z}_i\right\|^2 - \sum_{i=1}^{K}\|\bar{z}_i\|^2 \geq -K, \tag{93}$$

which gives $\sum_{i\neq j}\bar{z}_i\cdot\bar{z}_j \geq -K$, and hence $\frac{1}{K(K-1)}\sum_{i\neq j}\bar{z}_i\cdot\bar{z}_j \geq -\frac{1}{K-1}$. Combined with the monotonicity of ϕ, this yields (91).

Equality is attained if and only if both

(1) for all $i\neq j\in[K]$, $\bar{z}_i\cdot\bar{z}_j = -\frac{1}{K-1}$, and

(2) $\sum_{i=1}^{K}\bar{z}_i = 0$

hold. These conditions are identical to the regular simplex conditions of Definition 3 Therefore, attaining the minimum of $\mathcal{LB}_{\mathrm{FBCL}}$, and consequently of $\mathcal{L}$, yields a regular simplex.

$$\log\left(1+(K-1)\exp\left(-\tfrac{1}{K-1}-1\right)\right) = \log\left(1+(K-1)\exp\left(-\tfrac{K}{K-1}\right)\right). \tag{94}$$

□

C.4 Lower bounds of supervised contrastive objectives

The class-level lower bounds $\mathcal{LB}_{\mathrm{SCL}}$, $\mathcal{LB}_{\mathrm{BCL}}$, and $\mathcal{LB}_{\mathrm{FBCL}}$ (Propositions 1, 2, and 3) are derived through a common two-step procedure: an anchor-wise lower bound followed by a class-wise lower bound. We state these two steps as Lemmas 3 and 4, and then apply them sequentially to obtain each objective. The three losses differ only in the weights assigned to the negative classes in the denominator. We unify this difference by introducing a coefficient w_k on each negative class, where $w_k = B_k$ for SCL and $w_k = 1$ for BCL and FBCL.

C.4.1 Lower bound of the anchor-wise loss (Lemma 3)

Lemma 3 provides a lower bound for the anchor-wise loss of a single sample z_i.

Lemma 3. *(Lower bound of the anchor-wise loss) Let $Z, \bar{Z}, Y$ be defined as in Proposition 1. For any Z and Y, fix an anchor z_i with class label y. Then, the anchor-wise loss is bounded below by:*

$$-\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}\log\left(\frac{\exp(z_i\cdot z_j)}{\sum_{k\in[K]}\sum_{m\in\mathcal{B}_k}\frac{w_k}{B_k}\exp(z_i\cdot z_m)}\right)\geq\log\left(w_y+e^{-1}\sum_{k\in[K]\setminus\{y\}}w_k\exp(z_i\cdot\bar{z}_k)\right), \tag{95}$$

where B_k is replaced by B_k-1 when $k=y$. The case $w_k=B_k$ (with $w_y=B_y-1$ when $k=y$) corresponds to SCL, and $w_k=1$ for all $k\in[K]$ corresponds to BCL and FBCL. Equality is attained if and only if the following conditions are satisfied:

(C2) $z_i=z_j$, for all $i,j\in\mathcal{B}_y$.

(C3) For each $k\in[K]\setminus\{y\}$, there exists $\delta_{k,i}\in\mathbb{R}$ such that $z_i\cdot z_m=\delta_{k,i}$, for all $m\in\mathcal{B}_k$.

Proof. The anchor-wise loss can be rewritten as

$$-\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}\log\left(\frac{\exp(z_i\cdot z_j)}{\sum_{k\in[K]}\sum_{m\in\mathcal{B}_k}\frac{w_k}{B_k}\exp(z_i\cdot z_m)}\right) \tag{96}$$

$$=\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}\log\left(\frac{\sum_{k\in[K]}\sum_{m\in\mathcal{B}_k}\frac{w_k}{B_k}\exp(z_i\cdot z_m)}{\exp(z_i\cdot z_j)}\right) \tag{97}$$

$$=\log\left(\frac{\sum_{k\in[K]}\sum_{m\in\mathcal{B}_k}\frac{w_k}{B_k}\exp(z_i\cdot z_m)}{\exp\left(\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}z_i\cdot z_j\right)}\right) \tag{98}$$

$$=\log\left(\frac{\sum_{p\in\mathcal{B}_y\setminus\{i\}}\frac{w_y}{B_y-1}\exp(z_i\cdot z_p)+\sum_{k\in[K]\setminus\{y\}}\sum_{m\in\mathcal{B}_k}\frac{w_k}{B_k}\exp(z_i\cdot z_m)}{\exp\left(\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}z_i\cdot z_j\right)}\right). \tag{99}$$

Applying Jensen's inequality to the first term in the numerator, we obtain

$$\frac{1}{B_y-1}\sum_{p\in\mathcal{B}_y\setminus\{i\}}\exp\left(z_i\cdot z_p\right)\geq\exp\left(\frac{1}{B_y-1}\sum_{p\in\mathcal{B}_y\setminus\{i\}}z_i\cdot z_p\right), \tag{100}$$

where equality is attained if and only if

(a) there is a constant δ_i such that $z_i\cdot z_p=\delta_i,$ for all $p\in\mathcal{B}_y\setminus\{i\}$.

Applying Jensen's inequality to the second term in the numerator, we obtain

$$\sum_{k\in[K]\setminus\{y\}}\frac{1}{B_k}\sum_{m\in\mathcal{B}_k}\exp(z_i\cdot z_m)\geq\sum_{k\in[K]\setminus\{y\}}\exp\left(\frac{1}{B_k}\sum_{m\in\mathcal{B}_k}z_i\cdot z_m\right), \tag{101}$$

where equality is attained if and only if

(b) for each $k\in[K]\setminus\{y\}$, there is a constant $\delta'_{k,i}$ such that $z_i\cdot z_m=\delta'_{k,i},$ for all $m\in\mathcal{B}_k$.

