Latent Point Collapse on a Low Dimensional Embedding in Deep Neural Network Classifiers

Anonymous authors
Paper under double-blind review

Abstract

The topological properties of latent representations play a critical role in determining the performance of deep neural network classifiers. In particular, the emergence of well-separated class embeddings in the latent space has been shown to improve both generalization and robustness. In this paper, we propose a method to force the collapse of latent representations belonging to the same class into a single point, which enhances class separability in the latent space while confining network outputs to a bounded range. We demonstrate that this phenomenon, which we call *latent point collapse* (LPC), is induced by adding a strong L_2 penalty on the penultimate-layer representations and arises from the interplay between the L_2 penalty and the cross-entropy loss. We note that the added penalty makes the loss strongly convex with respect to the penultimate representations, and that increasing its strength tightens the confinement and reduces the within-class spread of the representations. In addition, we show the practical utility of applying this compressing loss term to the latent representations of a low-dimensional linear penultimate layer. LPC can be viewed as a stronger manifestation of *neural collapse* (NC): while NC only requires within-class representations to converge around their class means relative to between-class separation, LPC enforces absolute convergence to single points near the origin, yielding more pronounced improvements in robustness and generalization.

1 Introduction

Deep neural networks (DNNs) achieve high accuracy across many domains, but robustness and generalization remain difficult to guarantee. Progress on these challenges has come from understanding the geometry of the latent representations learned by DNNs. In particular, it has been observed that increasing the margin between classes, i.e., making classes more separable in the latent space, can improve both robustness and generalization (Elsayed et al., 2018; Wei & Ma, 2020; Ding et al., 2020; Sokolic et al., 2017). Indeed, the relationship between generalization and robustness is well established in the literature (Xu & Mannor, 2010; Szegedy et al., 2014; Achille & Soatto, 2018; Novak et al., 2018; Zhang et al., 2019).

DNNs naturally tend to improve the separation of different classes in the latent space during training, and this process occurs at a constant geometric rate (He & Su, 2023). Such evolving separation is associated with the phenomenon of *neural collapse* (NC) (Papyan et al., 2020; Han et al., 2022). While NC is primarily studied in the penultimate layer of DNN classifiers, its characteristics, propagation, and related phenomena have also been investigated in intermediate layers (Súkeník et al., 2023; Masarczyk et al., 2023; Súkeník et al., 2024; Parker et al., 2023; Rangamani et al., 2023). This phenomenon typically occurs in overparameterized models during the terminal phase of training (TPT). Even after the point of zero training error, the network further refines the representations by increasing their relative distances in the latent space. In practice, this means that the class means in the penultimate layer collapse to the vertices of an equiangular tight-frame simplex (ETFS). The occurrence of NC has been observed in large language models (Wu & Papyan, 2024), in graph neural networks (Kothapalli et al., 2023), with multivariate regression (Andriopoulos et al., 2024), and extensively investigated in unconstrained feature models (UFMs) (Yaras et al., 2022; Mixon et al., 2020; Ergen & Pilanci, 2020; Zhu et al., 2021; Graf et al., 2021; Ji et al., 2022; Tirer & Bruna, 2022; Zhou et al.,

2022a; Fisher et al., 2024; Thrampoulidis et al., 2022; Yang et al., 2022; Yan et al., 2024; Hong & Ling, 2024; Fang et al., 2021; Liu et al., 2023) and in the mean-field regime (Wu & Mondelli, 2025).

NC implies convergence to a geometry with enhanced separability, which improves generalization and robustness. In fact, it has been shown that these metrics continue to improve during the TPT (Papyan et al., 2020), precisely when the latent representations approach an ETFS. Subsequent research has identified further consequences of NC. For instance, NC has been linked to improved knowledge distillation (Zhang et al., 2025), few-shot class incremental learning (Yang et al., 2023) and transfer learning (Galanti et al., 2022a;b; Li et al., 2024; Munn et al., 2024). Building on this connection, Wang et al. (2023) use NC-based metrics to enhance the transferability of models. Another direction connects NC to out-of-distribution (OOD) detection (Haas et al., 2023; Ammar et al., 2024; Harun et al., 2025; Wu et al., 2025) and generalization (Chen et al., 2024).

The connection between margin-based approaches and robust generalization can also be understood through the framework of the information bottleneck (IB) principle (Tishby et al., 2000; Tishby & Zaslavsky, 2015). The IB principle posits that effective DNNs develop compact yet sufficiently informative latent representations by minimizing the mutual information between inputs and latent representations, while preserving information relevant for prediction. Empirically, it has been shown that IB improves network performance (Hu et al., 2024), and theoretical work provides rigorous arguments for IB's role in controlling generalization errors (Kawaguchi et al., 2023). In practice, DNN training proceeds in two phases: an empirical risk minimization phase, where the network primarily fits the data, followed by a compression phase, where the network constructs more compact embeddings layer by layer (Shwartz-Ziv & Tishby, 2017). This compression aligns with margin maximization and NC, suggesting that efficient representations emerge in both information-theoretic and geometric terms.

1.1 Contributions

We introduce *latent point collapse* (LPC), a phenomenon in which penultimate-layer representations of each class converge to distinct points near the origin under **strong** L_2 regularization. Unlike NC, which permits unbounded representation growth and only achieves *relative* convergence, LPC enforces strict geometric confinement with provable bounds on output variation, a property that standard NC cannot ensure. The key distinction lies in the collapse metrics. While NC theory formally defines collapse as the within-class covariance $\mathbf{\Sigma}_W \to 0$, practical implementations only achieve the weaker condition $\mathrm{Tr}(\mathbf{\Sigma}_W \mathbf{\Sigma}_B^{\dagger}) \to 0$, where $\mathbf{\Sigma}_B$ is the between-class covariance (Papyan et al., 2020). This relative metric allows substantial within-class variance as long as between-class separation dominates. Our theoretical analysis shows that cross-entropy loss alone drives unbounded norm growth, preventing true collapse. In contrast, LPC achieves $\mathbf{\Sigma}_W \to 0$ *absolutely*, with representations confined close to the origin. This confinement yields a uniform bound on output variation, a robustness property unattainable through standard NC. Our specific contributions are:

- **Discovery and theoretical characterization of latent point collapse:** We identify and analyze the emergent phenomenon of LPC, where strong L_2 regularization induces strict collapse of class representations to equilibrium points near the origin and provides a uniform output bound that is independent of input distance (Section 2). We further demonstrate that LPC represents a stronger manifestation of standard NC properties, with substantially tighter within-class collapse and full convergence to the ETFS structure (Section 5).

- **Empirical validation and performance improvements:** We experimentally confirm LPC in practice, observing collapse points near the origin and reduced output sensitivity (Section 4.2). We demonstrate substantial improvements in robustness and consistent gains in generalization (Section 4.3). Our approach uses a low-dimensional linear penultimate layer that acts as a bottleneck, amplifying both LPC and performance.

- **Information bottleneck connection:** We establish that LPC naturally induces an IB in the penultimate layer, providing an information-theoretic perspective on the observed generalization improvements (Section 3).

- **Binary encoding:** We document an emergent phenomenon where penultimate latent representations converge to hypercube vertices (Section 4.1).

1.2 Related Work

Our approach employs L_2 regularization on latent representations rather than network weights, an important distinction from conventional regularization strategies. While regularization strategies such as dropout (Srivastava et al., 2014) and weight-based penalties such as L_0 (Louizos et al., 2018), L_1 (Lemhadri et al., 2021), or weight decay are well established, they do not induce the geometric collapse phenomenon we observe. By targeting penultimate-layer representations directly, we achieve feature compression without architectural modifications.

The application of L_2 penalties to latent representations is common in theoretical analyses of UFMs (Tirer & Bruna, 2022; Fang et al., 2021; Zhu et al., 2021; Yaras et al., 2022; Graf et al., 2021; Hong & Ling, 2024; Yan et al., 2024; Yang et al., 2022; Thrampoulidis et al., 2022; Fisher et al., 2024; Zhou et al., 2022a), where both weight and feature regularization were employed to establish existence of global optimizers. However, these works used very small regularization coefficients (typically $\gamma \ll 1$, where γ denotes the strength of the L_2 penalty) that served purely as mathematical convenience without affecting network behavior. Such weak regularization neither induces observable LPC nor provides practical benefits. By contrast, we demonstrate the critical role of extreme regularization strengths ($\gamma = 10^6$ in our experiments) that fundamentally alter the optimization landscape. Our theoretical analysis (Appendix A) shows that in this regime the curvature of the feature-space loss is dominated by the isotropic penalty term, confining representations within radius $O(1/\gamma)$ of the origin and shrinking within-class fluctuations as $O(1/\sqrt{\gamma})$, effects that are negligible at the small regularization values used in prior work. The strong L_2 penalty establishes a confining potential that counteracts the unbounded growth inherent to cross-entropy minimization. This tension between opposing forces, with cross-entropy pushing representations outward and L_2 regularization pulling them toward the origin, induces LPC. It is not obvious that this severe confinement near the origin enhances rather than degrades performance, while simultaneously providing a provable bound on output variation. The novelty of our work lies not in using regularization itself, but in identifying and analyzing the collapse phenomenon it induces.

Adding loss terms to intermediate layers has also appeared in the context of deep supervision (Lee et al., 2015; Li et al., 2022), where intermediate outputs are trained to match target labels. However, our approach differs by seeking to compress the volume of latent representations rather than providing additional supervision signals.

Various methods have been devised to enlarge inter-class margins, such as contrastive learning (Wu et al., 2018; Hénaff et al., 2020; van den Oord et al., 2018; Tian et al., 2020; Hjelm et al., 2019; Chen et al., 2020; He et al., 2019) and supervised contrastive learning (SupCon) (Khosla et al., 2020), which pull together positive samples while pushing apart negative ones. Other techniques alter the loss function to reduce intra-class variance (Wen et al., 2016) or impose angular constraints (Liu et al., 2016; 2017), e.g., CosFace (Wang et al., 2018) and ArcFace (Deng et al., 2019). The latter two methods, ArcFace and CosFace, can be compared to our method in their simplicity, as they each introduce a single penalty term to the loss function to increase margins.

Bounding network outputs addresses several known failure modes of unconstrained architectures: the unbounded high-confidence predictions of deep networks far from training data (Hein et al., 2019), the systematic overconfidence of standard classifiers (Guo et al., 2017), and the violation of the bounded-loss assumption required by stability-based generalization theory (Bousquet & Elisseeff, 2002). Prior methods for controlling output sensitivity rely on architectural constraints such as spectral norm regularization (Miyato et al., 2017; Cisse et al., 2017; Tsuzuku et al., 2018), orthogonal weight matrices (Trockman & Kolter, 2021; Singla & Feizi, 2021; Singla et al., 2022), or norm-bound weights (Anil et al., 2018; Gouk et al., 2018), which can reduce model expressiveness or be computationally expensive. Other approaches project features onto a hypersphere (Pang et al., 2020; Deng et al., 2019; Wang et al., 2018). By contrast, our method imposes no specialized architectural constraints and confines representations to a region of controlled radius without forcing them onto a fixed surface.

2 Method and Theoretical Characterization

Given a labeled dataset $\{\boldsymbol{x}_i, \bar{y}_i\}_{i=1}^N$, where N denotes the number of training samples, we address the problem of multi-class classification using DNNs. We employ a DNN $\boldsymbol{f}(\boldsymbol{x})$ that learns a nonlinear mapping from input space to output space, approximating the underlying data distribution. DNNs consist of multiple layers arranged hierarchically, with each layer producing an intermediate latent representation. The network's output can be expressed as a composition of layer-wise transformations: $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}^{(M)} \circ \boldsymbol{f}^{(M-1)} \circ \cdots \circ \boldsymbol{f}^{(1)}(\boldsymbol{x})$, where M denotes the depth of the network.

For an input vector $\boldsymbol{x}$, the forward pass through the network can be conceptually divided into two stages. First, the nonlinear components transform the input into a high-dimensional latent representation $\boldsymbol{h}(\boldsymbol{x})$, corresponding to the output of the final hidden layer. Subsequently, a linear classifier maps this representation to the output space: $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{b}$, where $\boldsymbol{W}$ and $\boldsymbol{b}$ denote the weight matrix and bias vector of the classifier, respectively. The predicted class label y is obtained by applying the softmax function to the network's output, yielding a probability distribution over classes that quantifies the likelihood of input $\boldsymbol{x}$ belonging to each class. The network parameters are optimized by minimizing the cross-entropy loss: $\mathcal{L}_{\text{CE}}(\boldsymbol{f}(\boldsymbol{x}), \bar{y}) = -\log \frac{\exp(f_{\bar{y}}(\boldsymbol{x}))}{\sum_j \exp(f_j(\boldsymbol{x}))}$, which measures the divergence between predicted probabilities and ground-truth labels.

