NEURAL COLLAPSE BY DESIGN: LEARNING CLASS PROTOTYPES ON THE HYPERSPHERE

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural Collapse (NC) describes the global optimum of supervised learning, yet standard cross-entropy (CE) training rarely attains its geometry in practice. This is due to unconstrained radial degrees of freedom: cross-entropy is invariant to joint rescaling of features and weights, leaving radial directions underconstrained thus preventing convergence to a unique geometry. We show that constraining optimization to the unit hypersphere removes this degeneracy and reveals a unifying view of normalized softmax classifier learning (CL) and supervised contrastive learning (SCL) as the same prototype-contrast principle: both optimize angular similarity to class prototypes, using explicit learned weights for normalized softmax and implicit class means for SCL. Despite this shared foundation, existing objectives suffer from small effective negative sets and interference between positive and negative terms, which slows convergence to NC. We address these issues with two objectives: NTCE, which contrasts class prototypes against all batch instances to expand the negative set from K classes to M samples; and NONL, which normalizes only over negatives to decouple intra-class alignment from inter-class repulsion. Theoretically, we prove that SCL already learns an optimal prototype classifier under NC, eliminating the need for post-hoc typically hours-scale linear probing. Empirically, across four benchmarks including ImageNet-1K, our methods surpass CE accuracy, reach $\geq$95% on NC metrics, and match NC structure with substantially fewer iterations. Moreover, SCL with class-mean prototypes matches linear-probing accuracy while requiring no training. These results reframe supervised learning as prototype-based classification on the hypersphere, closing the theory–practice gap while simplifying training and accelerating convergence.

1 INTRODUCTION

Despite theoretical proofs that Neural Collapse (NC) is the global optimum of supervised learning objectives (Lu & Steinerberger, 2022; Zhou et al., 2022a; Graf et al., 2021), standard training with cross-entropy rarely achieves this configuration in practice. This failure is particularly striking because *NC delivers precisely the properties we seek*: when neural networks do approach this geometric configuration, where within-class representations collapse to their means, class means form an equiangular tight frame (ETF), and classifier weights align with these prototypes, they demonstrate improved generalization (Papyan et al., 2020; Bartlett et al., 2017; Neyshabur et al., 2018), adversarial robustness (Fawzi et al., 2016; Ding et al., 2020), enhanced transfer learning (Kornblith et al., 2019; Khosla et al., 2020), and converge toward max-margin classifiers (Soudry et al., 2018) with stronger robustness guarantees (Hein & Andriushchenko, 2017). *If NC is provably optimal and empirically beneficial, why does standard training consistently fail to achieve it?*

We identify the core issue as *unconstrained radial degrees of freedom*. Cross-entropy optimization allows features and weights to be jointly rescaled without changing predictions (Soudry et al., 2018). This leaves radial directions underconstrained, preventing convergence to a unique geometry. While explicit regularization of features, weights, and biases may theoretically resolve this (Zhu et al., 2021), it introduces multiple hyperparameters that complicate practical implementation. A more principled solution is to eliminate radial freedom entirely by constraining optimization to the unit hypersphere, where NC becomes the *unique* global optimum (Yaras et al., 2022).

This geometric perspective reveals a surprising *unity* between two learning paradigms traditionally viewed as fundamentally different. Classifier learning (CL) with normalized softmax (Wang et al., 2017) has been understood as directly learning decision boundaries through weight vectors, while supervised contrastive learning (SCL) (Khosla et al., 2020) has been viewed as learning representations through instance-to-instance comparisons followed by a separate classifier training phase. We show both are actually *prototype-contrast methods on the hypersphere*: normalized softmax optimizes angular similarity between normalized features and explicit class weight vectors serving as prototypes, while SCL optimizes angular similarity among normalized instances using implicit class-mean embeddings as prototypes.

Despite this shared geometric foundation, these methods inherit computational limitations that prevent efficient NC convergence. CL suffers from a small effective negative set (denominator contrasts against only K class weights) (He et al., 2020), while both paradigms couple positive and negative similarity terms through shared normalization (Yeh et al., 2022), creating interference that slows convergence to optimal geometry. These limitations suggest that achieving NC requires not just hyperspherical constraints but also algorithmic innovations in *how prototypes are contrasted*.

Building on the insight that both paradigms are prototype-based but computationally limited, we make four key contributions:

1. We **unify normalized softmax and SCL** under a single geometric framework, revealing both as *prototype-contrast methods on the unit hypersphere* that differ only in whether prototypes are explicit (learned weights) or implicit (class means). This framework explains why both can achieve NC while standard cross-entropy cannot.
2. We propose **two supervised objectives** that overcome existing computational limitations. **NTCE** (Normalized Temperature-scaled Cross Entropy) increases the effective number of negatives from K classes to M batch samples by contrasting the class prototype against all instances in the batch, strengthening inter-class separation. **NONL** (Negatives-Only Normalization Loss) eliminates interference between intra-class alignment and inter-class repulsion by normalizing only over negatives, accelerating NC convergence.
3. We prove that *SCL already learns an optimal classifier during pretraining*, **eliminating the need for linear probing**. The class-mean embeddings learned by SCL form an ETF-aligned prototype classifier under NC, implementing the self-duality condition by construction and yielding equivalent accuracy without incurring the computational cost of post-training probing.
4. We validate our approach across four benchmarks including ImageNet-1K. NTCE and NONL achieve $\geq 95\%$ *on NC metrics* while *surpassing standard cross-entropy accuracy*, and match cross-entropy's NC metrics with *substantially fewer training iterations*. Our prototype classifier maintains SCL's accuracy while eliminating hours of linear probing computation, a significant *practical saving for large-scale deployments*.

These results suggest a **fundamental shift** in how supervised learning should be understood: not as unconstrained optimization in Euclidean space, but as *prototype-based classification on the hypersphere*. By making this geometry explicit, we close the theory–practice gap, simplify training, accelerate convergence, and yield interpretable models that provably realize their optimal NC structure. The *practical impact is substantial*: faster training, elimination of extra compute phases, and models that reach the theoretical optimum. The *theoretical insight* provides a principled foundation for future advances in supervised learning.

2 Related Work

Neural Collapse. Neural Collapse (NC) describes a limiting geometry in which within-class features collapse to their means (NC1), class means form a centered simplex ETF (NC2), classifier weights align with the means (NC3), and biases collapse (NC4) (Papyan et al., 2020). Variants of this structure characterize global minimizers for several objectives and modeling assumptions, including MSE (Han et al., 2022; Zhou et al., 2022a), cross-entropy (CE) (Lu & Steinerberger, 2022), supervised contrastive learning (SCL) (Graf et al., 2021), and CE variants such as label smoothing and focal loss (Zhou et al., 2022b). In finite training, however, standard CE with weight decay often

fails to realize the optimal geometry: the loss is *scale-noncoercive* and can be driven toward zero by inflating logit magnitudes without improving angular structure (Albert & Anderson, 1984; Soudry et al., 2018). Class imbalance further distorts the ETF and slows convergence (Thrampoulidis et al., 2022; Hong & Ling, 2024); free bias terms obstruct NC4 and can exacerbate miscalibration unless controlled (e.g., logit adjustment) (Menon et al., 2021). While simultaneously penalizing features, weights, and biases can restore coercivity and yield NC in principle (Zhu et al., 2021; Zhou et al., 2022a), tuning multiple regularizers is brittle. *We show that contrasting instances against class prototypes on the hypersphere operationalizes NC in practice.*

Learning on the hypersphere. Constraining radial freedom is a principled route to NC. When both features and classifier lie on the unit hypersphere, CE over the product of spheres exhibits a benign strict-saddle landscape whose minima realize perfect NC (Yaras et al., 2022). Related evidence appears in contrastive objectives: SCL yields within-class collapse and simplex class means (Graf et al., 2021), while in self-supervised contrastive learning batch-level optima form a simplex ETF (Koromilas et al., 2024). A long line of face-recognition work, including SphereFace, CosFace, ArcFace, and NormFace (Liu et al., 2017; Wang et al., 2018; Deng et al., 2019; Wang et al., 2017), operationalizes direction-only discrimination by using angular/cosine margins. *We unify these appraches by showing that both normalized softmax and SCL perform prototype contrast on the hypersphere.* Building on this bridge, we extend normalized softmax with NTCE/NONL to import desirable properties.