Applying the Cauchy-Schwarz inequality to the denominator, we obtain

$$\frac{1}{\exp(\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}z_i\cdot z_j)}\geq\frac{1}{\exp(\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}\|z_i\|\|z_j\|)}=\frac{1}{\exp(1)}=e^{-1}, \tag{102}$$

where equality is attained if and only if

(c) $z_i = z_j$ for all $i, j \in \mathcal{B}_y$.

Substituting the three inequalities (100), (101), and (102) into the rewritten anchor-wise loss yields

$$\log\left(\frac{\sum_{p\in\mathcal{B}_y\setminus\{i\}}\frac{w_y}{B_y-1}\exp(z_i\cdot z_p)+\sum_{k\in[K]\setminus\{y\}}\sum_{m\in\mathcal{B}_k}\frac{w_k}{B_k}\exp(z_i\cdot z_m)}{\exp\left(\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}z_i\cdot z_j\right)}\right) \tag{103}$$

$$\overset{\text{Eq.(100)}}{\geq}\log\left(\frac{w_y\exp\left(\frac{1}{B_y-1}\sum_{p\in\mathcal{B}_y\setminus\{i\}}z_i\cdot z_p\right)+\sum_{k\in[K]\setminus\{y\}}\sum_{m\in\mathcal{B}_k}\frac{w_k}{B_k}\exp(z_i\cdot z_m)}{\exp\left(\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}z_i\cdot z_j\right)}\right) \tag{104}$$

$$=\log\left(w_y+\frac{\sum_{k\in[K]\setminus\{y\}}\sum_{m\in\mathcal{B}_k}\frac{w_k}{B_k}\exp(z_i\cdot z_m)}{\exp\left(\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}z_i\cdot z_j\right)}\right) \tag{105}$$

$$\overset{\text{Eq.(101)}}{\geq}\log\left(w_y+\frac{\sum_{k\in[K]\setminus\{y\}}w_k\exp\left(\frac{1}{B_k}\sum_{m\in\mathcal{B}_k}z_i\cdot z_m\right)}{\exp\left(\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}z_i\cdot z_j\right)}\right) \tag{106}$$

$$\overset{\text{Eq.(102)}}{\geq}\log\left(w_y+e^{-1}\sum_{k\in[K]\setminus\{y\}}w_k\exp\left(z_i\cdot\bar{z}_k\right)\right), \tag{107}$$

where equality is attained if and only if conditions (C2) and (C3) are satisfied. □

C.4.2 Lower bound of the class-wise loss (Lemma 4)

Lemma 4 provides a lower bound for the class-wise loss obtained by averaging the anchor-wise lower bound over a single class.

Lemma 4. *(Lower bound of the class-wise loss) Let $Z, \bar{Z}, Y$ be defined as in Proposition 1. For any Z and Y, fix a class y and let $w_k \geq 0$ for all $k \in [K]$. Then,*

$$\frac{1}{B_y}\sum_{i\in\mathcal{B}_y}\log\left(w_y+e^{-1}\sum_{k\in[K]\setminus\{y\}}w_k\exp(z_i\cdot\bar{z}_k)\right)\geq\log\left(w_y+e^{-1}\sum_{k\in[K]\setminus\{y\}}w_k\exp(\bar{z}_y\cdot\bar{z}_k)\right), \tag{108}$$

where equality is attained if and only if the following condition is satisfied:

(C4) For each $k \in [K]\setminus\{y\}$, there is a constant δ_k such that $z_i \cdot \bar{z}_k = \delta_k$, for all $i \in \mathcal{B}_y$.

Proof. The inequality follows from applying Jensen's inequality to the function

$$f(z)=\log\left(w_y+e^{-1}\sum_{k\in[K]\setminus\{y\}}w_k\exp(z\cdot\bar{z}_k)\right).$$

Since f is convex, we have $\frac{1}{B_y}\sum_{i\in\mathcal{B}_y}f(z_i)\geq f\left(\frac{1}{B_y}\sum_{i\in\mathcal{B}_y}z_i\right)$, which yields the desired result. □

C.4.3 Proofs of Propositions 1, 2, and 3

We now derive the class-level lower bounds of the three losses by applying Lemma 3 to bound each anchor-wise loss, and then Lemma 4 to bound each class-wise loss. The three propositions differ only in the choice of w_k: $w_k = B_k$ for SCL, and $w_k = 1$ for BCL and FBCL. Combining the equality conditions (C2), (C3), and (C4) across all $y \in [K]$ yields the class-collapse condition (C1).

(C1) (**Class-collapse condition**) For all $y \in [K]$, $z_i = \bar{z}_y$ for all $i \in \mathcal{B}_y$.

Proposition 1. *(Lower bound of SCL loss) Let $Z = (z_1, \dots, z_B) \in (\mathbb{S}^{h-1})^B$ be a B-point configuration with labels $Y = (y_1, \dots, y_B) \in [K]^B$. Then,*

$$\mathcal{L}_{\mathrm{SCL}}(Z; Y, \mathcal{B}) \geq \frac{1}{B} \sum_{y \in [K]} B_y \log \left(B_y - 1 + e^{-1} \sum_{k \in [K] \setminus \{y\}} B_k \exp(\bar{z}_y \cdot \bar{z}_k) \right) =: \mathcal{LB}_{\mathrm{SCL}}(\bar{Z}; K), \tag{109}$$

where $\bar{z}_y = \frac{1}{B_y} \sum_{i \in \mathcal{B}_y} z_i$ and $\bar{Z} = (\bar{z}_1, \dots, \bar{z}_K)$. Equality is attained if and only if condition (C1) is satisfied.

Proof of Proposition 1.