We propose augmenting the architecture with an additional linear transformation preceding the classifier: $\boldsymbol{z} = \boldsymbol{W}_{\text{L2}}\boldsymbol{h}(\boldsymbol{x}) + \boldsymbol{b}_{\text{L2}}$. This layer functions as the penultimate representation, with final classification performed via: $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}$. Beyond the standard cross-entropy loss, we introduce an L_2 regularization term applied to the penultimate layer: $\mathcal{L}_{\text{L2}}(\boldsymbol{z}) = \|\boldsymbol{z}\|^2$ where $||\cdot||$ denotes the Euclidean norm. The composite loss function becomes: $\mathcal{L} = \mathcal{L}_{\text{CE}} + \gamma\mathcal{L}_{\text{L2}}$, where $\gamma > 0$ is a regularization hyperparameter controlling the strength of the L_2 penalty. The phenomenon of LPC arises from the interplay between these competing objectives:

$$\mathcal{L} = -\log \frac{\exp((\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b})_{\bar{y}})}{\sum_j \exp((\boldsymbol{W}\boldsymbol{z} + \boldsymbol{b})_j)} + \gamma||\boldsymbol{z}||^2. \tag{1}$$

The *latent point collapse* phenomenon, illustrated in Figure 1, occurs when strong L_2 regularization applied to the representations $\boldsymbol{z}$ in the penultimate layer causes representations to collapse into distinct, class-specific points while maintaining separation between classes. LPC emerges from the interplay between competing optimization forces. We provide intuition here; rigorous proofs and complete theoretical characterization appear in Appendix A.

The necessity of regularization The unconstrained feature model (UFM) framework (Zhu et al., 2021; Fang et al., 2021; Tirer & Bruna, 2022) predicts that global minimizers of cross-entropy loss exhibit perfect NC, i.e. all within-class features converge to single points forming an ETFS. However, practical deep networks do not always reach such optima (Tirer et al., 2022).

Our analysis identifies the main obstacle: pure cross-entropy optimization exhibits an intrinsic instability. Its gradient contains an outward radial component that causes representations to grow unboundedly, preventing convergence to finite equilibria. We formalize this as a necessity result: without regularization ($\gamma = 0$), the outward radial component drives $\|\mathbf{z}\| \to \infty$, precluding any finite equilibria (proof in Appendix A). As norms increase, softmax probabilities saturate and gradients vanish, so representations never settle at finite equilibria and the collapse remains incomplete.

Strong L_2 regularization resolves this instability by introducing an inward restoring force $-2\gamma\mathbf{z}$ that counteracts the outward push. Let $\mathbf{w}_i^\top$ denote the i-th row of $\mathbf{W}$, so that the logit for class i is $\mathbf{w}_i^\top\mathbf{z} + b_i$, and let $p_i(\mathbf{z})$ denote the softmax probability assigned to class i. At equilibrium, these competing forces balance according to:

$$\mathbf{z}^* = \frac{1}{2\gamma}\left(\mathbf{w}_{\bar{y}} - \sum_{i=1}^{K} p_i(\mathbf{z}^*)\mathbf{w}_i\right) \tag{2}$$

This shows that the equilibrium radius scales inversely with regularization strength: $\|\mathbf{z}^*\| \leq M_W/\gamma$, where M_W bounds classifier weight norms (proof in Appendix A). Figure 2 (left) empirically validates this scaling.

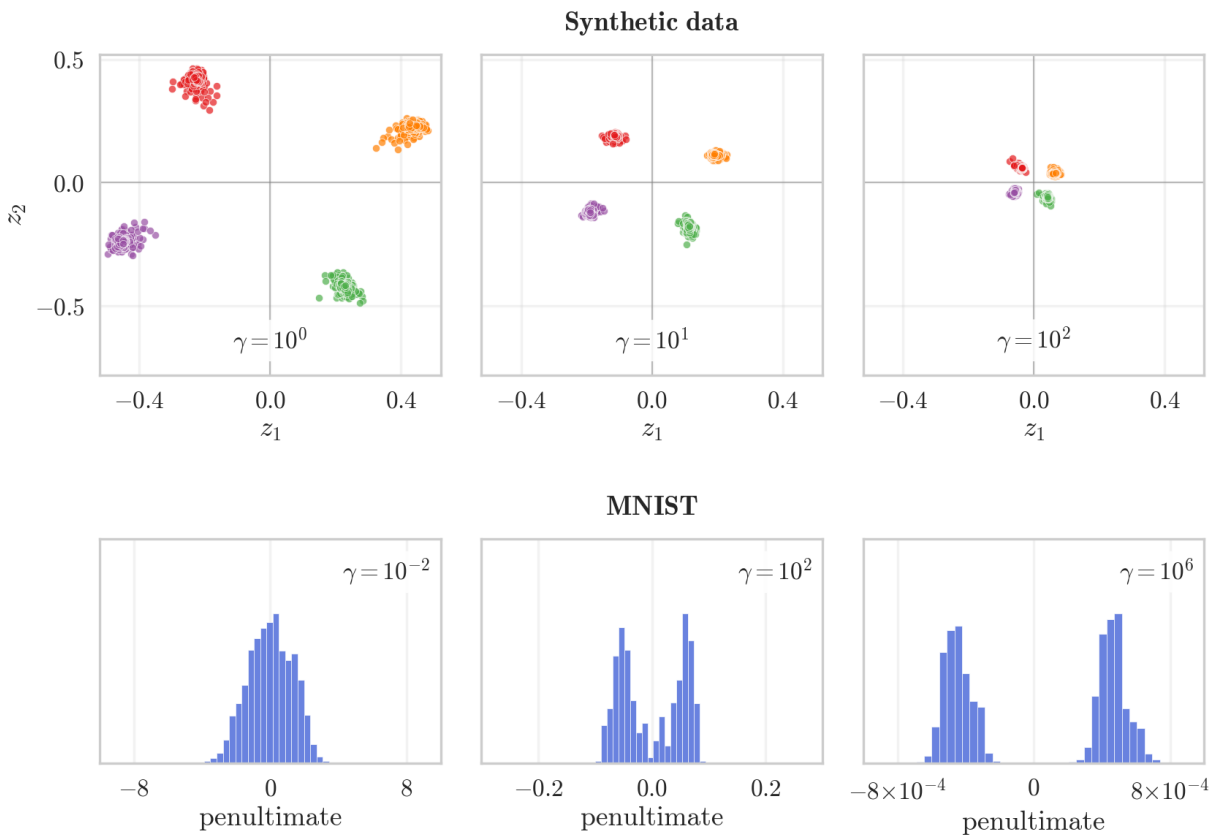

Figure 1: Latent point collapse progression on synthetic and MNIST datasets. *Top row:* 2D penultimate layer activations for a 4-class synthetic dataset at increasing γ regularization strengths. Each color represents a different class. The progression illustrates the competing forces at play: the cross-entropy loss drives class separation, while the $\gamma\|\mathbf{z}\|^2$ penalty compresses representations toward the origin. As γ increases, both the distance from the origin and the within-class spread decrease, yet class separability is preserved. *Bottom three rows:* Histograms of penultimate activations for MNIST classification across all penultimate nodes and datapoints. As γ increases, the activation distribution becomes increasingly concentrated around zero. At high γ values, the network responds to the extreme compression by converging toward binary encoding, with activations clustering at distinct points opposite to each other relative to the origin. Both experiments demonstrate how LPC regularization systematically reduces representation norms while maintaining discriminative power. A Jupyter notebook reproducing these results is available at https://anonymous.4open.science/r/lpc-0CEB/notebooks/lpc_demo.ipynb.

Strong convexity and collapse dynamics The cross-entropy Hessian $\nabla^2_{\mathbf{z}}\mathcal{L}_{\text{CE}}(\mathbf{z}) = \mathbf{W}^\top(\text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top)\mathbf{W}$ is positive semi-definite and bounded, $\mathbf{0} \preceq \nabla^2_{\mathbf{z}}\mathcal{L}_{\text{CE}}(\mathbf{z}) \preceq KM_W^2\,\mathbf{I}$, so that

$$2\gamma\mathbf{I} \;\preceq\; \nabla^2_{\mathbf{z}}\mathcal{L}(\mathbf{z}) \;\preceq\; \left(2\gamma + KM_W^2\right)\mathbf{I}. \tag{3}$$

The loss is therefore 2γ-strongly convex in the feature space for any $\gamma > 0$, guaranteeing a unique optimum with respect to the penultimate representations. The role of a large penalty is thus to make the curvature dominant and nearly isotropic: for $\gamma \gg KM_W^2$ the Hessian is well approximated by $2\gamma\mathbf{I}$, which both accelerates convergence and tightens the collapse.

Beyond establishing global convergence, strong regularization also controls local fluctuations. Near equilibrium, the dominant quadratic potential (second derivative $\approx 2\gamma\mathbf{I}$) tightly constrains stochastic fluctuations from gradient noise. Under standard assumptions of bounded classifier weights, fixed gradient noise variance σ^2, and the stability condition $\eta < 1/(2\gamma\bar{\lambda})$ (where $\bar{\lambda}$ characterizes the optimizer's preconditioning),

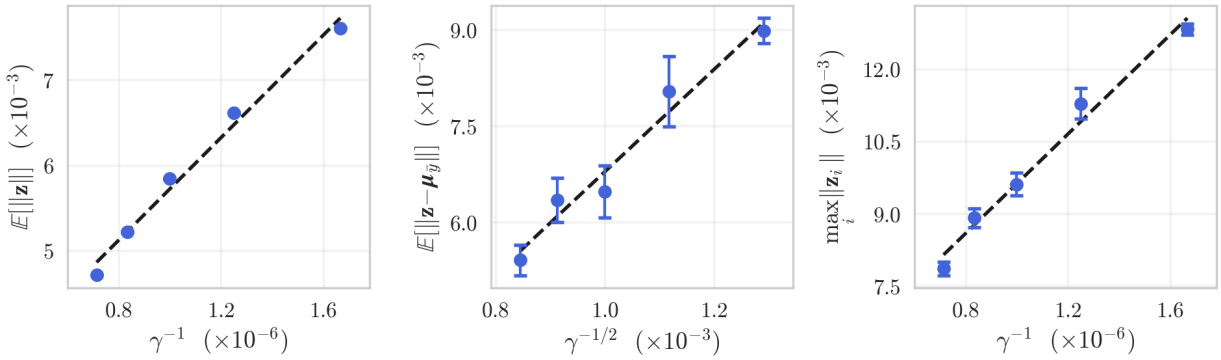

Figure 2: Scaling relationships between LPC regularization strength and penultimate layer geometry. CIFAR-10 trained as in Appendix B but with varying regularization strengths $\gamma_{max} \in \{6 \times 10^5, 8 \times 10^5, 10^6, 1.2 \times 10^6, 1.4 \times 10^6\}$ and fixed initial learning rate $\eta = 10^{-4}$. Blue points show empirical measurements with error bars indicating standard deviation across 5 runs. Dashed lines represent linear fits to the data. From left to right: Mean penultimate feature norm $\mathbb{E}[\|\mathbf{z}\|]$ as a function of $1/\gamma$. Mean within-class distance from class centroid $\mathbb{E}[\|\mathbf{z} - \boldsymbol{\mu}_{\bar{y}}\|]$ as a function of $1/\sqrt{\gamma}$. Full radius containing all penultimate representations $\max_i \|\mathbf{z}_i\|$ as a function of $1/\gamma$. Linear fits demonstrate strong scaling relationships with $R^2 = 0.9865$, $R^2 = 0.9647$, and $R^2 = 0.9808$, respectively. The tight linear relationships confirm theoretical predictions that LPC regularization systematically controls the geometric properties of learned representations.

our analysis establishes that steady-state variance decreases monotonically with γ. Specifically, the collapse radius scales as

$$r_{\text{collapse}} = O\left(\sqrt{P\sigma^2\eta/\gamma}\right), \tag{4}$$

where P is the representation dimensionality and η the learning rate (proof in Appendix A). Both the equilibrium radius $\|\mathbf{z}^*\| \leq M_W/\gamma$ and the collapse radius therefore decrease as γ grows. Figure 2 confirms both scalings.