Prototype-based classification and ETF classifiers. Prototype methods classify via distances to learned representatives (Snell et al., 2017). Motivated by NC, prior work fixes the classifier to a simplex ETF and learns only the encoder (Yang et al., 2022), enforces (non-negative) orthogonality (Kim & Kim, 2024), or guides the classifier toward the nearest ETF via a Riemannian inner optimization (Markou et al., 2024). *Our perspective is that CL and SCL already operate with prototypes: we modify the objectives to realize NC in practice, and we show that SCL's class-mean prototypes form an effective classifier, making linear probing unnecessary.*

3 PRELIMINARIES

Notation. Scalars are denoted by lowercase letters u, vectors by lowercase bold letters $\boldsymbol{u}$, and matrices by uppercase bold letters $\boldsymbol{U}$. Sets are represented by uppercase caligraphic letters $\mathcal{U}$. Individual elements are accessed using subscript notation: u_i for the i-th element of vector $\boldsymbol{u}$ and $U_{i,j}$ for the element at row i and column j of matrix $\boldsymbol{U}$. To denote vertical (row-wise) concatenation of matrices $\mathbf{X}$ and $\mathbf{Y}$, we use $[\mathbf{X}; \mathbf{Y}]$. We denote normalized vectors with $\hat{\boldsymbol{u}}_j = \boldsymbol{u}_j / \|\boldsymbol{u}_j\|$.

3.1 LEARNING PARADIGMS

Classifier Learning with Cross-Entropy. The cross-entropy loss is the standard Classifier Learning (CL) objective, optimizing representations and classifier weights simultaneously. An encoder $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Z}$, parameterized by $\boldsymbol{\theta} \in \Theta$, maps an input $\mathbf{x} \in \mathcal{X}$ to its representation $\mathbf{z} = f_{\boldsymbol{\theta}}(\mathbf{x}) \in \mathcal{Z}$. For a K-class task, y_i denotes the class assignment of sample $\mathbf{x}_i$. A linear classifier is placed on top of the encoder, with weight matrix $\mathbf{W} \in \mathbb{R}^{K \times h}$ and bias $\mathbf{b} \in \mathbb{R}^K$, where h is the embedding dimension. For a mini-batch of M samples with $\{\mathbf{z}_i\}_{i=1}^M$, the cross-entropy loss is defined as

$$\mathcal{L}_{\text{CE}}(\mathbf{Z}, \mathbf{W}) = \frac{1}{M} \sum_{i=1}^{M} -\log \left(\frac{e^{\mathbf{z}_i^\top \mathbf{w}_{y_i} + b_{y_i}}}{\sum_{j=1}^{K} e^{\mathbf{z}_i^\top \mathbf{w}_j + b_j}} \right), \tag{1}$$

where $\mathbf{w}_j$ denotes the j-th row of $\mathbf{W}$ and b_j the j-th component of $\mathbf{b}$.

Supervised Contrastive Learning. *Supervised Contrastive Learning (SCL)* takes a seemingly different direction: it learns representations by exploiting similarities between instances to learn class-invariant representations. Building on our notation, the contrastive framework augments the encoder $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Z}$ with a projection head $g_{\boldsymbol{\phi}} : \mathcal{Z} \to \mathcal{U}$, parameterized by $\boldsymbol{\phi} \in \Phi$, which maps representations onto the unit hypersphere, $\mathcal{U} = \mathbb{S}^{d-1} = \{\boldsymbol{u} \in \mathbb{R}^d \mid \|\boldsymbol{u}\| = 1\}$. We denote the projected representations as $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{U}$, where $\boldsymbol{u}_i$ comes from instance $\boldsymbol{x}_i$ and $\boldsymbol{v}_i$ from its alternative view produced via augmentation, a typical process in contrastive learning.

For SCL the objective is to pull together positive pairs while pushing apart negative pairs in the projection space. Typically alternative views of the same data point that originate from augmentation are considered as new data points, *i.e.* $\boldsymbol{A} = [\boldsymbol{U}; \boldsymbol{V}]$, and the supervised contrastive loss becomes:

$$\mathcal{L}_{\text{SCL}}(\boldsymbol{A}) = \frac{1}{2M} \sum_{i=1}^{2M} -\frac{1}{|\mathcal{C}(i)|} \sum_{l \in \mathcal{C}(i)} \log \left(\frac{e^{\boldsymbol{a}_i^\top \boldsymbol{a}_l / \tau}}{\sum_{\substack{j=1 \\ j \neq i}}^{2M} e^{\boldsymbol{a}_i^\top \boldsymbol{a}_j / \tau}} \right), \tag{2}$$

where $\mathcal{C}(i)$ denotes the set of indices corresponding to positive examples sharing the same class as $\boldsymbol{x}_i$ and $\tau > 0$ is a temperature parameter that controls the concentration of the distribution.

A crucial distinction emerges post-training: while learning with cross-entropy directly produces a classifier, contrastive learning requires an additional step. After optimizing Equation (2), the projection head is discarded and a linear classifier $\boldsymbol{W}, \boldsymbol{b}$ is trained on the frozen encoder representations $\boldsymbol{z}$ using Equation (1), a process known as **linear probing**.

3.2 Neural Collapse (NC).

Neural Collapse Papyan et al. (2020)is the late-training regime (on balanced data) where last-layer features and the linear classifier converge to a highly structured limit. Let $\boldsymbol{z}_i = f(\boldsymbol{x}_i) \in \mathbb{R}^h$, class means $\boldsymbol{\mu}_c = \frac{1}{n_c} \sum_{i:y_i=c} \boldsymbol{z}_i$, weights $\boldsymbol{w}_c$, and bias $\boldsymbol{b}$. NC holds when, up to common scalings:

(NC1) **Within-class collapse:** $\boldsymbol{z}_i = \boldsymbol{\mu}_{y_i}$ for all i.

(NC2) **Simplex ETF of class means:** the centered means $\tilde{\boldsymbol{\mu}}_c = \boldsymbol{\mu}_c - \frac{1}{K} \sum_{k=1}^{K} \boldsymbol{\mu}_k$ have equal norms and equal pairwise angles so the means span a centered $(K-1)$-simplex ETF.

(NC3) **Alignment of Class Representation and Classifier:** classifier columns align with the class means, $\boldsymbol{w}_c \parallel \boldsymbol{\mu}_c$ (there exists $\gamma > 0$ with $\boldsymbol{w}_c = \gamma \, \boldsymbol{\mu}_c$).

(NC4) **Bias collapse:** $\boldsymbol{b} = \beta \, \mathbf{1}$ for some scalar β.

Under NC, the decision rule reduces to nearest-class-mean classification. We assume balanced classes and $h \geq K - 1$ so a centered simplex ETF is feasible (Lu & Steinerberger, 2022).

Practical Challenges in reaching Neural Collapse Neural Collapse (NC) is now well documented in deep nets (Papyan et al., 2020) and characterizes global minima of balanced cross-entropy (Lu & Steinerberger, 2022). However standard pipelines does not enforce it in practice. For the typical paradigm of cross-entropy and classifier weight decay, the objective admits an *unbounded rescaling direction*: shrinking the classifier while amplifying features leaves logits unchanged, reduces the penalty, and drives the loss toward zero without achieving NC (Soudry et al., 2018; Albert & Anderson, 1984). It is shown by Zhu et al. (2021) that a well-posed objective arises when all radial degrees of freedom are constrained by penalizing weights, features, and biases simultaneously (Zhu et al., 2021). However this is practically brittle due to multiple regularizers to tune.