Proof. Setting $w_k = B_k$ (with $w_y = B_y - 1$ when $k = y$), we apply Lemmas 3 and 4 sequentially:

$$\mathcal{L}_{\mathrm{SCL}}(Z; Y, \mathcal{B}) = -\frac{1}{B} \sum_{y \in [K]} \sum_{i \in \mathcal{B}_y} \frac{1}{B_y - 1} \sum_{j \in \mathcal{B}_y \setminus \{i\}} \log \left(\frac{\exp(z_i \cdot z_j)}{\sum_{k \in \mathcal{B} \setminus \{i\}} \exp(z_i \cdot z_k)} \right) \tag{110}$$

$$\overset{\text{Lem. 3}}{\geq} \frac{1}{B} \sum_{y \in [K]} \sum_{i \in \mathcal{B}_y} \log \left(B_y - 1 + e^{-1} \sum_{k \in [K] \setminus \{y\}} B_k \exp(z_i \cdot \bar{z}_k) \right) \tag{111}$$

$$= \frac{1}{B} \sum_{y \in [K]} B_y \frac{1}{B_y} \sum_{i \in \mathcal{B}_y} \log \left(B_y - 1 + e^{-1} \sum_{k \in [K] \setminus \{y\}} B_k \exp(z_i \cdot \bar{z}_k) \right) \tag{112}$$

$$\overset{\text{Lem. 4}}{\geq} \frac{1}{B} \sum_{y \in [K]} B_y \log \left(B_y - 1 + e^{-1} \sum_{k \in [K] \setminus \{y\}} B_k \exp(\bar{z}_y \cdot \bar{z}_k) \right). \tag{113}$$

□

Proposition 2. *(Lower bound of BCL loss) Let $Z, \bar{Z}, Y$ be defined as in Proposition 1. Then,*

$$\mathcal{L}_{\mathrm{BCL}}(Z; Y, \mathcal{B}) \geq \frac{1}{B} \sum_{y \in [K]} B_y \log \left(1 + e^{-1} \sum_{k \in [K] \setminus \{y\}} \exp(\bar{z}_y \cdot \bar{z}_k) \right) =: \mathcal{LB}_{\mathrm{BCL}}(\bar{Z}; K). \tag{114}$$

Equality is attained if and only if condition (C1) is satisfied.

Proof of Proposition 2.

Proof. Setting $w_k = 1$ for all $k \in [K]$, we apply Lemmas 3 and 4 sequentially:

$$\mathcal{L}_{\mathrm{BCL}}(Z; Y, \mathcal{B}) = -\frac{1}{B} \sum_{y \in [K]} \sum_{i \in \mathcal{B}_y} \frac{1}{B_y - 1} \sum_{j \in \mathcal{B}_y \setminus \{i\}} \log \left(\frac{\exp(z_i \cdot z_j)}{\sum_{k \in [K]} \frac{1}{B_k} \sum_{m \in \mathcal{B}_k} \exp(z_i \cdot z_m)} \right) \tag{115}$$

$$\overset{\text{Lem. 3}}{\geq} \frac{1}{B} \sum_{y \in [K]} \sum_{i \in \mathcal{B}_y} \log \left(1 + e^{-1} \sum_{k \in [K] \setminus \{y\}} \exp(z_i \cdot \bar{z}_k) \right) \tag{116}$$

$$= \frac{1}{B} \sum_{y \in [K]} B_y \frac{1}{B_y} \sum_{i \in \mathcal{B}_y} \log \left(1 + e^{-1} \sum_{k \in [K] \setminus \{y\}} \exp(z_i \cdot \bar{z}_k) \right) \tag{117}$$

$$\overset{\text{Lem. 4}}{\geq} \frac{1}{B} \sum_{y \in [K]} B_y \log \left(1 + e^{-1} \sum_{k \in [K] \setminus \{y\}} \exp(\bar{z}_y \cdot \bar{z}_k) \right). \tag{118}$$

□

Proposition 3. *(Lower bound of FBCL loss) Let $Z, \bar{Z}, Y$ be defined as in Proposition 1. Then,*

$$\mathcal{L}_{\mathrm{FBCL}}(Z;Y,\mathcal{B}) \geq \frac{1}{K}\sum_{y\in[K]}\log\left(1+e^{-1}\sum_{k\in[K]\setminus\{y\}}\exp(\bar{z}_y\cdot\bar{z}_k)\right) =: \mathcal{LB}_{\mathrm{FBCL}}(\bar{Z};K). \tag{119}$$

Equality is attained if and only if condition (C1) is satisfied.

Proof of Proposition 3.

Proof. Setting $w_k = 1$ for all $k \in [K]$, we apply Lemmas 3 and 4 sequentially:

$$\mathcal{L}_{\mathrm{FBCL}}(Z;Y,\mathcal{B}) = -\frac{1}{K}\sum_{y\in[K]}\frac{1}{B_y}\sum_{i\in\mathcal{B}_y}\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}\log\left(\frac{\exp(z_i\cdot z_j)}{\sum_{k\in[K]}\frac{1}{B_k}\sum_{m\in\mathcal{B}_k}\exp(z_i\cdot z_m)}\right) \tag{120}$$

$$\overset{\text{Lem. 3}}{\geq} \frac{1}{K}\sum_{y\in[K]}\frac{1}{B_y}\sum_{i\in\mathcal{B}_y}\log\left(1+e^{-1}\sum_{k\in[K]\setminus\{y\}}\exp(z_i\cdot\bar{z}_k)\right) \tag{121}$$

$$\overset{\text{Lem. 4}}{\geq} \frac{1}{K}\sum_{y\in[K]}\log\left(1+e^{-1}\sum_{k\in[K]\setminus\{y\}}\exp(\bar{z}_y\cdot\bar{z}_k)\right). \tag{122}$$