Thus, γ serves a dual role: it guarantees a unique finite optimum for the representations while simultaneously controlling the tightness of intra-class clustering through increased curvature.

Bounded output sensitivity and robustness The bounded confinement directly limits output variation. Since all representations lie within radius $O(M_W/\gamma)$ of the origin, the maximum possible distance between any two network outputs is bounded independently of input distance. We prove that for any pair of inputs $\mathbf{x}_1, \mathbf{x}_2$ whose representations have reached the optimizer's steady state:

$$\mathbb{E}[\|\mathbf{f}(\mathbf{x}_1) - \mathbf{f}(\mathbf{x}_2)\|] \leq \frac{2\sqrt{K}M_W^2}{\gamma} + O\left(\sqrt{\frac{\sigma^2 P\eta}{\gamma}}\right) \tag{5}$$

regardless of $\|\mathbf{x}_1 - \mathbf{x}_2\|$ (proof in Appendix A). At fixed learning rate, the first term decreases as $O(1/\gamma)$ and the second as $O(1/\sqrt{\gamma})$, so the bound tightens monotonically with γ, as demonstrated empirically in Figure 2 (right). This uniform bound limits the sensitivity of network outputs to input perturbations.

Beyond confinement and collapse, the regularization induces a third geometric property in the TPT. Because the L_2 penalty acts purely in the radial direction, it does not influence the angular motion of the representations: the direction in which a representation points is driven entirely by the cross-entropy term. When classification confidence is high but the confinement prevents softmax saturation, this driving force aligns each representation with the corresponding classifier weight vector, and stronger regularization accelerates this alignment. The acceleration is visible as faster convergence of the NC property (Section 5); a formal statement and proof are given in Appendix A.

Table 1: Ablation study architectures. *Lin. Pen.*: linear penultimate-layer present. *Nodes Add. Layer*: dimensionality of the additional layer between backbone and classifier (exact values vary by dataset; see Appendix B). *Loss*: training objective.

Model	Lin. Pen.	Nodes Add. Layer	Loss
LPC	✓	Intermediate	$CE + L_2$
LPC-Wide	✓	Wide	$CE + L_2$
LPC-Narrow	✓	Narrow	$CE + L_2$
LPC-SCL	✓	Intermediate	$CE + L_2 + SCL$
LPC-NoPen	✗	✗	$CE + L_2$
LinPen	✓	Intermediate	CE
NonlinPen	✗	Intermediate	CE
SCL	✗	✗	$CE + SCL$
ArcFace	✗	✗	$ArcFace$
NoPenWD	✗	✗	CE
NoPen	✗	✗	CE

3 Information Bottleneck in Deterministic DNN Classifiers

The emergence of LPC creates an IB in the latent space, connecting this phenomenon to IB optimization in DNNs (Tishby & Zaslavsky, 2015; Alemi et al., 2017; Kolchinsky et al., 2019; Chalk et al., 2016; Achille & Soatto, 2017; Belghazi et al., 2018). Unlike many IB methods, which rely on variational approximations or noise injection, LPC implements a deterministic form of compression through a strong L_2 penalty on the features themselves, effectively shrinking their distribution and lowering its entropy. The optimization of the IB Lagrangian aims to maximize the following objective:

$$\mathcal{L}_{IB} = I(\boldsymbol{z}; y) - \beta I(\boldsymbol{z}; \boldsymbol{x}), \tag{6}$$

where $I(A; B)$ denotes the mutual information between two random variables A and B, and $\beta > 0$ controls the trade-off between compression and predictive accuracy. In a deterministic classifier both terms admit natural surrogates. Minimizing the cross-entropy loss maximizes a variational lower bound on $I(\boldsymbol{z}; y)$ (Alemi et al., 2017). Moreover, since $\boldsymbol{z}$ is a deterministic function of $\boldsymbol{x}$, $I(\boldsymbol{z}; \boldsymbol{x})$ reduces to the entropy of the representation itself ($H(\boldsymbol{z}|\boldsymbol{x}) = 0$ for discrete or quantized codes; for continuous representations this identification is formal, and we track the differential entropy of $\boldsymbol{z}$ as the compression measure). The corresponding surrogate objective is

$$\mathcal{L}_{\text{CE}}(\boldsymbol{f}(\boldsymbol{x}), \bar{y}) + \beta H(\boldsymbol{z}). \tag{7}$$

During training, the cross-entropy term is minimized explicitly, while $H(\boldsymbol{z})$ is reduced implicitly by the collapse of same-class representations. In the limit of exact collapse, $\boldsymbol{z}$ becomes a discrete variable supported on K points; for balanced classes the resulting code carries $\log K$ nats, the minimum information compatible with discriminating K classes, so the representation approaches a minimal sufficient statistic of $\boldsymbol{x}$ for y. For the continuous representations realized in practice, this compression manifests as a differential entropy far below that of unregularized baselines, which we quantify in Section 4.2.

4 Experiments

We empirically demonstrate that LPC regularization improves classifier performance through systematic ablation studies on CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), and ImageNet-1K (Deng et al., 2009) using ResNet (He et al., 2016) and Wide ResNet (Zagoruyko & Komodakis, 2017) backbones. Table 1 summarizes all architectures, which isolate the contributions of four factors relative to the NoPen baseline, which performs linear classification directly on the backbone output: L_2 regularization on representations (present only in LPC architectures), the presence and dimensionality of the linear penultimate layer (LPC-Narrow/LPC-Wide for dimensionality; LPC-NoPen for its absence), the addition of a penultimate layer

Table 2: Binary encoding metrics at the last training epoch. Score: average bimodal GMM log-likelihood $\overline{\ell}$ (mean $\pm$ std over runs). Peaks: relative peak separation $\overline{\mu}$ (mean $\pm$ std over runs). Min Peaks: worst-case peak separation across all runs and dimensions. CoV: coefficient of variation of representation norms. LPC-W and LPC-N denote LPC-Wide and LPC-Narrow.

Model	Score	Peaks	Min P.	CoV
CIFAR-10				
LPC	0.37 ± 0.26	24.9 ± 6.7	10.4	0.007
LPC-W	0.05 ± 0.18	18.1 ± 3.6	5.9	0.012
LPC-N	0.42 ± 0.32	26.1 ± 8.8	7.4	0.008
LinPen	-1.36 ± 0.05	1.9 ± 0.4	0.1	0.339
CIFAR-100				
LPC	0.65 ± 0.16	32.1 ± 5.8	23.3	0.005
LPC-W	0.52 ± 0.23	28.3 ± 6.6	9.1	0.021
LPC-N	0.82 ± 0.02	37.4 ± 0.8	29.2	0.005
LinPen	-1.42 ± 0.00	1.3 ± 0.1	0.5	0.258
ImageNet				
LPC	-0.47 ± 0.00	10.1 ± 0.1	9.5	0.202

without L_2 regularization, either linear or nonlinear (LinPen/NonlinPen), and loss function (supervised contrastive learning, SCL (Khosla et al., 2020); ArcFace (Deng et al., 2019); and the combined LPC-SCL). The NoPenWD variant differs from NoPen only in using elevated weight decay, and serves to verify that the benefits of LPC cannot be reproduced by stronger weight regularization alone. Full training details are in Appendix B.

4.1 Binary encoding

The collapse points align with vertices of a hypercube inscribed within a hypersphere. At each penultimate-layer node, latent representations approximately assume one of two values, forming a binary encoding. One possible explanation is that extreme compression forces the network to maximize relative distances between collapse points in each dimension to maintain discriminability, naturally leading to symmetric arrangements around the origin characteristic of hypercube vertices.

To verify this, we fit a bimodal Gaussian mixture model (GMM) on each dimension of the penultimate representations across the training set. For each dimension i, we obtain two modes with means $\mu_i^{(1,2)}$ and standard deviations $\sigma_i^{(1,2)}$. We then compute the average log-likelihood score $\overline{\ell}$ and the mean relative peak separation $\overline{\mu} = \frac{2}{P}\sum_i |\mu_i^{(2)} - \mu_i^{(1)}|/(\sigma_i^{(1)} + \sigma_i^{(2)})$, where P is the penultimate-layer dimensionality. Table 2 reports these quantities together with the worst-case peak separation observed across all experiments and the coefficient of variation of the representation norms.

For all LPC models, the bimodal fit produces high log-likelihood scores and large peak separations. Even in the worst case across all runs and dimensions, the minimum peak separation remains well above unity, confirming that the two modes are clearly resolved everywhere. The low coefficient of variation further confirms that representation norms cluster tightly around two values per node, supporting the hypothesis that class means collapse onto hypercube vertices. The same analysis applied to LinPen (which includes a linear penultimate layer but no L_2 penalty) shows no binary structure: low scores, small peak separations, and high coefficient of variation.

Table 3: All values in the table represent the means and standard deviations obtained from different experiments. Σ_W: mean across all entries of the within-class covariance matrix; Norm Mean: mean of the L_2 norm of latent representations; $\mathcal{R}$: class separation ratio, computed over a sample of 10,000 training samples; Avg. Grad Norm: average L_2 norm of the gradient of the logit-difference function $g_{\bar{y}}(\boldsymbol{x})$ with respect to the input $\boldsymbol{x}$, computed over the training set; H: estimated differential entropy on the testing set using the Kozachenko-Leonenko method (k=20), divided by the penultimate-layer dimension.

Dataset: CIFAR-10					
Model	Σ_W	Norm Mean	$\mathcal{R}$	Avg. Grad Norm	H
LPC	$2.55\text{E-}14 \pm 1.38\text{E-}14$	0.004 ± 0.001	636.05 ± 199.03	0.06 ± 0.03	-3.96 ± 0.31
LPC-Wide	$1.71\text{E-}14 \pm 5.88\text{E-}15$	0.003 ± 0.001	348.25 ± 148.10	0.15 ± 0.05	-3.4 ± 0.34
LPC-Narrow	$5.70\text{E-}14 \pm 1.05\text{E-}14$	0.003 ± 0.001	814.10 ± 245.34	0.04 ± 0.03	-4.13 ± 0.34
LPC-SCL	$9.89\text{E-}14 \pm 4.50\text{E-}14$	0.004 ± 0.001	288.88 ± 37.38	0.13 ± 0.04	-3.27 ± 0.11
LPC-NoPen	$1.33\text{E-}11 \pm 6.21\text{E-}12$	0.005 ± 0.001	55.43 ± 12.87	0.25 ± 0.26	-1.77 ± 0.33
LinPen	$1.10\text{E-}01 \pm 4.00\text{E-}02$	47.70 ± 7.89	3.18 ± 0.21	99.16 ± 10.61	0.58 ± 0.26
NonlinPen	$4.08\text{E-}01 \pm 1.42\text{E-}01$	39.39 ± 7.17	3.16 ± 0.14	62.06 ± 9.66	0.46 ± 0.02
SCL	$8.02\text{E-}03 \pm 4.13\text{E-}03$	25.42 ± 2.37	11.75 ± 0.36	44.09 ± 1.81	0.24 ± 0.4
ArcFace	$9.05\text{E-}02 \pm 2.32\text{E-}02$	22.81 ± 1.33	5.48 ± 0.20	78.91 ± 33.68	-0.06 ± 0.13
NoPen	$3.03\text{E-}02 \pm 1.08\text{E-}02$	13.88 ± 3.49	1.86 ± 0.11	84.55 ± 8.93	0.88 ± 0.05
NoPenWD	$2.32\text{E-}03 \pm 2.50\text{E-}04$	7.33 ± 0.19	2.81 ± 0.04	29.27 ± 1.69	0.59 ± 0.02
Dataset: CIFAR-100					
Model	Σ_W	Norm Mean	$\mathcal{R}$	Avg. Grad Norm	H
LPC	$2.62\text{E-}12 \pm 9.22\text{E-}13$	0.007 ± 0.001	327.63 ± 44.06	0.11 ± 0.02	-2.106 ± 0.119
LPC-Wide	$4.08\text{E-}12 \pm 4.13\text{E-}12$	0.008 ± 0.002	238.13 ± 53.61	0.22 ± 0.10	-1.914 ± 0.15
LPC-Narrow	$5.84\text{E-}12 \pm 3.37\text{E-}12$	0.006 ± 0.000	409.47 ± 19.60	0.06 ± 0.00	-2.319 ± 0.04
LPC-SCL	$3.94\text{E-}12 \pm 2.83\text{E-}12$	0.008 ± 0.001	172.19 ± 11.84	0.22 ± 0.01	-1.675 ± 0.089
LPC-NoPen	$3.39\text{E-}06 \pm 5.21\text{E-}06$	0.018 ± 0.018	21.79 ± 5.23	6.99 ± 10.90	5.559 ± 0.054
LinPen	$5.75\text{E-}01 \pm 1.34\text{E-}01$	74.28 ± 8.41	1.56 ± 0.02	86.82 ± 4.42	1.257 ± 0.022
NonlinPen	$3.50\text{E+}00 \pm 2.89\text{E-}01$	92.74 ± 1.87	1.65 ± 0.01	81.82 ± 2.32	1.208 ± 0.006
SCL	$2.87\text{E-}03 \pm 1.35\text{E-}03$	34.52 ± 5.34	8.43 ± 0.39	66.25 ± 9.16	5.123 ± 0.097
ArcFace	$4.40\text{E-}02 \pm 2.70\text{E-}03$	43.27 ± 1.16	4.80 ± 0.05	109.57 ± 11.50	5.055 ± 0.02
NoPen	$1.94\text{E-}02 \pm 1.44\text{E-}02$	25.08 ± 6.26	1.17 ± 0.02	70.81 ± 3.03	5.566 ± 0.029
NoPenWD	$8.48\text{E-}04 \pm 1.71\text{E-}04$	13.20 ± 0.69	1.42 ± 0.02	40.22 ± 1.7	5.529 ± 0.017
Dataset: ImageNet					
Model	Σ_W	Norm Mean	$\mathcal{R}$	Avg. Grad Norm	H
LPC	$4.18\text{E-}10 \pm 2.11\text{E-}11$	0.013 ± 0.000	12.96 ± 0.16	1.69 ± 0.03	1.321 ± 0.002
NoPen	$7.30\text{E-}03 \pm 2.00\text{E-}03$	29.14 ± 3.52	1.04 ± 0.00	11.93 ± 0.18	5.961 ± 0.002