Supervised contrastive training on the other hand can drive representations toward NC geometry (Graf et al., 2021). However, the subsequent *linear probing* step typically fits a softmax classifier with cross-entropy on *frozen* features, allowing free weight magnitudes and biases. This reintroduces the same scale and bias pathologies as cross-entropy even when training has already reached an NC.

4 Supervised Learning on the Hypersphere

In this section we present a common view-point bridging classifier learning and contrastive learning to accelerate neural collapse. Our approach leverages similarity-based optimization while eliminating radial degrees of freedom by constraining both feature and classifier norms to the hypersphere. This constraint transforms the optimization landscape into a benign geometry where all critical points become global optima (Yaras et al., 2022), enabling direct convergence to NC.

4.1 Revisiting Cross Entropy: Contrasting class prototypes to instances

The weight matrix of the final linear classifier in CL methods can be expressed as $\boldsymbol{W} = [\boldsymbol{w}_1; \boldsymbol{w}_2; \ldots; \boldsymbol{w}_K] \in \mathbb{R}^{K \times h}$, where each $\boldsymbol{w}_c$ represents a learnable class prototype. This formulation reveals an important insight: we can treat the classifier weights as *learnable prototypes* that

evolve through gradient descent to capture class-specific geometric structures. Building on this we design objectives that leverage such prototypes to help arrive at the optimal NC geometry.

Normalized Softmax Losses. Standard cross-entropy and contrastive learning represent two seemingly distinct paradigms: the former discriminates through learned magnitudes and biases in unconstrained space, while the latter operates purely on angular similarities on the hypersphere. This fundamental difference leads to a critical inefficiency: while both methods theoretically converge to neural collapse configurations, cross-entropy introduces unnecessary radial degrees of freedom that slow convergence to this optimal geometry (Yaras et al., 2022; Zhu et al., 2021).

Normalized softmax losses resolve this inefficiency by reformulating cross-entropy as a pure geometric objective. NormFace (Wang et al., 2017), a prominent example, achieves this through three coordinated modifications: (i) eliminating biases that merely translate decision boundaries without encoding semantic structure, (ii) projecting representations onto the hypersphere to focus exclusively on angular geometry, and (iii) introducing temperature scaling to control concentration of the softmax distribution. Formally, with $\boldsymbol{u}_i = \boldsymbol{z}_i/\|\boldsymbol{z}_i\|_2$ as the normalized representation and $\hat{\boldsymbol{w}}_j = \boldsymbol{w}_j/\|\boldsymbol{w}_j\|_2$ as the normalized classifier weight for class j, NormFace minimizes:

$$L_{\text{NormFace}}(\boldsymbol{U}, \boldsymbol{W}) = -\frac{1}{M}\sum_{i=1}^{M} \log\left(\frac{e^{\boldsymbol{u}_i^\top \hat{\boldsymbol{w}}_{y_i}/\tau}}{\sum_{j=1}^{K} e^{\boldsymbol{u}_i^\top \hat{\boldsymbol{w}}_j/\tau}}\right). \tag{3}$$

This reformulation transforms classification into contrastive learning between data instances and learnable class prototypes while maintaining cross-entropy's computational efficiency.

Normalized Temperature-scaled Cross Entropy (NTCE)

When utilizing NormFace to view CL from a contrastive learning perspective we end up with an inherent limitation of cross entropy: the number of negatives in the objective is limited to K, the number of class prototypes. It is very well investigated that contrastive objectives need very large numbers of negatives in order to converge (He et al., 2020). This is mostly due to the fact that fewer negatives provide a worse estimate to the expectation of the actual contrastive objective (Koromilas et al., 2024).

By inverting the contrastive direction from instance-to-class to class-to-instance discrimination we address this limitation through the Normalized Temperature-scaled Cross Entropy (NTCE). This modification fundamentally alters the learning dynamics: rather than each instance contrasting against K class prototypes, each class prototype now contrasts against M batch representations.

The key insight underlying NTCE is that *class prototypes themselves can serve as anchors* in the contrastive formulation. By anchoring on the class weight vector corresponding to each instance's ground-truth label and contrasting it against all batch representations, we dramatically expand the negative sampling space. Formally, NTCE takes the form:

$$L_{\text{NTCE}}(\boldsymbol{U}, \boldsymbol{W}) = \frac{1}{M}\sum_{i=1}^{M} -\log\left(\frac{e^{\hat{\boldsymbol{w}}_{y_i}^\top \boldsymbol{u}_i/\tau}}{\sum_{j=1}^{M} e^{\hat{\boldsymbol{w}}_{y_i}^\top \boldsymbol{u}_j/\tau}}\right), \tag{4}$$

where $\hat{\boldsymbol{w}}_{y_i}$ serves as the anchor for instance i, and critically, the denominator sums over all M instances in the batch rather than over K classes.

Negatives Only Normalization Loss. NTCE adds enhanced negative sampling on top of NormFace to directly transfer the principles of contrastive learning to cross entropy training. However, it also brings a fundamental drawback of popular contrastive objectives that compromises its optimization dynamics. The denominator in Equation (4) indiscriminately aggregates all instances sharing the same class anchor. That is the denominator, also known as the uniformity term, is optimized when instances of the same class have maximum distance (Wang & Isola, 2020), which contradicts the optimality of the numerator (alignment terms). More specifically, positive pairs explicitly appear as negative samples in the normalization term, generating gradients that actively repel instances from their own class prototype. When instance i and instance j share class $y_i = y_j$, the term $e^{\hat{\boldsymbol{w}}_{y_i}^\top \boldsymbol{u}_j/\tau}$ in the denominator produces gradients that decrease $\hat{\boldsymbol{w}}_{y_i}^\top \boldsymbol{u}_j$, directly opposing the alignment objective. This is a known behavior that is called *alignment-uniformity coupling* (Yeh et al., 2022).

In order to resolve this conflict we introduce the Negatives-Only Normalization Loss (NONL), which explicitly excludes same-class instances from the denominator:

$$L_{\text{NONL}}(\boldsymbol{U}, \boldsymbol{W}) = \frac{1}{M} \sum_{i=1}^{M} -\log \left(\frac{e^{\hat{\boldsymbol{w}}_{y_i}^\top \boldsymbol{u}_i / \tau}}{\sum_{\substack{j=1 \\ j \notin \mathcal{C}(i)}}^{M} e^{\hat{\boldsymbol{w}}_{y_i}^\top \boldsymbol{u}_j / \tau}} \right). \tag{5}$$

4.2 Revisiting Supervised Contrastive Learning: Contrasting Mean-Class Prototypes to Instances

SCL implicitly learns prototype classifiers. We follow Equation (2) to treat alternative views produced by data augmentation as distinct samples, i.e., $\boldsymbol{A} = [\boldsymbol{U}; \boldsymbol{V}]$. Let $\mathcal{B}_c = \{j \in [2M] : y_j = c\}$ denote the within-batch index set for class c, $n_c = |\mathcal{B}_c|$, and $\hat{\boldsymbol{\mu}}_c = \frac{1}{n_c} \sum_{j \in \mathcal{B}_c} \boldsymbol{a}_j$ the corresponding batch prototype (class mean). We define the *prototype loss*:

$$L_{\text{proto}}(\boldsymbol{A}) = -\frac{1}{2M} \sum_{i=1}^{2M} \log \left(\frac{e^{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_{y_i} / \tau}}{-e^{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_{y_i} / \tau} + \sum_{c=1}^{K} n_c \cdot e^{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_c / \tau}} \right), \tag{6}$$

where the numerator encourages alignment with the correct class prototype, while the denominator includes both positive and negative prototypes weighted by their batch frequencies n_c. Theorem 4.1 connects the optima of this loss to the ones of SCL. The proof can be found in Appendix A.1.