□

C.5 Decomposition of losses

We rewrite each loss by separating the numerator and the denominator inside the log into an alignment term $\mathcal{L}^{\text{alignment}}$ and a uniformity term $\mathcal{L}^{\text{uniformity}}$ (we discard $(Z;Y,\mathcal{B})$ for brevity):

$$\mathcal{L}_{\mathrm{SCL}} = \mathcal{L}^{\text{alignment}}_{\mathrm{SCL}} + \mathcal{L}^{\text{uniformity}}_{\mathrm{SCL}}, \tag{123}$$

$$\mathcal{L}_{\mathrm{BCL}} = \mathcal{L}^{\text{alignment}}_{\mathrm{BCL}} + \mathcal{L}^{\text{uniformity}}_{\mathrm{BCL}}, \tag{124}$$

$$\mathcal{L}_{\mathrm{FBCL}} = \mathcal{L}^{\text{alignment}}_{\mathrm{FBCL}} + \mathcal{L}^{\text{uniformity}}_{\mathrm{FBCL}}, \tag{125}$$

where the alignment terms are given by

$$\mathcal{L}^{\text{alignment}}_{\mathrm{SCL}} = \mathcal{L}^{\text{alignment}}_{\mathrm{BCL}} = -\frac{1}{B}\sum_{y\in[K]}\sum_{i\in\mathcal{B}_y}\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}(z_i\cdot z_j), \tag{126}$$

$$\mathcal{L}^{\text{alignment}}_{\mathrm{FBCL}} = -\frac{1}{K}\sum_{y\in[K]}\frac{1}{B_y}\sum_{i\in\mathcal{B}_y}\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}(z_i\cdot z_j), \tag{127}$$

and the uniformity terms are given by

$$\mathcal{L}^{\text{uniformity}}_{\mathrm{SCL}} = \frac{1}{B}\sum_{y\in[K]}\sum_{i\in\mathcal{B}_y}\log\left(\sum_{k\in\mathcal{B}\setminus\{i\}}\exp(z_i\cdot z_k)\right), \tag{128}$$

$$\mathcal{L}^{\text{uniformity}}_{\mathrm{BCL}} = \frac{1}{B}\sum_{y\in[K]}\sum_{i\in\mathcal{B}_y}\log\left(\sum_{k\in[K]}\frac{1}{B_k}\sum_{m\in\mathcal{B}_k}\exp(z_i\cdot z_m)\right), \tag{129}$$

$$\mathcal{L}^{\text{uniformity}}_{\mathrm{FBCL}} = \frac{1}{K}\sum_{y\in[K]}\frac{1}{B_y}\sum_{i\in\mathcal{B}_y}\log\left(\sum_{k\in[K]}\frac{1}{B_k}\sum_{m\in\mathcal{B}_k}\exp(z_i\cdot z_m)\right). \tag{130}$$

The uniformity term cannot be feasibly written with respect to the intra-class variance without any assumptions. We hope that future work could assume a distribution of the features and further analyze how the uniformity term penalizes intra-class variances.

Alignment term as a function of intra-class variance. Let $V_y = \frac{1}{B_y}\sum_{i\in\mathcal{B}_y}\|z_i - \bar{z}_y\|^2$. The alignment term of SCL or BCL (we only write for the SCL case) can be written using the following identities:

$$V_y = \frac{1}{B_y}\sum_{i\in\mathcal{B}_y}\|z_i - \bar{z}_y\|^2 \tag{131}$$

$$= \frac{1}{B_y}\sum_{i\in\mathcal{B}_y}\left(\|z_i\|^2 - 2z_i\cdot\bar{z}_y + \|\bar{z}_y\|^2\right) \tag{132}$$

$$= \frac{1}{B_y}\sum_{i\in\mathcal{B}_y}1 - \frac{2}{B_y}\sum_{i\in\mathcal{B}_y}z_i\cdot\bar{z}_y + \frac{1}{B_y}\sum_{i\in\mathcal{B}_y}\|\bar{z}_y\|^2 \tag{133}$$

$$= 1 - \|\bar{z}_y\|^2, \tag{134}$$

$$\sum_{i\in\mathcal{B}_y}\sum_{j\in\mathcal{B}_y\setminus\{i\}} z_i\cdot z_j = \sum_{i\in\mathcal{B}_y}\sum_{j\in\mathcal{B}_y} z_i\cdot z_j - \sum_{i\in\mathcal{B}_y}\|z_i\|^2 \tag{135}$$

$$= \left\|\sum_{i\in\mathcal{B}_y} z_i\right\|^2 - \sum_{i\in\mathcal{B}_y}\|z_i\|^2 \tag{136}$$

$$= B_y^2\|\bar{z}_y\|^2 - B_y \tag{137}$$

$$= B_y^2(\|\bar{z}_y\|^2 - 1) + B_y^2 - B_y \tag{138}$$

$$= -B_y^2(V_y) + B_y(B_y - 1). \tag{139}$$

Using the above,

$$\mathcal{L}_{\text{SCL}}^{\text{alignment}} = -\frac{1}{B}\sum_{y\in[K]}\sum_{i\in\mathcal{B}_y}\frac{1}{B_y - 1}\sum_{j\in\mathcal{B}_y\setminus\{i\}} z_i\cdot z_j \tag{140}$$

$$= -\frac{1}{B}\sum_{y\in[K]}\frac{1}{B_y - 1}\sum_{i\in\mathcal{B}_y}\sum_{j\in\mathcal{B}_y\setminus\{i\}} z_i\cdot z_j \tag{141}$$

$$= -\frac{1}{B}\sum_{y\in[K]}\frac{1}{B_y - 1}\big(-B_y^2(V_y) + B_y(B_y - 1)\big) \tag{142}$$

$$= \sum_{y\in[K]}\frac{B_y^2}{B(B_y - 1)}V_y - \sum_{y\in[K]}\frac{1}{B}\frac{B_y(B_y - 1)}{(B_y - 1)} \tag{143}$$