4.2 Latent point collapse: properties and output sensitivity

To investigate the occurrence of LPC in the penultimate layer, we examine the within-class covariance $\boldsymbol{\Sigma}_W$, defined as

$$\boldsymbol{\Sigma}_W = \frac{1}{N} \sum_{p=1}^{K} \sum_{i=1}^{N_p} \left(\boldsymbol{z}^{(i,p)} - \boldsymbol{\mu}^{(p)} \right) \left(\boldsymbol{z}^{(i,p)} - \boldsymbol{\mu}^{(p)} \right)^\top , \tag{8}$$

where N_p denotes the number of training samples with label p, $\boldsymbol{z}^{(i,p)}$ the i-th latent representation with label p, and $\boldsymbol{\mu}^{(p)}$ the mean of all latent representations with label p. A vanishing $\boldsymbol{\Sigma}_W$ indicates that latent representations within each class collapse toward their respective class mean. We additionally analyze the mean representation norm $\frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{z}_i\|$. Since within-class variability vanishes under LPC, this quantity

coincides with the average distance of the collapse points from the origin, which our theoretical analysis predicts to be small.

Table 3 presents the values of Σ_W and the mean norm at the final training epoch across various architectures. Notably, Σ_W approaches zero exclusively for architectures implementing an L_2 penalty on the penultimate layer, confirming that same-class latent representations indeed collapse to single points in these models. The exceptionally small mean norm values across all LPC architectures corroborate our theoretical prediction that collapse points reside near the origin. We observe that strong weight decay alone, as employed in the NoPenWD architecture, does not induce collapse. Table 3 also reports the estimated differential entropy H of penultimate-layer representations using the Kozachenko-Leonenko method. LPC-based models exhibit significantly lower entropy values compared to their non-penalized counterparts, confirming that point collapse sharply compresses the distribution of the representations. This compression connects LPC to the IB principle, as discussed in Section 3.

Beyond the geometric properties of the latent space, we analyze the resulting classification characteristics, particularly class separability and decision boundary stability. To quantify separability in the latent space, we introduce a *class separation ratio* $\mathcal{R}$: $\mathcal{R}^{(i,p)} = \frac{\min_{q \neq p} \|\boldsymbol{z}^{(i,p)} - \mu^{(q)}\|}{\|\boldsymbol{z}^{(i,p)} - \mu^{(p)}\|}$. This ratio measures the distance between a latent point and the nearest other-class centroid relative to its distance from its own-class centroid. Higher values indicate superior geometric separation between classes. As demonstrated in Table 3, architectures implementing the L_2 penalty on a penultimate linear layer achieve substantially higher class separation ratios than all other methods: more than two orders of magnitude above the NoPen baseline on CIFAR (and more than an order of magnitude above the strongest alternative, SCL), and roughly one order of magnitude above the baseline on ImageNet.

To evaluate classification boundary stability, we analyze the network's sensitivity to input perturbations through the function $g_{\bar{y}}(\boldsymbol{x}) = f_{\bar{y}}(\boldsymbol{x}) - \max_{j \neq \bar{y}} f_j(\boldsymbol{x})$, where $f_{\bar{y}}(\boldsymbol{x})$ denotes the logit of the true class $\bar{y}$. This function captures the classification margin in the output space, with $g_{\bar{y}}(\boldsymbol{x}) > 0$ indicating correct classification.

We quantify network sensitivity by computing $\frac{1}{N} \sum_{i=1}^{N} \|\nabla_{\boldsymbol{x}} g_{\bar{y}}(\boldsymbol{x}_i)\|$ over the training set, which provides an empirical measure of the average local sensitivity of $g_{\bar{y}}(\boldsymbol{x})$ to input perturbations.

Table 3 reports this average gradient norm. Models employing the L_2 penalty on a linear penultimate layer exhibit average gradient norms two to three orders of magnitude lower than alternative configurations on CIFAR, and roughly seven times lower on ImageNet, indicating substantially reduced sensitivity to input perturbations and enhanced output margin stability. The LPC-NoPen configuration, which applies the L_2 penalty directly to the backbone output without a linear penultimate layer, sits between these two regimes: clearly below the baselines, but above the other LPC variants. This pattern indicates that the L_2 penalty alone reduces output sensitivity, and that the additional dimensionality reduction provided by a linear penultimate layer amplifies the effect. This reduced sensitivity, coupled with enhanced class separation, shows that LPC yields classifications that are geometrically well separated and exhibit bounded output sensitivity, providing improved guarantees on network behavior under perturbations.

4.3 Robustness and generalization

Table 4 presents the magnitude of minimal perturbations required to induce misclassification, quantified using the DeepFool algorithm (Moosavi-Dezfooli et al., 2016). Our results show a large improvement in network robustness when L_2 regularization is applied to the penultimate layer, with minimal adversarial perturbations up to two orders of magnitude larger than the baselines on CIFAR and one order of magnitude larger on ImageNet, particularly pronounced when the penultimate layer is linear. This robustness improvement extends to adversarial attacks: LPC models maintain non-trivial adversarial accuracy under projected gradient descent (PGD) attacks (Madry et al., 2018) (e.g., 14.6% at $\epsilon = 4/255$ on CIFAR-10, 6.5% at $\epsilon = 4/255$ on CIFAR-100, and 19.2% at $\epsilon = 1/255$ and 2.1% at $\epsilon = 2/255$ on ImageNet), while the standard baselines exhibit essentially zero robustness under comparable perturbation budgets. Among the alternative regularization techniques, SCL improves the DeepFool metric only marginally and retains near-zero PGD robustness; ArcFace attains non-trivial PGD accuracy on CIFAR-100 (10.5% at $\epsilon = 4/255$, exceeding the

Table 4: All values in the table represent the means and standard deviations obtained from different experiments. *DeepFool*: Norm of the minimal perturbation to cause a prediction change, divided by the norm of the input (Moosavi-Dezfooli et al., 2016), averaged over 1000 test set samples. *PGD columns*: Adversarial accuracy under PGD attacks (Madry et al., 2018) with $\epsilon \in \{4/255, 8/255, 12/255\}$ for CIFAR and $\epsilon \in \{1/255, 2/255\}$ for ImageNet, evaluated on 1000 test set samples. PGD uses 100 iterations (50 for ImageNet) with 5 random restarts, step size $\alpha = \epsilon/4$, DLR loss, cosine schedule, and ℓ_∞ norm constraint. *Accuracy*: Classification accuracy on the testing set. For ImageNet, the generalization gap (Gen. Gap) is also included, representing the difference between training and testing accuracy.

Dataset: CIFAR-10					
Model	DeepFool	PGD ϵ=4/255	PGD ϵ=8/255	PGD ϵ=12/255	Accuracy
LPC	1.272 ± 0.406	0.146 ± 0.021	0.034 ± 0.011	$\mathbf{0.013} \pm 0.007$	94.86 ± 0.08
LPC-Wide	0.732 ± 0.409	0.180 ± 0.059	0.033 ± 0.025	0.010 ± 0.008	94.73 ± 0.04
LPC-Narrow	$\mathbf{1.542} \pm 0.411$	0.153 ± 0.029	$\mathbf{0.040} \pm 0.018$	$\mathbf{0.013} \pm 0.006$	94.90 ± 0.13
LPC-SCL	0.519 ± 0.053	0.170 ± 0.025	0.034 ± 0.007	0.009 ± 0.004	$\mathbf{94.91} \pm 0.10$
LPC-NoPen	0.737 ± 0.263	$\mathbf{0.195} \pm 0.034$	0.014 ± 0.006	0.001 ± 0.002	94.86 ± 0.09
LinPen	0.013 ± 0.001	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	94.58 ± 0.08
NonlinPen	0.015 ± 0.001	0.001 ± 0.001	0.000 ± 0.000	0.000 ± 0.000	94.50 ± 0.05
SCL	0.025 ± 0.002	0.012 ± 0.009	0.001 ± 0.001	0.000 ± 0.000	94.77 ± 0.11
ArcFace	0.019 ± 0.001	0.031 ± 0.015	0.000 ± 0.001	0.000 ± 0.000	94.54 ± 0.08
NoPen	0.013 ± 0.001	0.000 ± 0.001	0.000 ± 0.000	0.000 ± 0.000	94.57 ± 0.14
NoPenWD	0.016 ± 0.001	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	94.25 ± 0.04

Dataset: CIFAR-100					
Model	DeepFool	PGD ϵ=4/255	PGD ϵ=8/255	PGD ϵ=12/255	Accuracy
LPC	0.421 ± 0.047	0.065 ± 0.019	0.013 ± 0.007	0.004 ± 0.002	77.64 ± 0.17
LPC-Wide	0.377 ± 0.020	0.076 ± 0.019	0.015 ± 0.004	0.002 ± 0.002	77.75 ± 0.24
LPC-Narrow	$\mathbf{0.459} \pm 0.031$	0.040 ± 0.005	0.004 ± 0.002	0.000 ± 0.000	77.29 ± 0.18
LPC-SCL	0.183 ± 0.027	0.049 ± 0.005	0.007 ± 0.001	0.003 ± 0.001	$\mathbf{78.17} \pm 0.23$
LPC-NoPen	0.136 ± 0.044	$\mathbf{0.135} \pm 0.049$	$\mathbf{0.034} \pm 0.018$	$\mathbf{0.008} \pm 0.005$	77.27 ± 0.36
LinPen	0.007 ± 0.000	0.001 ± 0.001	0.000 ± 0.000	0.000 ± 0.000	76.69 ± 0.17
NonlinPen	0.007 ± 0.000	0.000 ± 0.001	0.000 ± 0.000	0.000 ± 0.000	76.38 ± 0.30
SCL	0.011 ± 0.000	0.004 ± 0.002	0.000 ± 0.000	0.000 ± 0.000	78.00 ± 0.15
ArcFace	0.012 ± 0.000	0.105 ± 0.023	0.013 ± 0.004	0.001 ± 0.002	77.21 ± 0.14
NoPen	0.007 ± 0.000	0.001 ± 0.001	0.000 ± 0.000	0.000 ± 0.000	76.83 ± 0.15
NoPenWD	0.007 ± 0.000	0.001 ± 0.001	0.000 ± 0.000	0.000 ± 0.000	76.97 ± 0.14

Dataset: ImageNet					
Model	DeepFool	PGD ϵ=1/255	PGD ϵ=2/255	Accuracy	Gen. Gap
LPC	$\mathbf{0.020} \pm 0.001$	$\mathbf{0.192} \pm 0.015$	$\mathbf{0.021} \pm 0.013$	$\mathbf{75.63} \pm 0.20$	$\mathbf{15.60} \pm 0.20$
NoPen	0.002 ± 0.000	0.001 ± 0.002	0.001 ± 0.002	74.21 ± 0.09	22.53 ± 0.15

6.5% of LPC, though below the 13.5% of LPC-NoPen) but achieves only 3.1% on CIFAR-10 and, like all baselines, remains one to two orders of magnitude below the LPC models on the minimal-perturbation (DeepFool) metric.