Theorem 4.1 (Equivalence of SCL and prototype–softmax minimizers)**.** *For unit-norm representations and balanced labels the supervised contrastive loss L_{SCL} and the prototype loss L_{proto} in Equation* (6) *share the same set of global minimizers (up to rotation and label permutation). In particular, at every global minimizer the representations exhibit in-class collapse and the class means form a centered simplex ETF.*

This result clarifies our understanding of SCL: rather than merely learning good representations for classification, *SCL directly optimizes for classifier-feature alignment* through its contrastive objective. The learned prototypes are not just byproducts but the optimal classifiers themselves.

Connection to Classifier Learning. The n_c weighting in the denominator of Equation (6) captures the effect of utilizing multiple negative instances, matching the structure of Equation (4). When discarding the n_c weights, this loss reduces to Equation (3), establishing a direct **correspondence between the prototype weights and class means**. Adding that the optimal solution of Equation (3) holds when $\boldsymbol{w}_c = \hat{\boldsymbol{\mu}}_c$ (Yaras et al., 2022) the connection becomes even more prevalent.

In other words, despite SCL converging to collapsed class representations forming an ETF, its optima can also be attained by contrasting instances to class-mean prototypes. This connects CL techniques to SCL, where the learnable classifier weights in the former are free parameters while in the latter they emerge implicitly from the learned representations.

Why linear probing fails for SCL features. In practice linear probing is used to train a classifier for the learned SCL representations. This approach introduces **unnecessary degrees of freedom** that disrupt the geometric optimality achieved by SCL. Specifically this process introduces: (i) **geometric mismatch**: SCL features live on the hypersphere with collapsed, ETF-structured class means. Linear probing operates in unconstrained Euclidean space, allowing weight rescaling and bias shifts that break classifier-feature alignment (Soudry et al., 2018). (ii) **Redundancy:** Our theorem shows SCL has already learned optimal classifier weights, *i.e.* the class prototypes themselves.

Class-mean Prototypes inplace of Linear Probing. We observe that class prototypes $\hat{\boldsymbol{\mu}}_c$ serve as natural classifier weights that satisfy NC3 (classifier-feature alignment) by construction. Rather than retrofitting a linear head to pre-collapsed features, we directly impose the NC-optimal classifier from the learned geometry. We *discard linear probing entirely* and set the classifier weights to the learned prototypes: $\boldsymbol{w}_c = \hat{\boldsymbol{\mu}}_c$. Doing so, we alleviate the need for an extra training phase, and we show empirically (Section 5) that this prototype-based classification matches linear probing performance.

5 Experiments

In this section we empirically validate our methods against cross-entropy (CE) and NormFace(Wang et al., 2017) for Classifier Learning paradigms and supervised contrastive learning (SCL), evaluat-

Table 1: Performance comparison of learning paradigms and objectives across datasets. **Bold**: best within method family; **green**: overall best per dataset.

(I) Classifier Learning Methods

Loss	CIFAR-10	CIFAR-100	ImageNet-100	ImageNet-1K
CE	94.6	72.1	84.4	75.4
NORMFACE	94.8	72.4	84.4	75.6
NTCE (ours)	94.7	72.9	84.7	**76.0**
NONL (ours)	**94.9**	**73.6**	**84.9**	75.0

(II) Supervised Contrastive Learning Methods

Classifier Learning	Loss	Forward Passes	CIFAR-10	CIFAR-100	ImageNet-100	ImageNet-1K
LINEAR PROBING	SCL	$T \times N$	**95.0**	**73.9**	84.8	**75.1**
NORMALIZED LINEAR PROBING	SCL	$T \times N$	94.9	73.6	84.8	**75.1**
FIXED PROTOTYPES	SCL	N	**95.0**	**73.9**	**86.8**	**75.1**

ing: (i) classification accuracy, (ii) proximity to neural collapse geometry, and (iii) NC convergence speed. Experiments are conducted on four standard datasets: *CIFAR10, CIFAR100, ImageNet-100, and ImageNet1K*, following common representation learning benchmarking practices (Khosla et al., 2020; Markou et al., 2024; Wang et al., 2021; Yeh et al., 2022). We use ResNet50 for ImageNet datasets and ResNet18 for CIFAR. Implementation details are provided in Appendix A.2.

5.1 CLASSIFICATION PERFORMANCE

Classifier Learning Methods. As can be inferred from Table 1(I), normalized losses *outperform cross-entropy* (CE) in 11 out of 12 cases, while our losses outperform NormFace in 7/8 cases. NONL achieves the strongest gains on datasets with few (10) to medium (100) number of classes but underperforms on ImageNet-1K. We hypothesize this degradation stems from a fundamental gradient imbalance under uniform bacth sampling: with batch size 2048 and K=1000 classes, a large number of classes are absent per batch in expectation. For these missing classes, NONL produces exclusively negative gradients, which means their weights $\boldsymbol{w}_j$ appear in all other samples' normalization terms but receive no positive signal from their own instances. In such cases our NTCE circumvents this limitation through normalizing over the batch, achieving the best ImageNet-1K accuracy (76.0%).

Supervised Contrastive Learning Methods. The accuracy from three classifier learning strategies on SCL representations is presented in Table 1(II): (i) standard linear probing with learnable weights and bias, (ii) normalized linear probing using NormFace loss, and (iii) fixed prototypes computed as class-mean embeddings. Fixed prototypes match linear probing performance on 3 of 4 datasets, and mark a considerable +2.0% improvement on ImageNet-100 **requiring only N forward passes versus $T \times N$** for training-based methods, where T is the number of epochs. Normalized linear probing achieves comparable accuracy to standard linear probing, validating that the *discriminative information in SCL features resides primarily in their angular structure* rather than magnitude or biases. These findings validate that angular structure alone suffices for discrimination in well-trained representations, enabling *training-free classification* in SCL via fixed prototypes that **eliminate huge computational costs** by discrarding a, typically hours long, training phase.

5.2 QUANTIFYING NEURAL COLLAPSE

We quantify NC1–NC3 with complementary, condition-specific metrics; we omit NC4 (bias collapse) as our models enforce zero bias by design.

Effective Rank (NC1, NC2). For matrix $\mathbf{A}$ with singular values $\{\sigma_i\}$ the effective rank (Roy & Vetterli, 2007) is defined as $\text{erank}(\mathbf{A}) = \exp\{-\sum_i p_i \log p_i\}$ where $p_i = \sigma_i / \sum_j \sigma_j$. We compute the intra and inter class effective ranks (Zhang et al., 2024) as: $\text{erank}_{\text{intra}} = \frac{1}{K}\sum_{c=1}^{K} \text{erank}(\text{Cov}[\mathbf{z}_i - \mu_c \mid y_i = c])$ and $\text{erank}_{\text{inter}} = \text{erank}(\text{Cov}[\mu_c - \mu_G])$, where Cov is the covariance matrix. These metrics quantify **NC1** (within-class variability collapse): $\text{erank}_{\text{intra}} \to 0$ indicates $\boldsymbol{z}_i \to \mu_{y_i}$, and

Table 2: NC metrics on CIFAR-10/100 (training). **Bold** marks the best within each learning family; **green** marks the overall best per dataset. Theoretical optima: Intra ER $0/0$, Inter ER $9/99$, Weights ER $9/99$, Weight Align $0/0$, Instance Align $0/0$, MIR $1/1$, HDR $0/0$.