$$= \sum_{y\in[K]}\frac{B_y^2}{B(B_y - 1)}V_y - 1. \tag{144}$$

Similarly,

$$\mathcal{L}_{\text{FBCL}}^{\text{alignment}} = -\frac{1}{K}\sum_{y\in[K]}\frac{1}{B_y}\sum_{i\in\mathcal{B}_y}\frac{1}{B_y - 1}\sum_{j\in\mathcal{B}_y\setminus\{i\}} z_i\cdot z_j \tag{145}$$

$$= -\frac{1}{K}\sum_{y\in[K]}\frac{1}{B_y(B_y - 1)}\sum_{i\in\mathcal{B}_y}\sum_{j\in\mathcal{B}_y\setminus\{i\}} z_i\cdot z_j \tag{146}$$

$$= -\frac{1}{K}\sum_{y\in[K]}\frac{1}{B_y(B_y - 1)}\big(-B_y^2(V_y) + B_y(B_y - 1)\big) \tag{147}$$

$$= \sum_{y\in[K]}\frac{B_y}{K(B_y - 1)}V_y - \sum_{y\in[K]}\frac{1}{K}\frac{B_y(B_y - 1)}{B_y(B_y - 1)} \tag{148}$$

$$= \sum_{y\in[K]}\frac{B_y}{K(B_y - 1)}V_y - 1. \tag{149}$$

C.6 Balancing components in FBCL

As described in Section 4.2, FBCL contains three balancing components: positive balancing (PB), which averages over positive samples as in SCL; negative balancing (NB), which averages negative samples per class in the denominator as in BCL; and structural balancing (SB), which normalizes the aggregation of anchor-wise losses. The following equation marks each method within the FBCL loss:

$$\mathcal{L}_{\mathrm{FBCL}}(Z;Y,\mathcal{B}) = -\frac{1}{K}\sum_{y\in[K]} \underbrace{\frac{1}{B_y}}_{\text{SB}} \sum_{i\in\mathcal{B}_y} \underbrace{\frac{1}{B_y-1}\sum_{j\in\mathcal{B}_y\setminus\{i\}}}_{\text{PB}} \log\Big(\frac{\exp(z_i\cdot z_j)}{\sum_{k\in[K]} \underbrace{\frac{1}{B_k}}_{\text{NB}} \sum_{m\in\mathcal{B}_k}\exp(z_i\cdot z_m)}\Big).$$

For positive balancing, we include the summation of positive samples so that a method without balancing corresponds to SimCLR.

D Experimental Details

D.1 Figure 1 Details

We use three random classes from ImageNet, and resize the images to 32×32 and train a 2-layer MLP that maps each flattened image through a hidden layer of dimension 256 to a two-dimensional normalized embedding. Thus, the learned features lie on the unit circle in the two-dimensional feature space. We therefore visualize the embeddings directly on this circle. We train all models using the SGD optimizer with learning rate 0.1, weight decay $5e-4$, momentum 0.9.

D.2 Implementation Details and Hyperparameters

FBCL loss in practice. As we discussed in Section 6, our analysis follows the standard setting where each batch contains all classes. We also note that the temperature rate is fixed to $\tau = 1$ in our theory, since the value does not affect the theory. In our settings, we adopt the temperature rate used in the baseline methods for a fair comparison. But in practice, the representation space is usually affected by the value of the temperature rate under both balanced and imbalanced distributions (Wang & Liu, 2021; Kukleva et al., 2023). We hope that further studies investigate the attainability of a regular simplex under a large-batch training regime and its relation to temperature. Since the goal of our work is not to attain state-of-the-art results, we carefully conduct experiments to do a fair comparison with the baseline methods. We use the following loss in practice:

$$\mathcal{L}_{\mathrm{FBCL}}(Z;Y,\mathcal{B}) = -\frac{1}{|\mathcal{Y}_\mathcal{B}|}\sum_{y\in\mathcal{Y}_\mathcal{B}} \frac{1}{B_y}\sum_{i\in\mathcal{B}_y}\frac{1}{B_y}\sum_{j\in\{\mathcal{B}_y\cup\{c_y\}\}\setminus\{i\}} \log\left(\frac{\exp(z_i\cdot z_j/\tau)}{\sum_{k\in[K]}\frac{1}{B_k+1}\sum_{m\in\{\mathcal{B}_k\cup\{c_k\}\}}\exp(z_i\cdot z_m/\tau)}\right), \tag{150}$$

where $\mathcal{Y}_\mathcal{B}$ is the index set of class labels in the batch $\mathcal{B}$, and $\{c_y\}$ are class-complemented pairs. The difference between practice and our theory lies in the class distribution of the anchors. Our theory assumes that the anchor classes cover the entire label set, $[K]$, whereas in practice, the set of anchor classes depends on the mini-batch composition. **Thus, increasing the batch size is expected to improve class coverage and make the practical objective potentially encouraging a regular simplex geometry even on large-scale datasets**. In our experiments, however, the batch size was limited by GPU memory, and we leave a systematic investigation of this regime to future work. To further analyze this regime, simply increasing the batch size may still be insufficient for severe long-tailed datasets. For example, ImageNet-LT contains 1,000 classes, and even with a batch size of 1,024, few-shot classes may appear only rarely as anchors. One possible extension is therefore to augment the anchor set with class prototypes of few-shot classes. In particular, the same class prototypes introduced as complementary contrastive pairs could serve as anchors for few-shot classes, increasing their anchor coverage without requiring substantially larger mini-batches. We leave the theoretical and empirical study of such prototype-based anchor augmentation for future work.