LPC also improves generalization, achieving performance comparable to or exceeding established regularization methods such as SCL. The combination of LPC with SCL yields the highest generalization performance on both CIFAR-10 and CIFAR-100, showing that LPC can be combined with existing techniques for additional gains. The generalization improvements are particularly pronounced on ImageNet, where LPC reduces

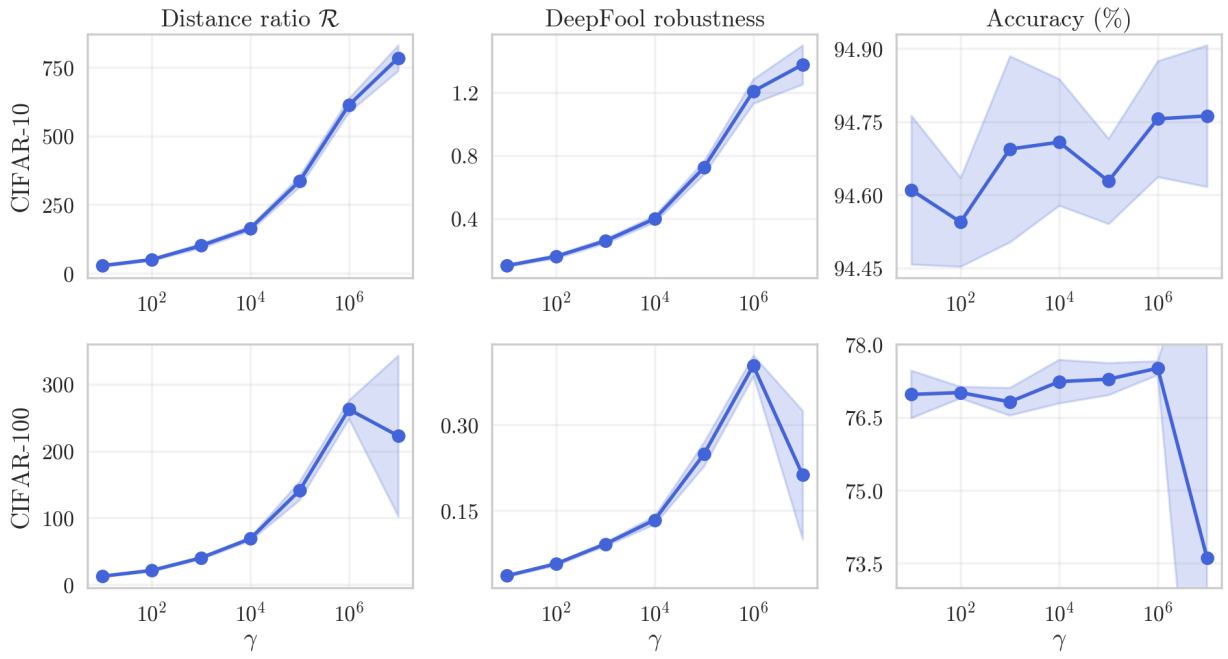

Figure 3: Effect of LPC regularization strength on classification performance and robustness. The figure shows accuracy, DeepFool robustness, and class separation ratio $\mathcal{R}$ as functions of the LPC regularization parameter γ for CIFAR-10 (top row) and CIFAR-100 (bottom row) datasets. Results are averaged across multiple independent runs with error bars indicating standard deviation. As γ increases from 10^1 to 10^6, both geometric margins and robustness improve significantly while classification accuracy remains stable. CIFAR-10 maintains stable accuracy across the entire range ($\gamma \leq 10^7$), while CIFAR-100 shows degradation emerging only at the extreme value $\gamma = 10^7$.

the generalization gap from 22.53% to 15.60%, a relative reduction of over 30%. These improvements are stable across a range of regularization strengths (Figure 3).

Our architectural analysis further reveals practical insights regarding the penultimate-layer design. Specifically, we observe that lower-dimensional penultimate layers yield substantially superior robustness as measured by DeepFool, with the LPC-Narrow configuration achieving the highest scores across both CIFAR datasets. Interestingly, adversarial robustness under PGD attacks exhibits a different pattern: the LPC-NoPen variant attains the highest adversarial accuracy at $\epsilon = 4/255$ on both CIFAR datasets and across all budgets on CIFAR-100, although on CIFAR-10 it becomes the weakest LPC variant at larger budgets.

5 Neural Collapse

We evaluate LPC against the first two properties of NC identified by Papyan et al. (2020): (NC1) within-class variability converges to zero, and (NC2) class means form an ETFS. We adopt the same metrics as Papyan et al. (2020) for NC2 (equinorm, equiangularity, and maximal-angle deviation from ETFS) and measure NC1 via $\mathrm{Tr}(\boldsymbol{\Sigma}_W \boldsymbol{\Sigma}_B^{\dagger}/K)$, where $\boldsymbol{\Sigma}_W$ is defined in Section 4.2 and $\boldsymbol{\Sigma}_B$ is the between-class covariance of the centered class means. Because our architectures use penultimate layers of different dimensionality, we report a renormalized version $\mathrm{Tr}(\boldsymbol{\Sigma}_W \boldsymbol{\Sigma}_B^{\dagger}/K/P)$, dividing by the number of penultimate nodes P. The dashed vertical lines in Figure 4 mark the epoch at which each architecture enters the TPT, as defined in Appendix B.

Figure 4 (top) shows that the renormalized NC1 value is orders of magnitude lower in LPC architectures than in baselines. SCL and ArcFace also improve over the baseline, but their gains are smaller than those

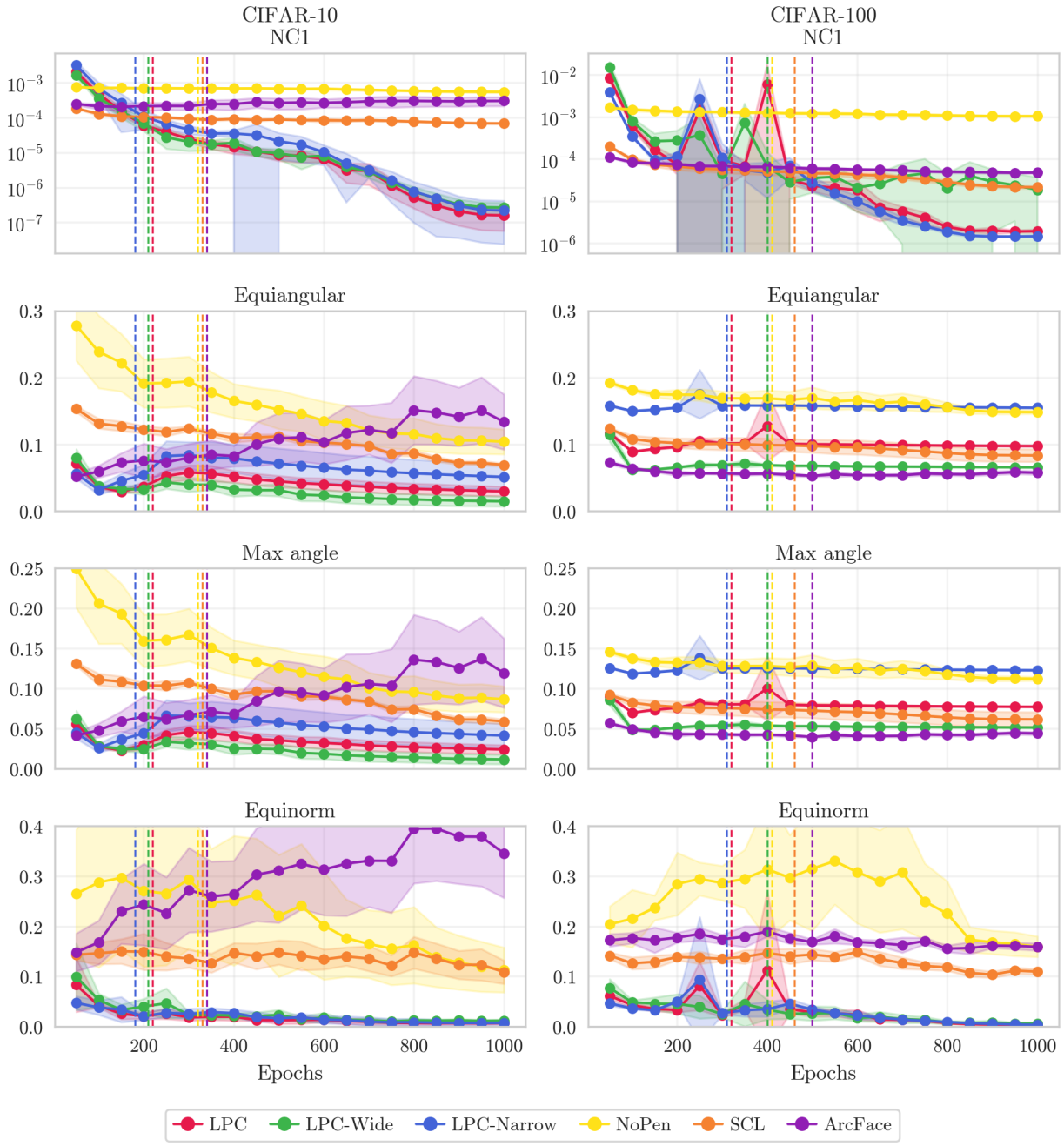

Figure 4: Convergence towards Neural Collapse. Top: renormalized NC1 metric $\mathrm{Tr}(\boldsymbol{\Sigma}_W \boldsymbol{\Sigma}_B^{\dagger}/K/P)$, where division by the penultimate-layer dimension P ensures fair comparison across architectures. Dashed lines mark the epoch at which each architecture enters the TPT (see Appendix B), averaged across runs. Bottom three panels: NC2 metrics (equinorm, equiangularity, maximal-angle deviation from ETFS), computed as in Papyan et al. (2020).

of LPC. The bottom three panels show convergence of class means toward an ETFS. All metrics plateau in the TPT, confirming that NC is fully realized. The LPC-Narrow architecture exhibits slightly elevated angular metrics (equiangularity and maximal-angle), reflecting the geometric difficulty of inscribing an ETFS within a low-dimensional hypercube.

Since the networks enter the TPT well before the end of training, the improvements documented throughout this paper are in addition to those typically associated with standard NC.

6 Discussion and Limitations

6.1 Discussion

The LPC phenomenon is associated with strong empirical class separability and large improvements in robustness, going beyond the collapse predicted by traditional NC. While NC ensures that within-class representations converge around their class means (achieving small within-class variance relative to between-class variance), LPC enforces absolute collapse to single points. We demonstrate particular effectiveness when applying this regularization to low-dimensional linear penultimate layers, where the dimensional bottleneck amplifies both the collapse effect and its benefits. A key distinction is that our approach provides a provable uniform bound on output variation without architectural constraints, a property unattainable through standard NC.

Regarding accuracy, we note that LPC consistently improves classification performance over the corresponding baselines, with gains typically larger than the variation across runs. In absolute terms, however, the improvements on CIFAR amount to only a fraction of a percentage point. On ImageNet the gain is more substantial, with LPC raising accuracy by 1.4 percentage points and reducing the generalization gap by roughly 30%. We emphasize that accuracy is not the principal claim of this work. With the same backbone and training budget, the addition of a linear penultimate layer and the L_2 penalty yields small but consistent accuracy gains, while improving adversarial robustness by orders of magnitude and producing the geometric and information-theoretic properties characterized in Sections 2 and 3. The contribution of LPC lies in this combination of properties.

While our CIFAR-10 and CIFAR-100 experiments operated largely in the TPT, ImageNet training did not reach full convergence of training-set accuracy within the training budget. Even so, LPC yielded superior class separation, reduced gradient norms, and a substantially smaller generalization gap on ImageNet (a relative improvement of about 30%), indicating that the geometric benefits of the method develop progressively during optimization and do not require extended training in the TPT.