Learning Family	Method	Effective Rank			Alignment		Information Theory Metrics	
		Intra ↓	Inter ↑	Weights ↑	Weight ↓	Instance ↓	MIR ↑	HDR ↓
Classifier Learning	CE	22.5 / 96.4	8.6 / 57.1	8.9 / 89.7	0.59 / 0.83	0.69 / 1.05	**0.98** / 0.97	**0.03** / **0.13**
	NormFace	10.5 / 13.6	**9.0** / 96.2	**9.0** / 96.1	0.12 / **0.01**	0.14 / 0.06	0.95 / **1.00**	0.04 / 0.30
	NTCE	9.0 / 12.6	**9.0** / **99.0**	8.9 / 98.9	**0.08** / **0.01**	**0.10** / **0.05**	0.96 / **1.00**	0.05 / 0.30
	NONL	**4.0** / **11.4**	**9.0** / **99.0**	**9.0** / **99.0**	0.11 / **0.01**	0.16 / 0.06	0.95 / **1.00**	0.05 / 0.30
Contrastive Learning	SCL (w probing)	**4.5** / **7.5**	**9.0** / **66.7**	8.3 / **77.8**	0.99 / 1.03	**0.10** / **0.34**	0.99 / **0.95**	**0.07** / **0.11**
	SCL (w/o probing)	**4.5** / **7.5**	**9.0** / **66.7**	**9.0** / 66.5	**0.00** / **0.00**	**0.10** / **0.34**	**1.00** / 0.87	0.09 / 0.14

Table 3: Convergence speed (% of training iters): **(I)** time to reach the 95% NC threshold; **(II)** time to match CE's final value; "0%" indicates the target is met at the first logged eval.

Method	Instance alignment	Weight alignment	Weights erank	Intra erank	Inter erank
(I) NC convergence to 95% threshold (ratio to max iterations)					
NormFace	79.4%	8.2%	52.6%	45.4%	56.2%
NTCE	79.4%	6.8%	56.4%	36.6%	52.4%
NONL	79.4%	7.4%	34.6%	14.6%	47.2%
(II) CE convergence to converged value (ratio to CE converged iteration)					
NormFace	2.2%	2.0%	66.3%	0%	7.4%
NTCE	2.2%	1.8%	73.9%	0%	7.4%
NONL	2.2%	1.8%	35.4%	0%	6.0%

NC2 (ETF structure): Zhang et al. (2024) proved that when $\text{erank}_{\text{inter}} = K - 1$ the class means form a simplex with equal pairwise angles. We also report $\text{erank}(\mathbf{W})$ to assess whether classifier weights approximate an equiangular tight frame (ETF).

Alignment (NC3). We quantify feature–classifier alignment by $\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{z}_i - \mathbf{w}_{y_i}\|_2^2$ and also report instance-to-instance alignment to probe per-class collapse.

Information Metrics (NC2, NC3). For normalized Gram matrices $\mathbf{G}_W$ (weights), $\mathbf{G}_M$ (class means) and H being the matrix entropy, Song et al. (2024) connects Neural Collapse to the metrics:

$$\text{MIR} = \frac{\text{H}(\mathbf{G}_W) + \text{H}(\mathbf{G}_M) - \text{H}(\mathbf{G}_W \odot \mathbf{G}_M)}{\min\{\text{H}(\mathbf{G}_W), \text{H}(\mathbf{G}_M)\}}, \qquad \text{HDR} = \frac{|\text{H}(\mathbf{G}_W) - \text{H}(\mathbf{G}_M)|}{\max\{\text{H}(\mathbf{G}_W), \text{H}(\mathbf{G}_M)\}} \tag{7}$$

These capture the information-theoretic signatures of **NC2** and **NC3** where under full collapse MIR $\to 1$ and HDR $\to 0$, reflecting perfect structural alignment.

In Table 2 four key findings are revealed: **(i) CE fails to achieve NC:** high intra-class variance (erank 22.5/96.4), suboptimal inter-class separation (erank 8.6/57.1 vs. theoretical K-1=9/99), and poor weight-feature alignment (w-inst 0.59/0.83, inst-inst 0.69/1.05). **(ii) Normalized softmax losses satisfy NC2-NC3** since they achieve perfect inter-class separation (erank 9.0/99.0), near-zero alignment errors (NTCE: w-inst 0.08/0.01, inst-inst 0.10/0.05), and optimal weight dimensionality matching the simplex ETF, with **NONL being the overall best** mostly due to its better intra class structure. **(iii) SCL with linear probing violates NC3:** despite superior within-class collapse (erank 4.5/7.5), inter-class structure degrades (erank 9.0/66.7) and classifier-feature alignment fails (w-inst 0.99/1.03). **(iv) Fixed prototypes restore NC3 in SCL:** removing the trainable classifier enforces perfect alignment by construction, though inter-class separation remains suboptimal.

Convergence Dynamics. On CIFAR-100, we track NC metrics and define convergence as the earliest iteration where the exponentially-weighted moving average enters and remains within a metric-specific tolerance around the 95% NC threshold.

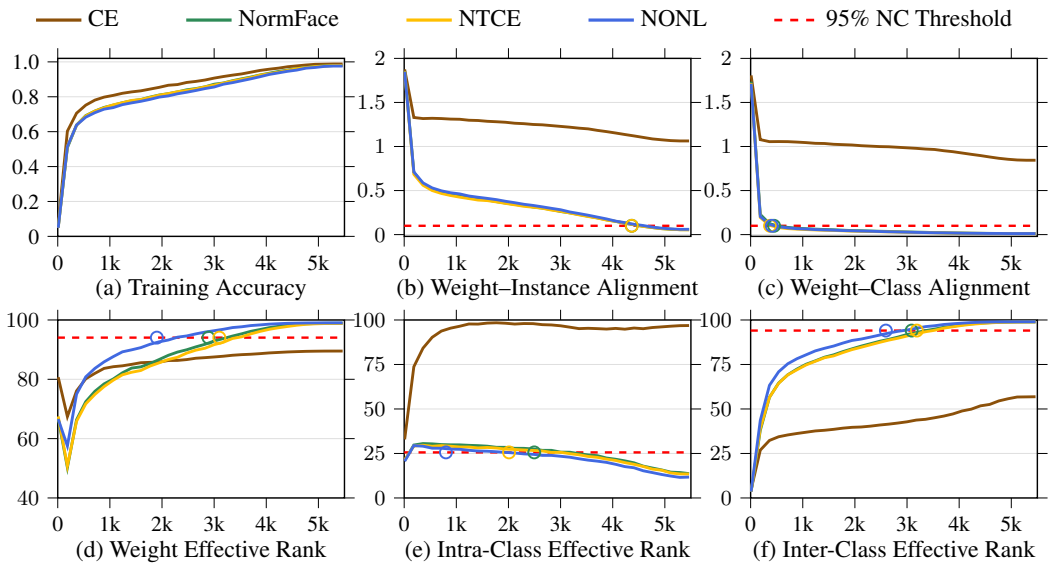


Figure 1: **NC convergence on CIFAR-100.** Six metrics vs. training iterations; red dashed lines mark the 95% NC threshold and circles denote each method's convergence.

In Figure 1 the training dynamics are denomstrated. While cross-entropy (CE) achieves perfect training accuracy, it fails to reach neural collapse geometry, plateauing at suboptimal metric values. CE's accuracy improvements appear to *derive solely from magnitude and bias adjustments* rather than geometric reorganization. In contrast, our methods *simultaneously optimize all NC metrics throughout training*, converging to proper NC geometry while maintaining optimal accuracy.

In Table 3(I) the convergence speed to 95% of theoretical NC thresholds is quantified. Normalized losses reach these thresholds, *typically early in training*. NONL converges faster with **gains over NormFace for the rank metrics** (1.2-3.1 speedup), benefiting from simplified optimization without competing terms. Table 3(II) benchmarks against CE's converged values. The acceleration is dramatic: normalized losses reach CE-equivalent values in under 7.5% of CE's required iterations across 4/5 metrics, while **NONL converges faster**. This demonstrates that normalized losses fundamentally restructure the optimization landscape, *enabling direct paths to neural collapse*.