ImageNet-LT and iNaturalist. Our setting mainly follows Zhu et al. (2022). When combining with other methods, we follow their settings and add structural balancing. For ImageNet-LT, we train the model for 90 epochs with learning rate 0.1, and weight decay 5×10^{-4}. For iNaturalist, we train for 100 epochs with learning rate 0.2, and weight decay 1×10^{-4}. For both datasets, we

use a batch size of 256, cosine scheduling for learning rate, and set $\lambda = 1.0$ and $\mu = 0.35$. Class-complement (Zhu et al., 2022) is used during training to incorporate class centers as additional contrastive pairs, allowing all classes to appear in every mini-batch. For SCL, class-complement is not used as Zhu et al. (2022) reported performance degradation when used. SimCLR also does not use class-complement as it is a self-supervised method. The temperature is set to $\tau = 0.07$. All models are trained using SGD with momentum 0.9, using four RTX 4090 GPUs.

CIFAR-100-LT. For CIFAR-100, we train the model for 200 epochs with learning rate warmed up to 0.15 within the first 5 epochs, decaying at 160 and 180 with a step size 0.1, and weight decay 5×10^{-4}. We set the temperature to $\tau = 0.1$, and set $\lambda = 2.0$ and $\mu = 0.6$. All models are trained using SGD with momentum 0.9, using one RTX 4090 GPU.

D.3 Transfer Learning Evaluation protocol

Linear evaluation The linear evaluation protocol follows Chen et al. (2020); Oh & Lee (2024). Specifically, we divide the training dataset into a train set and a validation set as shown in Table 5 and tune the regularization parameter by minimizing the L2-regularized cross-entropy objective using L-BFGS. Images are only resized and 224×224 center-cropped. With the best parameter, we retrain the linear classifier using both training and validation splits and then report the test accuracy.

Fine-tuning protocol We divide the training dataset into a train set and a validation set as in linear evaluation. Training images are resized and randomly cropped to 224×224, and then randomly horizontally flipped. We perform grid search over the number of epochs, learning rate, and weight decay using the validation split with SGD with momentum 0.9. With the best hyperparameters, we retrain using both training and validation splits and then report the model accuracy on the test set.

Table 5: Dataset information

Dataset	# of classes	Training	Validation	Test	Metric
CIFAR-10	10	45000	5000	10000	Top-1 accuracy
CIFAR-100	100	45000	5000	10000	Top-1 accuracy
MIT67	67	4690	670	1340	Top-1 accuracy
DTD (split1)	47	1880	1880	1880	Top-1 accuracy
CUB200	200	4994	1000	5794	Top-1 accuracy
Food	101	68175	7575	25250	Top-1 accuracy
Cars	196	6494	1650	8041	Top-1 accuracy
Pets	37	2940	740	3669	Mean per-class accuracy
Caltech101	101	2525	505	5647	Mean per-class accuracy
Flowers	102	1020	1020	6149	Mean per-class accuracy
Aircraft	100	3334	3333	3333	Mean per-class accuracy

D.4 Toy Experiment

The detailed settings of the toy dataset used in Section 4.3 follow Oh & Lee (2024). The dataset consists of five classes, each following a Gaussian distribution with orthogonal mean vectors and a shared isotropic covariance matrix with a scale of 0.15. The mean vectors are obtained from the left singular vectors of a random matrix sampled from a standard Gaussian. The data dimension is set to 6, and the class cardinalities are $500, 250, 100, 50, 25$. The first class is treated as the many class, the second class as the medium class, and the remaining three classes as few classes. The test dataset consists of 500 samples, with 100 samples per class.

We use a two-layer MLP backbone with 32 hidden dimensions, followed by a classifier. For the cross-entropy baseline, we train the backbone and classifier for 2000 epochs. For contrastive objectives, we first train the backbone model for 2000 epochs and then train the classifier for 400 epochs. We use SGD with learning rate 0.1, weight decay $5e-4$, and temperature 0.25 for contrastive training.

For the intra-class variance analysis, we first compute the within-class variance and the between-class variance from the learned representations. Let S_W and S_B denote the within-class and

between-class variances, respectively, and let $S_T = S_W + S_B$ be the total variance. For each class y, we compute the class-wise intra-class variance and normalize it using S_T. We then report the average normalized intra-class variance for each group. This normalization allows us to compare how much variance each class retains relative to the overall representation variance.

D.5 More details and standard deviations

D.5.1 FBCL on ImageNet-LT and iNaturalist

Here, statistics and iNaturalist results for Table 1 are provided.

Table 6: Statistics over three runs, on contrastive objectives on ImageNet-LT and iNaturalist.

Method	ImageNet-LT				iNaturalist			
	Many	Medium	Few	All	Many	Medium	Few	All
SimCLR (Chen et al., 2020)	65.0 ±0.1	52.6 ±0.2	36.2 ±0.1	55.2 ±0.2	68.1 ±0.3	70.7 ±0.2	69.9 ±0.3	70.1 ±0.2
SCL (Khosla et al., 2020)	65.8 ±0.4	53.3 ±0.4	36.1 ±0.5	55.8 ±0.4	68.4 ±0.3	71.1 ±0.4	70.2 ±0.5	70.4 ±0.3
BCL (Zhu et al., 2022)	65.8 ±0.2	53.7 ±0.2	36.2 ±0.1	56.0 ±0.1	70.0 ±0.7	72.0 ±0.4	71.1 ±0.2	71.3 ±0.6
FBCL (Ours)	66.4±0.2	54.1 ±0.2	37.5 ±0.4	56.6 ±0.1	70.4 ±0.3	72.8 ±0.2	71.9 ±0.3	72.2 ±0.1

D.5.2 Intra-class variance and accuracy

Here, statistics for Table 4, including intra-class variance and accuracy on synthetic data, ImageNet-LT and CIFAR-100-LT (imbalance ratio=100), are provided in Tables 7, 8, and 9.