6.2 Limitations and outlook

While we demonstrate LPC under cross-entropy loss with L_2 regularization, NC has been observed with various loss functions (Han et al., 2022; Zhou et al., 2022b). Exploring alternative loss formulations that might induce or enhance LPC could reveal more efficient training procedures or stronger guarantees.

We note that LPC is not designed as a substitute for adversarial training. The robustness gains reported here are relative to standard training and other non-adversarial regularization methods. The combination of LPC with adversarial training is a natural direction for future work.

Our analysis primarily examines geometric and robustness metrics established in the original NC literature (Papyan et al., 2020). However, NC enhances numerous other properties including knowledge distillation (Zhang et al., 2025), few-shot learning (Yang et al., 2023), transfer learning (Galanti et al., 2022a;b), and out-of-distribution detection (Haas et al., 2023; Ammar et al., 2024). The stronger absolute collapse of LPC may amplify these benefits, presenting promising avenues for investigation.

The empirically observed binary encoding phenomenon (Section 4.1), wherein collapse points align precisely with hypercube vertices inscribed within a hypersphere, is an intriguing finding. Under extreme L_2 regularization, the network spontaneously organizes class representations into a discrete binary scheme where each penultimate node assumes one of two values, effectively creating a $\{-1, +1\}^d$ encoding. Understanding

this spontaneous quantization could bridge connections to binary neural networks and discrete representation learning, though the precise mechanisms driving this phenomenon remain an open theoretical question warranting further investigation.

6.3 Conclusion

We introduce *latent point collapse*, a phenomenon where strong L_2 regularization applied to penultimate-layer representations causes the features within each class to collapse to a single point near the origin. This simple modification, i.e. adding a regularized, low-dimensional linear penultimate layer, strongly improves model robustness while yielding consistent gains in generalization.

The method's simplicity and effectiveness make it immediately practical. By inducing a stronger form of *neural collapse* with provable bounds on output sensitivity, LPC provides principled robustness guarantees while maintaining or improving classification accuracy. The early emergence of benefits, demonstrated in our ImageNet experiments, indicates that LPC's advantages are accessible even without complete training convergence, which makes the approach useful in resource-constrained settings.

References

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation, 2017. URL https://arxiv.org/abs/1611.01353.

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9, 2018. doi: 10.1109/ITA.2018.8503149.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyxQzBceg.

Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. NECO: NEural collapse based out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=9ROuKblmi7.

George Andriopoulos, Zixuan Dong, Li Guo, Zifan Zhao, and Keith W. Ross. The prevalence of neural collapse in neural multivariate regression. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=Wq6aY6fC2H.

Cem Anil, James Lucas, and Roger Baker Grosse. Sorting out lipschitz function approximation. In *International Conference on Machine Learning*, 2018. URL https://api.semanticscholar.org/CorpusID:53292059.

Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, R. Devon Hjelm, and Aaron C. Courville. Mine: Mutual information neural estimation. *ArXiv*, abs/1801.04062, 2018. URL https://api.semanticscholar.org/CorpusID:3614357.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.

Matthew Chalk, Olivier Marre, and Gašper Tkačik. Relevant sparse codes with variational information bottleneck. In *Neural Information Processing Systems*, 2016. URL https://api.semanticscholar.org/CorpusID:6019591.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, Proceedings of Machine Learning Research, pp. 1597–1607. PMLR, 2020. URL http://proceedings.mlr.press/v119/chen20j.html.

Zhikang Chen, Min Zhang, Sen Cui, Haoxuan Li, Gang Niu, Mingming Gong, Changshui Zhang, and Kun Zhang. Neural collapse inspired feature alignment for out-of-distribution generalization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=wQpNG9JnPK.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 854–863. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/cisse17a.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2019. doi: 10.1109/CVPR.2019.00482.

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.

Gamaleldin F. Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 850–860, Red Hook, NY, USA, 2018. Curran Associates Inc.

Tolga Ergen and Mert Pilanci. Convex duality of deep neural networks. *CoRR*, abs/2002.09773, 2020. URL https://arxiv.org/abs/2002.09773.

Cong Fang, Hangfeng He, Qi Long, and Weijie J. Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118 (43), October 2021. ISSN 1091-6490. doi: 10.1073/pnas.2103091118. URL http://dx.doi.org/10.1073/pnas.2103091118.

Quinn LeBlanc Fisher, Haoming Meng, and Vardan Papyan. Pushing boundaries: Mixup's influence on neural collapse. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jTSKkcbEsj.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2022a. URL https://openreview.net/forum?id=SwIp410B6aQ.

Tomer Galanti, András György, and Marcus Hutter. Improved generalization bounds for transfer learning via neural collapse. In *First Workshop on Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022*, 2022b. URL https://openreview.net/forum?id=VrK7pKwOhT_.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J. Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110:393 – 416, 2018. URL https://api.semanticscholar.org/CorpusID:4811672.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3821–3830. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/graf21a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, pp. 1321–1330, 2017.

Jarrod Haas, William Yolland, and Bernhard Rabus. Linking neural collapse and l2 normalization with improved out-of-distribution detection in deep neural networks, 2023.

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.

Md Yousuf Harun, Jhair Gallardo, and Christopher Kanan. Controlling neural collapse enhances out-of-distribution detection and transfer learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=8AGdUCdDyI.

Hangfeng He and Weijie J. Su. A law of data separation in deep learning. *Proceedings of the National Academy of Sciences*, 120(36):e2221704120, 2023. doi: 10.1073/pnas.2221704120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2221704120.

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

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735, 2019. URL https://api.semanticscholar.org/CorpusID:207930212.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41–50, 2019.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bklr3j0cKX.

Wanli Hong and Shuyang Ling. Neural collapse for unconstrained feature model under cross-entropy loss with imbalanced data. *Journal of Machine Learning Research*, 25(192):1–48, 2024. URL http://jmlr.org/papers/v25/23-1215.html.

Shizhe Hu, Zhengzheng Lou, Xiaoqiang Yan, and Yangdong Ye. A survey on information bottleneck. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5325–5344, 2024. doi: 10.1109/TPAMI.2024.3366349.

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding, 2020. URL https://arxiv.org/abs/1905.09272.

Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J. Su. An unconstrained layer-peeled perspective on neural collapse. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=WZ3yjh8coDg.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Artemy Kolchinsky, Brendan D. Tracey, and David H. Wolpert. Nonlinear information bottleneck. *Entropy*, 21(12):1181, November 2019. ISSN 1099-4300. doi: 10.3390/e21121181. URL http://dx.doi.org/10.3390/e21121181.

Vignesh Kothapalli, Tom Tirer, and Joan Bruna. A neural collapse perspective on feature evolution in graph neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=sxao2udWXi.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *arXiv preprint*, 2009.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-Supervised Nets. In Guy Lebanon and S. V. N. Vishwanathan (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 562–570, San Diego, California, USA, 09–12 May 2015. PMLR. URL https://proceedings.mlr.press/v38/lee15a.html.

Ismael Lemhadri, Feng Ruan, Louis Abraham, and Robert Tibshirani. Lassonet: A neural network with feature sparsity. *Journal of Machine Learning Research*, 22(127):1–29, 2021.

Renjie Li, Xinyi Wang, Guan Huang, Wenli Yang, Kaining Zhang, Xiaotong Gu, Son N. Tran, Saurabh Garg, Jane Alty, and Quan Bai. A comprehensive review on deep supervision: Theories and applications, 2022.

Xiao Li, Sheng Liu, Jinxin Zhou, Xinyu Lu, Carlos Fernandez-Granda, Zhihui Zhu, and Qing Qu. Understanding and improving transfer learning of deep models via neural collapse, 2024. URL https://arxiv.org/abs/2212.12206.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 507–516. JMLR.org, 2016.

Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6738–6746, 2017. URL https://api.semanticscholar.org/CorpusID:206596594.

Weiyang Liu, Longhui Yu, Adrian Weller, and Bernhard Schölkopf. Generalizing and decoupling neural collapse via hyperspherical uniformity gap. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=inU2quhGdNU.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through L_0 regularization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1Y8hhg0b.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Wojciech Masarczyk, Mateusz Ostaszewski, Ehsan Imani, Razvan Pascanu, Piotr Miłoś, and Tomasz Trzcinski. The tunnel effect: Building data representations in deep neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=xQOHOpe1Fv.

Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *CoRR*, abs/2011.11619, 2020. URL https://arxiv.org/abs/2011.11619.

Takeru Miyato, Shin ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1979–1993, 2017. URL https://api.semanticscholar.org/CorpusID:17504174.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks, 2016.

Michael Munn, Benoit Dherin, and Javier Gonzalvo. The impact of geometric complexity on neural collapse in transfer learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=PLbFid00aU.

Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study, 2018. URL https://arxiv.org/abs/1802.08760.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Jun Zhu, and Hang Su. Boosting adversarial training with hypersphere embedding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 7779–7792, 2020.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *PNAS*, 117:24652–24663, 2020.

Liam Parker, Emre Onal, Anton Stengel, and Jake Intrater. Neural collapse in the intermediate hidden layers of classification neural networks, 2023. URL https://arxiv.org/abs/2308.02760.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions, 2017.

Akshay Rangamani, Marius Lindegaard, Tomer Galanti, and Tomaso A Poggio. Feature learning in deep classifiers through intermediate neural collapse. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28729–28745. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/rangamani23a.html.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information, 2017. URL https://arxiv.org/abs/1703.00810.

Sahil Singla and Soheil Feizi. Skew orthogonal convolutions, 2021. URL https://arxiv.org/abs/2105.11417.

Sahil Singla, Surbhi Singla, and Soheil Feizi. Improved deterministic l2 robustness on CIFAR-10 and CIFAR-100. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=tD7eCtaSkR.

Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, August 2017. ISSN 1941-0476. doi: 10.1109/tsp.2017.2708039. URL http://dx.doi.org/10.1109/TSP.2017.2708039.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56): 1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Peter Súkeník, Marco Mondelli, and Christoph H Lampert. Deep neural collapse is provably optimal for the deep unconstrained features model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=v9yC7sSXf3.

Peter Súkeník, Christoph H. Lampert, and Marco Mondelli. Neural collapse vs. low-rank bias: Is deep neural collapse really optimal? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0jld45XGgJ.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014.

Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=utahaTbcHdP.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*, pp. 776–794, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58620-1. doi: 10.1007/978-3-030-58621-8_45. URL https://doi.org/10.1007/978-3-030-58621-8_45.

Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21478–21505. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tirer22a.html.

Tom Tirer, Haoxiang Huang, and Jonathan Niles-Weed. Perturbation analysis of neural collapse. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:253237344.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5. IEEE, 2015.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method, 2000. URL https://arxiv.org/abs/physics/0004057.

Asher Trockman and J. Zico Kolter. Orthogonalizing convolutional layers with the cayley transform, 2021. URL https://arxiv.org/abs/2104.07167.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Neural Information Processing Systems*, 2018. URL https://api.semanticscholar.org/CorpusID:3655212.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

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 *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 5265–5274. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00552. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Wang_CosFace_Large_Margin_CVPR_2018_paper.html.

Zijian Wang, Yadan Luo, Liang Zheng, Zi Huang, and Mahsa Baktashmotlagh. How far pre-trained models are from neural collapse on the target dataset informs their transferability. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5526–5535, 2023. doi: 10.1109/ICCV51070.2023.00511.

Colin Wei and Tengyu Ma. Improved sample complexities for deep neural networks and robust classification via an all-layer margin. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJe_yR4Fwr.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 499–515, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46478-7.

Diyuan Wu and Marco Mondelli. Neural collapse beyond the unconstrained features model: Landscape, dynamics, and generalization in the mean-field regime. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ZrhGq664om.

Robert Wu and Vardan Papyan. Linguistic collapse: Neural collapse in (large) language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=G0LfcMiRkc.

Yingwen Wu, Ruiji Yu, Xinwen Cheng, Zhengbao He, and Xiaolin Huang. Pursuing feature separation based on neural collapse for out-of-distribution detection. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=mUXdysoxEP.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018. URL https://api.semanticscholar.org/CorpusID:4591284.

Huan Xu and Shie Mannor. Robustness and generalization, 2010. URL https://arxiv.org/abs/1005.2243.