6 Conclusion

In this work, we address the mismatch between the theoretical optima of supervised objectives and their behavior in practice. Constraining learning to the unit hypersphere removes the radial degeneracy of cross-entropy and unifies normalized softmax and supervised contrastive learning (SCL) as a single prototype-contrast paradigm. Building on this view, we propose two objectives (NTCE and NONL) that accelerate convergence to Neural Collapse. Theoretically, we prove SCL already yields an optimal prototype classifier during contrastive training, eliminating the typical linear probing phase. Empirically, across four benchmarks including ImageNet-1K, our methods surpass CE accuracy, reach $\geq$95% on NC metrics, and attain NC geometry in substantially fewer iterations. Overall, supervised learning is recast as prototype-based classification on the hypersphere, narrowing the theory–practice gap while simplifying and speeding up training.

Reproducibility Statement

Our approach modifies only loss functions within standard pipelines. Results can be replicated by pluging configutrations from Appendix A.2 into popular codebases (*e.g.* `https://github.com/HobbitLong/SupContrast`) with minimal effort substituting the original loss with ours.

REFERENCES

Arthur Albert and John A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6241–6250, 2017.

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

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019.

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HkeryxBtPB.

Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. URL https://papers.neurips.cc/paper/6331-robustness-of-classifiers-from-adversarial-to-random-noise.pdf.

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

X.Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=w1UbdvWH_R3.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. *Advances in neural information processing systems*, 30, 2017.

Yuan Hong and Shuyang Ling. Neural collapse for unconstrained feature model under class-imbalance. *Journal of Machine Learning Research*, 25(180):1–48, 2024. URL https://www.jmlr.org/papers/volume25/23-1215/23-1215.pdf.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://proceedings.neurips.cc/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Hoyong Kim and Kangil Kim. Fixed non-negative orthogonal classifier: Inducing zero-mean neural collapse with feature dimension separation. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/pdf?id=F4bmOrmUwc.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2661–2671, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/papers/Kornblith_Do_Better_ImageNet_Models_Transfer_Better_CVPR_2019_paper.pdf.

Panagiotis Koromilas, Giorgos Bouritsas, Theodoros Giannakopoulos, Mihalis Nicolaou, and Yannis Panagakis. Bridging mini-batch and asymptotic analysis in contrastive learning: From infonce to kernel-based losses. In *International Conference on Machine Learning*, pp. 25276–25301. PMLR, 2024.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6738–6746, 2017.

Jianfeng Lu and Stefan Steinerberger. Neural collapse under cross-entropy loss. *Applied and Computational Harmonic Analysis*, 59:224–241, 2022. doi: 10.1016/j.acha.2021.12.011.

Evan Markou, Thalaiyasingam Ajanthan, and Stephen Gould. Guiding neural collapse: Optimising towards the nearest simplex equiangular tight frame. *Advances in Neural Information Processing Systems*, 37:35544–35573, 2024.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=37nvvqkCo5.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Kun Song, Zhiquan Tan, Bochao Zou, Huimin Ma, and Weiran Huang. Unveiling the dynamics of information interplay in supervised learning. In *International Conference on Machine Learning*, pp. 46156–46167. PMLR, 2024.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70): 1–57, 2018. URL https://jmlr.org/papers/v19/18-188.html.

Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems*, 35:27225–27238, 2022.

Feng Wang, Xiang Xiang, Jian Cheng, and Alan L. Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM International Conference on Multimedia (MM)*, pp. 1041–1049, 2017. doi: 10.1145/3123266.3123359.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5265–5274, 2018.

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

Xudong Wang, Ziwei Liu, and Stella X Yu. Unsupervised feature learning by cross-level instance-group discrimination. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12586–12595, 2021.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/f7f5f501282771c96bb3fedcc96bedfe-Paper-Conference.pdf.

Can Yaras, Peng Wang, Zhihui Zhu, Laura Balzano, and Qing Qu. Neural collapse with normalized features: A geometric analysis over the riemannian manifold. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *European Conference on Computer Vision*, pp. 668–684. Springer, 2022.

Yifan Zhang, Zhiquan Tan, Jingqin Yang, Weiran Huang, and Yang Yuan. Matrix information theory for self-supervised learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 59897–59918, 2024.

Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under mse loss: Global optimality with unconstrained features. In *International Conference on Machine Learning*, pp. 27179–27202. PMLR, 2022a.

Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/cdce17de141c9fba3bdf175a0b721941-Paper-Conference.pdf.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.

A Appendix

A.1 Proof

Here we provide the proof of Theorem 4.1.

Proof. Fix $i \in [2M]$ with label y_i. Let $\mathcal{C}(i) = \{j \in [2M] : j \neq i,\ y_j = y_i\}$, $\mathcal{B}_c = \{j \in [2M] : y_j = c\}$, $n_c = |\mathcal{B}_c|$, and $\hat{\boldsymbol{\mu}}_c = \frac{1}{n_c}\sum_{j\in\mathcal{B}_c} \boldsymbol{a}_j$.

(A) SCL lower bound. By unfolding the SCL loss defined in Equation (2), the per-example loss term can be written as

$$\ell_i^{\text{SCL}} = -\frac{1}{|\mathcal{C}(i)|}\sum_{l\in\mathcal{C}(i)} \frac{\boldsymbol{a}_i^\top \boldsymbol{a}_l}{\tau} + \log \sum_{j\in[2M]\setminus\{i\}} \exp(\boldsymbol{a}_i^\top \boldsymbol{a}_j/\tau).$$

For the first term, using $\frac{1}{|\mathcal{C}(i)|}\sum_{l\in\mathcal{C}(i)} \boldsymbol{a}_l = \frac{n_{y_i}\hat{\boldsymbol{\mu}}_{y_i} - \boldsymbol{a}_i}{n_{y_i}-1}$ and $\|\boldsymbol{a}_i\| = 1$ gives $-\frac{\boldsymbol{a}_i^\top}{\tau}\big(\frac{1}{|\mathcal{C}(i)|}\sum_{l\in\mathcal{C}(i)} \boldsymbol{a}_l\big) \geq -\frac{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_{y_i}}{\tau}$.

For the second term, we group by class, subtract the self term and then apply Jensen classwise due to convexity of the exponential function:

$$\sum_{j\in[2M]\setminus\{i\}} e^{\boldsymbol{a}_i^\top \boldsymbol{a}_j/\tau} = \sum_{c=1}^{K}\sum_{l\in\mathcal{B}_c} e^{\boldsymbol{a}_i^\top \boldsymbol{a}_l/\tau} - e^{1/\tau} \ \geq\ \sum_{c=1}^{K} n_c\, e^{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_c/\tau} - e^{1/\tau}.$$

Combining,

$$\ell_i^{\text{SCL}} \ \geq\ -\frac{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_{y_i}}{\tau} + \log\left(\sum_{c=1}^{K} n_c\, e^{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_c/\tau} - e^{1/\tau}\right) \ =:\ \ell_i^\star. \tag{8}$$

Equality in equation 8 holds iff every class-wise sum is collapsed, i.e., $\boldsymbol{a}_j = \hat{\boldsymbol{\mu}}_c$ for all $j \in \mathcal{B}_c$, because the positive-term bound is tight only when $\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_{y_i} = 1$ (so $\boldsymbol{a}_i = \hat{\boldsymbol{\mu}}_{y_i}$) and the classwise Jensen step is tight only when all within-class logits $\{\boldsymbol{a}_i^\top \boldsymbol{a}_l : l \in \mathcal{B}_c\}$ are equal.