Table 7: Statistics over five runs on intra-class variance and Top-1 accuracy on synthetic data.

Method	Many	Medium	Few	Acc
CE	0.459 ± 0.02	0.533 ± 0.03	0.721 ± 0.03	73.8 ± 0.16
SCL	0.272 ± 0.01	0.346 ± 0.02	0.575 ± 0.05	74.1 ± 2.74
BCL	0.250 ± 0.02	0.385 ± 0.03	0.749 ± 0.07	74.1 ± 2.36
FBCL (Ours)	0.409 ± 0.02	0.453 ± 0.02	0.438 ± 0.03	75.5 ± 3.10

Table 8: Statistics over five runs of the ablation study over balancing components on CIFAR-100-LT (IR=100).

Method	**PB**	**NB**	**SB**	**Intra-class variance**			**Acc.**
				Many	Medium	Few	
PB (SCL)	✓	✗	✗	0.211±0.001	0.342±0.003	0.420±0.001	51.5±0.18
PB+NB (BCL)	✓	✓	✗	0.210±0.005	0.382±0.004	0.465±0.011	51.7±0.43
PB+SB	✓	✗	✓	0.340±0.004	0.387±0.002	0.467±0.003	51.5±0.25
NB+SB	✗	✓	✓	0.660±0.009	0.620±0.006	0.601±0.009	51.4±0.16
PB+NB+SB (Ours)	✓	✓	✓	0.267±0.003	0.320±0.006	0.379±0.003	52.2±0.23

Table 9: Ablation study of positive balancing (PB), negative balancing (NB), and structural balancing (SB) on ImageNet-LT. Statistics over three runs are given.

Method	**PB**	**NB**	**SB**	**Intra-class variance**			**Acc.**			
				Many	Medium	Few	Many	Medium	Few	All
PB (SCL)	✓	✗	✗	0.189±0.001	0.298±0.003	0.357±0.001	65.76±0.38	53.33±0.40	36.13±0.46	55.78±0.41
PB+NB (BCL)	✓	✓	✗	0.167±0.021	0.293±0.003	0.362±0.006	65.83±0.15	53.74±0.17	36.23±0.06	56.03±0.07
PB+SB	✓	✗	✓	0.368±0.004	0.502±0.003	0.509±0.003	66.07±0.12	53.40±0.30	36.22±0.36	55.92±0.19
NB+SB	✗	✓	✓	0.597±0.002	0.623±0.002	0.602±0.001	65.20±0.20	53.03±0.29	35.47±0.32	55.35±0.17
PB+NB+SB (Ours)	✓	✓	✓	0.196±0.008	0.281±0.010	0.339±0.008	66.44±0.15	54.12±0.21	37.50±0.40	56.64±0.12

D.5.3 STRUCTURAL BALANCING WITH OTHER METHODS

Here, statistics for the structural balancing results in Table 2 are provided in Tables 10 and 11.

Table 10: Statistics over three runs of FBCL and BCL combined with DBM.

Method	ImageNet-LT				iNaturalist			
	Many	Medium	Few	All	Many	Medium	Few	All
DBM+BCL	64.4 ± 0.1	54.3 ± 0.2	41.4 ± 0.3	56.5 ± 0.2	66.3 ± 0.3	71.8 ± 0.5	73.3 ± 0.3	71.8 ± 0.1
DBM+FBCL (Ours)	64.9 ± 0.2	54.5 ± 0.2	42.3 ± 0.3	56.9 ± 0.2	66.5 ± 0.2	72.4 ± 0.4	73.8 ± 0.2	72.3 ± 0.2

Table 11: Statistics over three runs of structural balancing combined with ProCo on ImageNet-LT and iNaturalist.

Method	ImageNet-LT				iNaturalist			
	Many	Medium	Few	All	Many	Medium	Few	All
ProCo	66.5 ± 0.1	54.8 ± 0.5	38.1 ± 0.4	57.1 ± 0.2	70.4 ± 0.4	73.1 ± 0.1	71.9 ± 0.6	72.3 ± 0.2
ProCo + SB (Ours)	66.8 ± 0.2	55.2 ± 0.1	38.8 ± 0.2	57.4 ± 0.1	70.7 ± 0.6	73.4 ± 0.4	72.3 ± 0.4	72.7 ± 0.5

We additionally report results on CIFAR-100-LT, with imbalance ratios in $\{200, 100, 50, 10\}$, for Son et al. (2025) and Wang et al. (2024) in Tables 12 and 13. In the official implementation of Wang et al. (2024), code for ImageNet-LT or iNaturalist were not given. We will add the results when the codes are available. Note that FBCL corresponds to BCL+SB.

Method	CIFAR-100-LT		
	IR=100	IR=50	IR=10
BCL	51.7 ± 0.5	55.7 ± 0.4	63.7 ± 0.1
FBCL	52.1 ± 0.2 (+0.4)	56.0 ± 0.2 (+0.3)	64.0 ± 0.1 (+0.3)
DBM + BCL	50.4 ± 1.0	55.6 ± 0.1	65.7 ± 0.2
DBM + FBCL	51.2 ± 0.5 (+0.8)	55.9 ± 0.3 (+0.3)	65.8 ± 0.1 (+0.1)
LLM-AutoDA + BCL	50.6 ± 0.6	55.5 ± 0.2	64.8 ± 0.1
LLM-AutoDA + FBCL	51.3 ± 0.3 (+0.7)	55.9 ± 0.3 (+0.4)	65.2 ± 0.5 (+0.4)

Table 12: Statistics over five runs of FBCL or BCL with other methods on CIFAR-100-LT.