Hongren Yan, Yuhua Qian, Furong Peng, Jiachen Luo, zheqing zhu, and Feijiang Li. Neural collapse to multiple centers for imbalanced data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=RJEC9fZ9Ma.

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 Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=A6EmxI3_Xc.

Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=y5W8tpojhtJ.

Can Yaras, Peng Wang, Zhihui Zhu, Laura Balzano, and Qing Qu. Neural collapse with normalized features: A geometric analysis over the riemannian manifold. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Zvh6lF5b26N.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks, 2017. URL https://arxiv.org/abs/1605.07146.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy, 2019.

Shuoxi Zhang, Zijian Song, and Kun He. Neural collapse inspired knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(21):22542–22550, Apr. 2025. doi: 10.1609/aaai.v39i21.34412. URL https://ojs.aaai.org/index.php/AAAI/article/view/34412.

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 Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27179–27202. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/zhou22c.html.

Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=8rfYWE3nyXl.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29820–29834. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf.

A Theoretical Analysis of Latent Point Collapse

This Appendix includes a theoretical analysis of how strong L_2 regularization induces LPC in DNN classifiers, establishing both its necessity for achieving theoretical optima and its sufficiency for ensuring bounded output sensitivity. We employ the UFM framework to demonstrate that a large regularization parameter γ fundamentally reshapes the optimization landscape and dynamics of penultimate-layer representations.

Our analysis proceeds through a characterization of the interplay between cross-entropy loss and L_2 regularization. We first establish equilibrium conditions (Section A.2), proving that regularization confines representations to a ball of radius $O(M_W/\gamma)$, where M_W bounds the classifier weights. We then analyze the Hessian structure (Section A.3) to show that the feature-space loss is 2γ-strongly convex for every $\gamma > 0$, with curvature dominated by the isotropic regularization term when $\gamma \gg KM_W^2$. This strong convexity enables us to derive tight bounds on the steady-state distribution of representations, establishing that same-class features concentrate within radius $r_{\text{collapse}} = O(\sqrt{\sigma^2 d\eta/\gamma})$ of their class-specific equilibria, whose size reflects a balance between the noise level σ, the dimensionality of the latent space d, the magnitude of learning rate η, and the strength of the confining potential γ.

The geometric implications of this confinement are explored through our alignment analysis (Section A.4), where we show that representations not only collapse but also align with their corresponding classifier weight vectors in the terminal phase of training. The L_2 gradient is purely radial and therefore does not enter the angular dynamics; angular evolution is driven entirely by the cross-entropy term, which in the terminal phase aligns each representation with its classifier weight without requiring explicit geometric constraints.

These results lead to our main theorem establishing that strong L_2 regularization applied specifically to the penultimate layer induces a stronger manifestation of NC characterized by simultaneous confinement around the origin and class-specific point collapse. We show that this mechanism yields a uniform bound on output variation that is independent of input distance, a property that emerges from the bounded collapse around the origin.

A.1 Setup and Preliminaries

We analyze penultimate-layer representations using the unconstrained feature model (UFM) framework (Zhu et al., 2021; Fang et al., 2021; Tirer & Bruna, 2022). In this framework, features after the penultimate layer are treated as free optimization variables $\boldsymbol{z} \in \mathbb{R}^d$ (throughout the appendix, d denotes the penultimate-layer dimensionality, written P in the main text), allowing us to study their optimization dynamics independently. The final linear classifier has weights $\boldsymbol{W} \in \mathbb{R}^{K\times d}$ and biases $\boldsymbol{b} \in \mathbb{R}^K$, yielding logits $\boldsymbol{l} = \boldsymbol{W}\boldsymbol{z} + \boldsymbol{b}$.

Previous work (Zhu et al., 2021; Fang et al., 2021) established that for balanced datasets with bounded representations and classifier weights, global minimizers of cross-entropy loss exhibit NC. This phenomenon involves three key properties: (1) all features within a class converge to a single point, (2) the K class means form an ETFS, and (3) these means align with their corresponding classifier weights. However, reaching such global optima in deep networks with millions of parameters is practically infeasible (Tirer et al., 2022).

Our analysis investigates how strong L_2 regularization with coefficient γ drives the system toward this theoretical ideal. We analyze the combined loss:

$$\mathcal{L}(\boldsymbol{z}) = \mathcal{L}_{\text{CE}}(\boldsymbol{z}) + \gamma\|\boldsymbol{z}\|^2 \tag{9}$$

where $\mathcal{L}_{\text{CE}}(\boldsymbol{z}) = -\log p_{\bar{y}}(\boldsymbol{z})$ is the cross-entropy loss with softmax probabilities $p_i(\boldsymbol{z}) = \frac{\exp(\ell_i)}{\sum_{j=1}^K \exp(\ell_j)}$, and $\gamma > 0$ is the regularization strength.

Assumption 1 (Bounded Classifier Weights)**.** The classifier weight vectors satisfy $\|\boldsymbol{w}_i\| \leq M_W$ for all classes $i \in [K]$, where $M_W < \infty$. This mild condition holds under standard training practices with weight decay, which keeps classifier weights bounded.

We consider a general class of stochastic optimizers that update representations according to:

$$\boldsymbol{z}_{t+1} = \boldsymbol{z}_t - \eta_t \boldsymbol{P}_t(\nabla_{\boldsymbol{z}} \mathcal{L}(\boldsymbol{z}_t) + \mathbf{a}_t) \tag{10}$$

where η_t is the learning rate, $\boldsymbol{P}_t$ is a preconditioning matrix (identity for SGD, diagonal adaptive for Adam/AdamW), and $\mathbf{a}_t$ is gradient noise from mini-batch sampling with $\mathbb{E}[\mathbf{a}_t|\boldsymbol{z}_t] = \mathbf{0}$.

Assumption 2 (Optimizer Properties)**.** The optimizer satisfies:

1. **Bounded preconditioning:** $\lambda_{\min}\boldsymbol{I} \preceq \boldsymbol{P}_t \preceq \lambda_{\max}\boldsymbol{I}$ for some $0 < \lambda_{\min} \leq \lambda_{\max} < \infty$

2. **Bounded noise:** The noise covariance matrix is bounded: $\mathbb{E}[\mathbf{a}_t\mathbf{a}_t^T|\boldsymbol{z}_t] \preceq \sigma^2\boldsymbol{I}$ for some constant $\sigma^2 > 0$. This implies $\mathbb{E}[\|\mathbf{a}_t\|^2|\boldsymbol{z}_t] \leq d\sigma^2$.

3. **Convergence:** The learning rate schedule ensures convergence to a neighborhood of local minima

These assumptions are mild and satisfied by standard optimizers (SGD, Adam, AdamW) with appropriate hyperparameters. We will leverage these properties throughout our analysis to establish bounds on the behavior of representations under stochastic optimization.

A.2 Equilibrium Analysis

We first characterize the equilibrium points that arise from the interplay between cross-entropy loss and L_2 regularization.

Definition 1 (Stochastic Equilibrium)**.** A point $\boldsymbol{z}^*$ is a stochastic equilibrium if $\mathbb{E}[\boldsymbol{z}_{t+1}|\boldsymbol{z}_t = \boldsymbol{z}^*] = \boldsymbol{z}^*$ under the update rule equation 10.

For optimizers with positive definite preconditioning matrices (Assumption 2), this condition requires $\nabla_{\boldsymbol{z}}\mathcal{L}(\boldsymbol{z}^*) = \mathbf{0}$.

At a stochastic equilibrium $\boldsymbol{z}^*$, for a sample with true label $\bar{y}$, the gradient vanishes:

$$\nabla_{\boldsymbol{z}}\mathcal{L}(\boldsymbol{z}^*) = \sum_{i=1}^{K} p_i(\boldsymbol{z}^*)\boldsymbol{w}_i - \boldsymbol{w}_{\bar{y}} + 2\gamma\boldsymbol{z}^* = \mathbf{0} \tag{11}$$

Rearranging yields an explicit characterization:

$$\boldsymbol{z}^* = \frac{1}{2\gamma}\left(\boldsymbol{w}_{\bar{y}} - \sum_{i=1}^{K} p_i(\boldsymbol{z}^*)\boldsymbol{w}_i\right) \tag{12}$$

This equation shows that the equilibrium position is controlled by γ: larger values of γ yield equilibria closer to the origin.

Theorem 1 (Bounded Equilibrium)**.** *Any stochastic equilibrium $\boldsymbol{z}^*$ satisfies:*

$$\|\boldsymbol{z}^*\| \leq \frac{M_W}{\gamma} \tag{13}$$

Proof. From Equation 12, taking norms:

$$2\gamma\|\boldsymbol{z}^*\| = \left\|\boldsymbol{w}_{\bar{y}} - \sum_{i=1}^{K} p_i(\boldsymbol{z}^*)\boldsymbol{w}_i\right\| \tag{14}$$

$$= \left\|(1 - p_{\bar{y}}(\boldsymbol{z}^*))\boldsymbol{w}_{\bar{y}} - \sum_{i\neq\bar{y}} p_i(\boldsymbol{z}^*)\boldsymbol{w}_i\right\| \tag{15}$$

$$\leq (1 - p_{\bar{y}}(\boldsymbol{z}^*))\|\boldsymbol{w}_{\bar{y}}\| + \sum_{i\neq\bar{y}} p_i(\boldsymbol{z}^*)\|\boldsymbol{w}_i\| \tag{16}$$

$$\leq M_W\left[(1 - p_{\bar{y}}(\boldsymbol{z}^*)) + \sum_{i\neq\bar{y}} p_i(\boldsymbol{z}^*)\right] \tag{17}$$

$$= 2M_W(1 - p_{\bar{y}}(\boldsymbol{z}^*)) \tag{18}$$

The general bound follows since $1 - p_{\bar{y}}(\boldsymbol{z}^*) \leq 1$. □

This theorem establishes that L_2 regularization creates a confining potential that prevents representation explosion, with the confinement radius inversely proportional to γ.

A.3 Hessian Analysis and Stochastic Dynamics

To understand how representations behave near equilibrium under stochastic optimization, we analyze the local curvature of the loss landscape. The Hessian of the combined loss is:

$$\nabla^2_{\boldsymbol{z}}\mathcal{L}(\boldsymbol{z}) = \nabla^2_{\boldsymbol{z}}\mathcal{L}_{\text{CE}}(\boldsymbol{z}) + \nabla^2_{\boldsymbol{z}}(\gamma\|\boldsymbol{z}\|^2) \tag{19}$$

The Hessian of the L_2 regularization term is:

$$\nabla^2_{\boldsymbol{z}}(\gamma\|\boldsymbol{z}\|^2) = 2\gamma\boldsymbol{I} \tag{20}$$

The cross-entropy Hessian takes the form:

$$\nabla^2_{\boldsymbol{z}}\mathcal{L}_{\text{CE}}(\boldsymbol{z}) = \boldsymbol{W}^T(\text{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T)\boldsymbol{W} \tag{21}$$

with $\boldsymbol{p} = [p_1(\boldsymbol{z}), \ldots, p_K(\boldsymbol{z})]^T$ being the softmax probability vector.

Lemma 1 (Bounded Cross-Entropy Hessian)**.** *Under Assumption 1, the spectral norm of the cross-entropy Hessian satisfies:*

$$\|\nabla^2_{\boldsymbol{z}}\mathcal{L}_{CE}(\boldsymbol{z})\| \leq KM_W^2 \tag{22}$$

Proof. The matrix $\boldsymbol{D} = \text{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T$ is the covariance matrix of a categorical distribution with spectral norm $\|\boldsymbol{D}\|_2 \leq 1$. The spectral norm of the Hessian is:

$$\|\nabla^2_{\boldsymbol{z}}\mathcal{L}_{\text{CE}}(\boldsymbol{z})\|_2 = \|\boldsymbol{W}^T\boldsymbol{D}\boldsymbol{W}\|_2 \leq \|\boldsymbol{W}^T\|_2\|\boldsymbol{D}\|_2\|\boldsymbol{W}\|_2 = \|\boldsymbol{W}\|_2^2 \tag{23}$$

The spectral norm of $\boldsymbol{W}$ is bounded by its Frobenius norm:

$$\|\boldsymbol{W}\|_2^2 \leq \|\boldsymbol{W}\|_F^2 = \sum_{i=1}^{K}\|\boldsymbol{w}_i\|_2^2 \leq KM_W^2 \tag{24}$$

where the last inequality follows from Assumption 1. □

This lemma shows that the cross-entropy contribution to the Hessian is bounded, allowing the L_2 regularization term to dominate for large γ.