(B) Prototype loss lower bound. Since $\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_{y_i} \le 1$ for unit vectors, $e^{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_{y_i}/\tau} \le e^{1/\tau}$. Therefore

$$\underbrace{\sum_{c=1}^{K} n_c\, e^{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_c/\tau} - e^{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_{y_i}/\tau}}_{=:D_i^{\text{proto}}} \;\ge\; \underbrace{\sum_{c=1}^{K} n_c\, e^{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_c/\tau} - e^{1/\tau}}_{=:D_i^{\star}},$$

and thus, with the *same* numerator $e^{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_{y_i}/\tau}$,

$$\ell_i^{\text{proto}} = -\frac{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_{y_i}}{\tau} + \log D_i^{\text{proto}} \;\ge\; -\frac{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_{y_i}}{\tau} + \log D_i^{\star} = \ell_i^{\star}.$$

Averaging over i gives the following inequalities for any batch $\boldsymbol{A}$:

$$\begin{aligned} L_{\text{SCL}}(\boldsymbol{A}) &\ge L_{\star}(\boldsymbol{A}) \\ L_{\text{proto}}(\boldsymbol{A}) &\ge L_{\star}(\boldsymbol{A}). \end{aligned}$$

(C) Collapse–simplex makes all three equal. By Graf et al. (2021, Theorem 2), any SCL global minimizer exhibits class-wise collapse, $\boldsymbol{a}_j = \boldsymbol{\zeta}_{y_j}$, and the directions $\{\boldsymbol{\zeta}_c\}$ form a centered regular $(K-1)$-simplex. Hence $\hat{\boldsymbol{\mu}}_c = \boldsymbol{\zeta}_c$ and $\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_{y_i} = 1$ for all i, making both inequalities above tight:

$$L_{\text{SCL}}(\boldsymbol{A}^\star) = L_{\star}(\boldsymbol{A}^\star) = L_{\text{proto}}(\boldsymbol{A}^\star).$$

Therefore $\min L_{\text{SCL}} = \min L_{\star} = \min L_{\text{proto}}$, all attained at the collapsed-simplex configurations.

(D) Equality of argmin sets. Let $\boldsymbol{A}$ minimize L_{proto}. Then $L_{\text{proto}}(\boldsymbol{A}) = \min L_{\text{proto}} = \min L_{\star}$, so $L_{\star}(\boldsymbol{A}) = L_{\text{proto}}(\boldsymbol{A})$, which forces $e^{\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_{y_i}/\tau} = e^{1/\tau}$ for every i, i.e., $\boldsymbol{a}_i^\top \hat{\boldsymbol{\mu}}_{y_i} = 1$ and hence $\boldsymbol{a}_i = \hat{\boldsymbol{\mu}}_{y_i}$ (class-wise collapse). Moreover $L_{\text{SCL}}(\boldsymbol{A}) = L_{\star}(\boldsymbol{A}) = \min L_{\text{SCL}}$, so $\boldsymbol{A}$ also minimizes SCL.

Graf's theorem then implies the class means form a centered simplex ETF. Thus the argmin sets of L_{SCL} and L_{proto} coincide (up to rotation and label permutation).

□

A.2 Implementation Details

Experiments are conducted on four standard image classification datasets: *CIFAR10, CIFAR100, ImageNet-100, and ImageNet1K*, following common representation learning benchmarking practices (Khosla et al., 2020; Markou et al., 2024; Wang et al., 2021; Yeh et al., 2022). We use ResNet50 for ImageNet-100/ImageNet1K and ResNet18 for CIFAR10/CIFAR100. All models are trained using SGD optimizer for 500 epochs on ImageNet1K (batch size 2048, temperature 0.1) and ImageNet-100 (batch size 1024, temperature 0.1) and 1000 epochs on CIFAR10/CIFAR100. For CIFAR10/100 we set the batch size to 512 and evaluate all 11 temperatures in the set [0.07, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1]. In Table 1 and Table 2 we report for each method the best performing temperature. For Supervised Contrastive Learning we perform the linear probing phase for the typical 90 epochs.

A.2.1 Classifier Learning Methods (CE, NormFace, NTCE, NONL)

For the family of classifier learning methods, we employ the following hyperparameters across datasets:

CIFAR10/CIFAR100. Models are trained for 1000 epochs with batch size 512. We use SGD optimizer with momentum 0.9, weight decay 10^{-4}, and initial learning rate 0.2. The learning rate follows a cosine annealing schedule throughout training, decaying to a minimum value of $\eta_{\min} = \eta_0 \times 0.1^3$ where η_0 is the initial learning rate. Data augmentation consists of RandomResizedCrop with scale (0.2, 1.0), RandomHorizontalFlip, and standard normalization with dataset-specific mean and standard deviation values.

ImageNet-100. ResNet50 models are trained for 500 epochs with batch size 1024 (256 per GPU with 4 GPUs). We employ SGD optimizer with momentum 0.9, weight decay 10^{-4}, and initial learning rate 0.1, which is automatically scaled based on the total batch size. We use cosine annealing scheduler with 10 epochs of linear warmup from 0.01 to the target learning rate. After warmup, the learning rate follows a cosine decay to $\eta_{\min} = \eta_0 \times 0.1^3$. Synchronized BatchNorm is enabled across GPUs. Data augmentation includes RandomResizedCrop(224) with scale (0.2, 1.0), RandomHorizontalFlip, and standard ImageNet normalization.

ImageNet1K. ResNet50 models are trained for 500 epochs with batch size 2048 (256 per GPU with 8 GPUs). Hyperparameters follow the same configuration as ImageNet-100, with SGD optimizer (momentum 0.9, weight decay 10^{-4}), initial learning rate 0.1 with automatic scaling based on batch size. We apply 10 epochs of linear warmup followed by cosine annealing to $\eta_{\min} = \eta_0 \times 0.1^3$. Data augmentation and normalization follow ImageNet-100 settings.

A.2.2 Supervised Contrastive Learning

For supervised contrastive methods, we implement a two-phase training procedure:

Phase 1: Contrastive Training.

CIFAR10/CIFAR100: Models are trained for 1000 epochs with batch size 512. SGD optimizer is used with momentum 0.9, weight decay 10^{-4}, and initial learning rate 0.05. The learning rate follows cosine annealing schedule throughout training, decaying to $\eta_{\min} = \eta_0 \times 0.1^3$. We use extensive data augmentation including RandomResizedCrop with scale (0.2, 1.0), RandomHorizontalFlip, ColorJitter(0.4, 0.4, 0.4, 0.1) with probability 0.8, and RandomGrayscale with probability 0.2. Each image generates two augmented views for contrastive learning.

ImageNet-100: ResNet50 encoder with 128-dimensional projection head is trained for 500 epochs with batch size 1024. We use SGD optimizer with momentum 0.9, weight decay 10^{-4}, and base learning rate 0.8 (automatically scaled by batch size). Learning rate follows cosine annealing with 10 epochs linear warmup from 0.01, then decays following a cosine schedule to $\eta_{\min} = \eta_0 \times 0.1^3$. Data augmentation extends CIFAR settings with the addition of Gaussian blur for ImageNet scale images.

ImageNet1K: Training spans 500 epochs with batch size 2048 using the same optimizer configuration as ImageNet-100. Base learning rate is set to 0.1 with automatic scaling. We employ cosine annealing with 5 epochs warmup from 0.01, followed by cosine decay to $\eta_{\min} = \eta_0 \times 0.1^3$. The same augmentation pipeline as ImageNet-100 is used.

Phase 2: Linear Evaluation. For all datasets, we freeze the learned encoder and train a linear classifier on top of the representations:

CIFAR10/CIFAR100: Linear classifier is trained for 100 epochs using SGD with learning rate 5.0, momentum 0.9, and zero weight decay. Learning rate is decayed by factor 0.2 at epochs 60, 75, 90 using a step scheduler.

ImageNet-100: Linear evaluation runs for 90 epochs with SGD optimizer, learning rate 2.0, momentum 0.9, and zero weight decay. Learning rate decay by factor 0.2 occurs at epochs 30, 60, 80 using a step scheduler.