	CIFAR-100-LT		
Imbalance Factor	100	50	10
Top-1 accuracy			
ProCo	51.9 ± 0.2	56.2 ± 0.2	64.6 ± 0.4
ProCo + SB	52.1 ± 0.3	56.3 ± 0.21	64.7 ± 0.1
Few-class accuracy			
ProCo	32.1 ± 0.6	43.2 ± 0.7	-
ProCo + SB	33.9 ± 0.6	44.8 ± 0.2	-

Table 13: Statistics over five runs of structural balancing applied to ProCo (Du et al., 2024) on CIFAR-100-LT dataset. Adding structural balancing improves few class accuracy.

D.6 COSINE SIMILARITY OF THE REPRESENTATION SPACE

As discussed in section 6, the attainability of a regular simplex in practice is beyond the scope of this work, so we instead evaluate the cosine similarities between class centers to analyze the representation space in which the contrastive objective operates, i.e., the output of the contrastive head. On CIFAR-100-LT, Table 14 shows that FBCL consistently yields lower average cosine similarity across different imbalance ratios and classes. Lower cosine similarity indicates greater separation

Table 14: Cosine similarity across class centers on CIFAR-100-LT.

(a) Average cosine similarity across imbalance ratios.

Method	CIFAR-100-LT		
	100 (↓)	50 (↓)	10 (↓)
BCL	0.298±0.003	0.296±0.003	0.294±0.007
FBCL	**0.233**±0.004	**0.240**±0.004	**0.242**±0.006

(b) Average cosine similarity across classes.

Method	CIFAR-100-LT (IR=100)		
	Many (↓)	Medium (↓)	Few (↓)
BCL	0.318±0.002	0.249±0.008	0.331±0.002
FBCL	**0.259**±0.002	**0.205**±0.004	**0.230**±0.007

Table 15: Cosine similarity across class centers on ImageNet-LT.

Method	**ImageNet-LT**			
	Many (↓)	Medium (↓)	Few (↓)	All (↓)
BCL	0.154 ±0.003	0.106 ±0.004	0.104 ±0.005	0.124 ±0.004
FBCL	**0.137** ±0.001	**0.088** ±0.002	**0.079** ±0.005	**0.103** ±0.002

between class centers. Table 14b further shows the difference particularly in few-shot classes which implies that FBCL better separates tail classes than BCL. We also provide results on the large-scale dataset, ImageNet-LT. The results show the same trend, with FBCL consistently yielding lower cosine similarity across many-, medium-, and few-shot classes, as well as overall, indicating a more separated inter-class geometry than BCL.

D.7 Class balanced sampling and other reweighting strategies

One could expect structural balancing to be similar to conventional re-weighting, since both introduce class-dependent weighting, or that class-balanced sampling could provide a similar effect by constructing batches with a more uniform class distribution. However, prior studies have shown that re-balancing through class-balanced sampling or loss re-weighting (classification losses) can impair representation quality (Kang et al., 2020; Zhou et al., 2020). This motivates examining whether similar strategies are effective for contrastive objectives and how structural balancing differs from them.

We therefore compare structural balancing with class-balanced sampling and representative re-weighting strategies. Class-balanced sampling (*CB-sampling*) constructs batches with a more uniform class distribution, but changes the original data distribution and can degrade performance (Kang et al., 2020). *Re-weighting* applies a weight determined by the class prior to each loss $\frac{N}{KN_y}$, while class-balanced weighting (*CB-weighting*) uses the effective number of samples, $\frac{1-\beta}{1-\beta^{N_y}}$ (Cui et al., 2019), with $\beta = 0.99$. (SCL with CB-weighting corresponds to Zhong et al. (2022)). *RCL* (De Alvis et al., 2025) instead introduces balancing based on class priors within the logarithmic term of the contrastive loss.

Table 16: Comparison with class-balanced sampling and other re-weighting methods. Statistics over five runs are provided.

Method	**CIFAR-100-LT (IR=100)**			
	Many	Medium	Few	Overall
BCL	66.04±0.55	52.57±0.49	33.78±1.04	51.65±0.45
BCL + CB Sampling	51.32±0.25	56.63±0.38	30.17±1.60	46.84±0.39
BCL + Re-weighting	65.51±1.01	52.69±0.54	33.56±0.98	51.45±0.45
SCL + CB-weighting	64.84±0.80	51.44±0.52	33.14±0.64	50.65±0.27
BCL + CB-weighting	64.77±0.59	51.46±0.85	33.02±0.03	50.59±0.49
RCL	66.05±0.29	52.01±0.76	33.97±0.38	51.52±0.40
BCL + SB	66.47±0.50	53.28±0.42	35.05±0.60	52.14±0.15

As shown in Table 16, class-balanced sampling yields the lowest overall accuracy, and interestingly, CB-weighting performs worse than conventional re-weighting, whereas structural balancing achieves the best performance. These results show that simply applying existing re-weighting strate-

gies to a contrastive objective is not sufficient, as they do not provide the theoretical guarantee established for structural balancing. Note that in the full-batch setting, both are same, but such a setting is not practical for large-scale datasets, and in a mini-batch setting when re-weighting does not exactly compensate for the class imbalance, the resulting objective retains dependency on class frequency. In addition, **re-weighting does not consistently improve performance as the imbalance becomes more severe**. At IR=200, BCL+re-weighting achieves an overall accuracy of 46.35 ± 0.34, which is lower than both BCL (47.63 ± 0.30) and FBCL (48.18 ± 0.32). This suggests that conventional re-weighting is not always beneficial for contrastive objectives and can even degrade performance depending on class imbalance. Although structural balancing takes a similar form, it is derived from the class frequency dependence arising from the aggregation of contrastive losses. Our analysis shows that it removes this dependence in the lower bound and recovers the optimal representation geometry. Also, while re-weighting has been reported to impair representation learning when applied to classification losses, our results show that *a re-weighting mechanism derived from the structure of a contrastive objective with a theoretical guarantee can instead be beneficial*, motivating further re-weighting strategies specifically designed for contrastive learning.