ImageNet1K: Linear classifier training spans 90 epochs with SGD, learning rate 0.8, momentum 0.9, and zero weight decay. The same step decay schedule as ImageNet-100 is applied.

A.2.3 Additional Implementation Details

For distributed training on ImageNet datasets, we employ DistributedDataParallel with one process per GPU. Random seed is fixed at 42 for reproducibility. The cosine annealing scheduler is implemented following the standard formulation: $\eta_t = \eta_{\min} + \frac{1}{2}(\eta_0 - \eta_{\min})(1 + \cos(\frac{\pi t}{T}))$, where t is the current epoch and T is the total number of epochs. For experiments with warmup, the warmup period linearly interpolates from the warmup starting learning rate to the initial learning rate before transitioning to cosine annealing. Temperature parameter τ is searched over the range [0.07, 0.1, 0.2, ..., 1.0] for CIFAR experiments, while ImageNet experiments use the optimal temperature found

through preliminary experiments (0.1 for supervised contrastive, 0.2 for classifier learning methods). All models use standard weight initialization and no additional regularization beyond weight decay.

A.3 Extra Ablation Studies

A.3.1 Role of the Projection Head

Table 4: **Contrastive Learning Results - Without Projection Head.** Performance comparison across different classifier learning approaches without projection head.

Classifier Learning	Loss	CIFAR-10	CIFAR-100	ImageNet-100	ImageNet-1K
Linear Probing	SCL	95	70.6	84.1	71
Normalized Linear Probing	SCL	95	71.4	84.3	72.1
Fixed Prototypes	SCL	95	71.4	84.7	70.1

In Table 4 we demonstrate the importance of the projection head in contrastive training. Across three datasets, except on the relatively simple CIFAR-10 benchmark, removing the head consistently reduces accuracy by more than 2 points. At first glance, one might expect the opposite: discarding the head should let the loss act directly on the final encoder embeddings on the unit hypersphere. We hypothesize that the projection head helps primarily by imposing a beneficial dimensionality bottleneck. With ResNet-50, the encoder's representation is 2048-dimensional, whereas the projection head maps it to 128 dimensions. For a K-class problem (e.g., $K = 100$), the ideal equiangular tight frame (ETF) geometry lives in a $(K-1)$-dimensional subspace. Encouraging embeddings to adopt this structure is plausibly easier in a 128-dimensional space than in a 2048-dimensional one, where the optimizer has many more irrelevant directions to explore.

A.3.2 Effective Hyperparameter ranges

Normalized softmax losses introduce too hyperparameters that originate from contrastive learning (i) temperature, and (ii) need for larger batch size. Here we test whether and how these haperparameters affect the downstream performance.

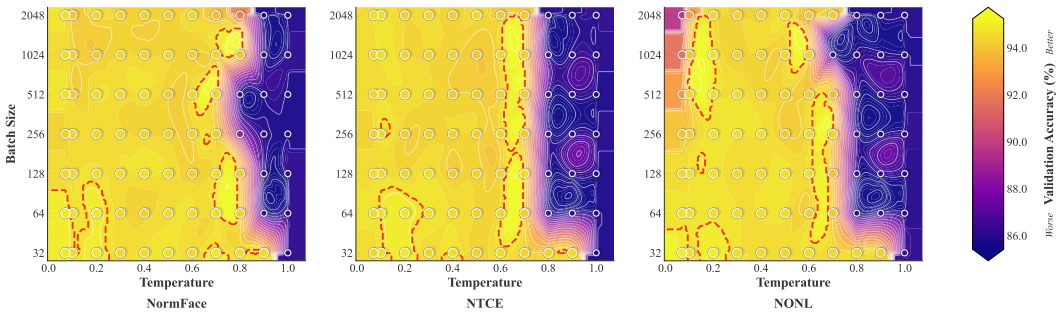
We conduct hyperparameter optimization experiments on CIFAR-10 and CIFAR-100, evaluating all combinations of 11 temperatures in the range [0.07, 1] and 7 batch sizes in the range [32, 2048]. The results show that different contrastive learning methods exhibit distinct optimal hyperparameter regions with minimal overlap in their peak performance zones across both datasets.

In Figure 2 we can see that normalized softmax losses exhibit the same behavior in terms of downstream performance compared to self-supervised contrastive learning (Chen et al., 2020), which means that there are trustable goto to setups for instance $\tau = \{0.1, 0.2\}$ for small to medium number of classes datasets and $\tau = \{0.07, 0.1\}$ for large. For that reason normalized softmax methods despite introducing extra hyperparameters, this is not a problem in practice

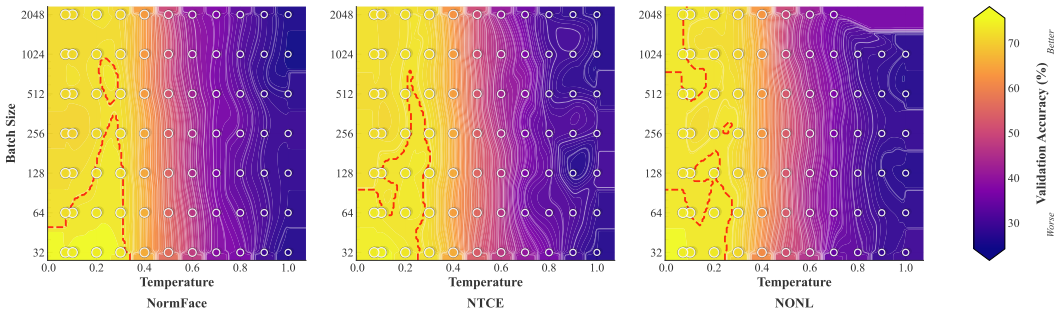
A.3.3 Effective hyperparameter ranges

Normalized softmax losses introduce two hyperparameters inherited from contrastive learning: the temperature τ, which controls the sharpness of the similarity distribution, and the need for larger batch size B, which governs the number of in-batch negatives. We assess their impact by grid–searching $\tau \in [0.07, 1.0]$ (11 values) and $B \in \{32, 64, 128, 256, 512, 1024, 2048\}$ on CIFAR-10 and CIFAR-100 with NormFace, NTCE, and NONL.

Figure 2 shows consistent "sweet spots" across methods: accuracy forms a pronounced band at moderate temperatures, with performance degrading for overly large τ and, to a lesser extent, for very small τ. The location of this band shifts toward slightly smaller temperatures as the number of classes increases (CIFAR-100 vs. CIFAR-10), mirroring observations in self-supervised contrastive learning (Chen et al., 2020). Within the effective temperature range, performance is comparatively



(a) CIFAR-10



(b) CIFAR-100

Figure 2: **Validation Accuracy (%) Phase Diagrams.** Classification accuracy on validation set. Higher values indicate better generalization performance. Each subplot shows the performance landscape across temperature and batch size hyperparameters for different loss functions: NormFace, NTCE, and NONL. Brighter regions indicate superior performance. White contour lines indicate iso-performance curves for detailed analysis. Red dashed contours highlight optimal parameter regions (top 10% performance). Scatter points represent individual experimental runs with performance-based sizing. Each dataset uses its own optimal colorbar range. Results originate from grid runs across temperature values in [0.07, 1.0] and batch sizes in 32, 64, 128, 256, 512, 1024, 2048.

insensitive to B, yielding a broad plateau over batch sizes—large batches can help, but are not strictly required.

In practice, these trends provide the same reliable defaults as in self-supervised contrastive learning (Chen et al., 2020): $\tau \in \{0.1, 0.2\}$ works well for small- to medium-class datasets, while $\tau \in \{0.07, 0.1\}$ is preferable for larger-class settings. Thus, although normalized softmax losses expose additional hyperparameters, their effective ranges are narrow and stable, so a small amount of tuning (or even these defaults) is typically sufficient to reach near-peak accuracy.