A.3.1 Equilibrium Neighborhood Dynamics

The bounded Hessian structure established in Lemma 1 enables us to analyze how representations behave near equilibrium points under stochastic optimization.

The regularization parameter γ reshapes the loss landscape quantitatively: while the feature-space loss is convex even without regularization, only the penalty produces a finite minimizer and controls the curvature around it. Consider the Hessian at an arbitrary point $\boldsymbol{z}$:

$$\nabla_{\boldsymbol{z}}^2 \mathcal{L}(\boldsymbol{z}) = \nabla_{\boldsymbol{z}}^2 \mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}) + 2\gamma \boldsymbol{I} \tag{25}$$

Because $\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^T$ is positive semi-definite, the cross-entropy Hessian is itself positive semi-definite, and the combined Hessian satisfies $\nabla_{\boldsymbol{z}}^2 \mathcal{L}(\boldsymbol{z}) \succeq 2\gamma \boldsymbol{I} \succ 0$ for every $\gamma > 0$: the loss landscape in the feature variable is globally 2γ-strongly convex, with a unique minimizer for each sample. Large regularization therefore does not create convexity; its role is quantitative. By Lemma 1, $\nabla_{\boldsymbol{z}}^2 \mathcal{L}(\boldsymbol{z}) \preceq (2\gamma + KM_W^2)\boldsymbol{I}$, so for $\gamma \gg KM_W^2$ the curvature is dominated by the isotropic term $2\gamma \boldsymbol{I}$ and the landscape becomes a nearly harmonic well, which is the regime analyzed below.

Near an equilibrium point $\boldsymbol{z}^*$, the dynamics are governed by the local curvature. For large γ satisfying $\gamma \gg KM_W^2$, the Hessian becomes dominated by the regularization term:

$$\boldsymbol{H} = \nabla_{\boldsymbol{z}}^2 \mathcal{L}(\boldsymbol{z}^*) \approx 2\gamma \boldsymbol{I} \tag{26}$$

This creates a strong, isotropic quadratic potential around the equilibrium, analogous to a harmonic oscillator with spring constant $k = 2\gamma$. Under stochastic optimization, representations do not converge to exact points but rather to steady-state distributions around equilibria. The concentration of these distributions is determined by the balance between the confining potential controlled by γ and the stochastic excitation from gradient noise.

Consider the linearized dynamics around equilibrium:

$$\boldsymbol{d}_{t+1} = (\boldsymbol{I} - \eta \boldsymbol{P}_t \boldsymbol{H})\boldsymbol{d}_t - \eta \boldsymbol{P}_t \mathbf{a}_t \tag{27}$$

where $\boldsymbol{d}_t = \boldsymbol{z}_t - \boldsymbol{z}^*$ is the deviation from equilibrium, $\boldsymbol{P}_t$ is the optimizer's preconditioning matrix, and $\mathbf{a}_t$ represents gradient noise.

For the simplified case where $\boldsymbol{H} = 2\gamma \boldsymbol{I}$ and uniform preconditioning $\boldsymbol{P}_t = \bar{\lambda}\boldsymbol{I}$, the steady-state variance of each component satisfies:

$$\mathbb{E}[\delta_{i,\infty}^2] = \frac{\eta \bar{\lambda} \sigma^2}{4\gamma(1 - \eta\gamma\bar{\lambda})} \tag{28}$$

This expression shows that for any fixed learning rate η satisfying the stability condition $\eta < \frac{1}{\gamma\bar{\lambda}}$, the steady-state variance decreases monotonically as γ increases. Taking the derivative with respect to γ yields:

$$\frac{\partial}{\partial \gamma} \mathbb{E}[\delta_{i,\infty}^2] = -\frac{\eta \bar{\lambda} \sigma^2 (1 - 2\eta\bar{\lambda}\gamma)}{4\gamma^2 (1 - \eta\gamma\bar{\lambda})^2} \tag{29}$$

Under the overdamped stability condition $\eta < \frac{1}{2\gamma\bar{\lambda}}$, the factor $(1 - 2\eta\bar{\lambda}\gamma) > 0$, confirming that increasing γ reduces the variance of representations around their equilibria. This stronger condition (compared to the basic stability requirement $\eta < \frac{1}{\gamma\bar{\lambda}}$) ensures that the discrete-time dynamics do not oscillate, which is necessary for monotonic variance reduction. In the underdamped regime $\frac{1}{2\gamma\bar{\lambda}} < \eta < \frac{1}{\gamma\bar{\lambda}}$, the system remains stable but increasing γ can amplify fluctuations through resonance with the discrete time step.

The total expected squared deviation across all d dimensions is:

$$\mathbb{E}[\|\boldsymbol{d}_\infty\|^2] = \sum_{i=1}^{d} \mathbb{E}[\delta_{i,\infty}^2] = \frac{d\eta\bar{\lambda}\sigma^2}{4\gamma(1 - \eta\gamma\bar{\lambda})} \tag{30}$$

For small learning rates satisfying $\eta\gamma\bar{\lambda} \ll 1$ (which is typically required for stability), we can approximate:

$$\mathbb{E}[\|\boldsymbol{d}_\infty\|^2] \approx \frac{d\eta\bar{\lambda}\sigma^2}{4\gamma} \tag{31}$$

Therefore, the root mean square deviation, which characterizes the typical collapse radius, is:

$$r_{\text{collapse}} = \sqrt{\mathbb{E}[\|\boldsymbol{d}_\infty\|^2]} \approx \frac{\sigma}{2}\sqrt{\frac{d\eta\bar{\lambda}}{\gamma}} = O\left(\sqrt{\frac{\sigma^2 d\eta}{\gamma}}\right) \tag{32}$$

where the last equality follows since $\bar{\lambda}$ and σ are $O(1)$ constants independent of γ and η.

This analysis shows that a large regularization parameter γ serves a crucial dual purpose. First, it renders the feature-space landscape 2γ-strongly convex, replacing the minimizer-at-infinity of the pure cross-entropy problem with a unique finite optimum. Second, it induces feature collapse. The strong quadratic potential not only creates this single basin of attraction but also tightly confines representations within it, counteracting the stochasticity from gradient noise. This results in a steady-state variance that decreases monotonically with γ, leading to a tighter clustering of same-class representations.

Consequently, under the overdamped stability condition ($\eta < \frac{1}{2\gamma\bar{\lambda}}$), the optimizer is guaranteed to converge to a neighborhood of the global optimum whose radius, $r_{\text{collapse}} = O(\sqrt{\sigma^2 d\eta/\gamma})$, is explicitly controlled by the regularization strength, with monotonically decreasing variance as γ increases. This provides a principled explanation for the efficacy of strong L_2 regularization in achieving NC, as it ensures both convergence to the correct geometric configuration and control over the degree of intra-class feature concentration.

A.4 Alignment Convergence Analysis in the Terminal Phase

In the terminal phase of training, when samples are classified with high confidence, we analyze how strong regularization induces alignment between representations and their corresponding classifier weights. Throughout this subsection we work in a regime where $p_{\bar{y}}(\boldsymbol{z})$ is close to one so that the approximation $\sum_i p_i \boldsymbol{w}_i \approx p_{\bar{y}}\boldsymbol{w}_{\bar{y}}$ applies, while remaining bounded away from one on the confinement region.

To analyze this process, we consider the gradient flow dynamics $\dot{\boldsymbol{z}} = -\nabla_{\boldsymbol{z}}\mathcal{L}(\boldsymbol{z})$, a valid approximation for small learning rates. We decompose the dynamics into radial and angular components. Let $r(t) = \|\boldsymbol{z}(t)\|$ and $\boldsymbol{u}(t) = \boldsymbol{z}(t)/r(t)$ denote the representation's magnitude and direction, respectively.

Definition 2 (Representation-Weight Alignment)**.** For a representation $\boldsymbol{z}$ with direction $\boldsymbol{u} = \boldsymbol{z}/\|\boldsymbol{z}\|$ and true label $\bar{y}$, the alignment with the corresponding classifier weight is:

$$a(t) = \boldsymbol{u}(t)^T\hat{\boldsymbol{w}}_{\bar{y}} = \frac{\boldsymbol{z}(t)^T\boldsymbol{w}_{\bar{y}}}{\|\boldsymbol{z}(t)\|\|\boldsymbol{w}_{\bar{y}}\|} \tag{33}$$

where $\hat{\boldsymbol{w}}_{\bar{y}} = \boldsymbol{w}_{\bar{y}}/\|\boldsymbol{w}_{\bar{y}}\|$ is the normalized classifier weight. Note that $a(t) \in [-1, 1]$ by the Cauchy-Schwarz inequality, with $a(t) = 1$ indicating perfect alignment.

We previously discussed that the L_2 term $2\gamma\boldsymbol{z}$ creates a strong restoring force in the radial direction, causing $r(t)$ to converge to an equilibrium radius $r^* = O(M_W/\gamma)$.

Theorem 2 (Alignment under L_2 Regularization)**.** *Consider the terminal phase of training in which the approximation $\sum_i p_i\boldsymbol{w}_i \approx p_{\bar{y}}\boldsymbol{w}_{\bar{y}}$ holds, and assume γ lies in a regime where representations are confined to $\|\boldsymbol{z}\| \le M_W/\gamma$ while the correct-class probability remains uniformly bounded away from one: $1 - p_{\bar{y}}(\boldsymbol{z}) \ge \epsilon_0$ for some $\epsilon_0 > 0$ on the confinement region. Under gradient flow dynamics, the alignment $a(t) = \boldsymbol{u}(t)^T\hat{\boldsymbol{w}}_{\bar{y}}$ with the corresponding classifier weight direction is non-decreasing and, whenever $a(0) \ge 0$, satisfies*

$$1 - a(t) \le (1 - a(0))\, e^{-ct}, \tag{34}$$

with rate $c = \epsilon_0\gamma\|\boldsymbol{w}_{\bar{y}}\|/M_W > 0$.

Proof. The angular velocity of the unit vector $\boldsymbol{u}$ is given by projecting the velocity $\dot{\boldsymbol{z}}$ onto the tangent space of the unit sphere:

$$\dot{\boldsymbol{u}} = \frac{1}{r}(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T)\dot{\boldsymbol{z}} = -\frac{1}{r}(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T)\nabla_{\boldsymbol{z}}\mathcal{L}(\boldsymbol{z}) \tag{35}$$

The gradient is $\nabla_{\boldsymbol{z}}\mathcal{L}(\boldsymbol{z}) = (\sum_i p_i \boldsymbol{w}_i - \boldsymbol{w}_{\bar{y}}) + 2\gamma\boldsymbol{z}$. Since $(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T)\boldsymbol{z} = \boldsymbol{0}$, the regularization term does not affect the angular dynamics:

$$\dot{\boldsymbol{u}} = \frac{1}{r}(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T)(\boldsymbol{w}_{\bar{y}} - \sum_i p_i \boldsymbol{w}_i) \tag{36}$$

The rate of change of alignment $a(t) = \boldsymbol{u}^T\hat{\boldsymbol{w}}_{\bar{y}}$ can be derived explicitly. We have:

$$\dot{a} = \dot{\boldsymbol{u}}^T\hat{\boldsymbol{w}}_{\bar{y}} = \frac{1}{r}\left((\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T)(\boldsymbol{w}_{\bar{y}} - \sum_i p_i \boldsymbol{w}_i)\right)^T \hat{\boldsymbol{w}}_{\bar{y}} \tag{37}$$

Since the projection matrix $(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T)$ is symmetric, this simplifies to:

$$\dot{a} = \frac{1}{r}(\boldsymbol{w}_{\bar{y}} - \sum_i p_i \boldsymbol{w}_i)^T(\boldsymbol{I} - \boldsymbol{u}\boldsymbol{u}^T)\hat{\boldsymbol{w}}_{\bar{y}} = \frac{1}{r}(\boldsymbol{w}_{\bar{y}} - \sum_i p_i \boldsymbol{w}_i)^T(\hat{\boldsymbol{w}}_{\bar{y}} - \boldsymbol{u}(\boldsymbol{u}^T\hat{\boldsymbol{w}}_{\bar{y}})) \tag{38}$$

Recalling that $a = \boldsymbol{u}^T\hat{\boldsymbol{w}}_{\bar{y}}$, we get:

$$\dot{a} = \frac{1}{r}(\boldsymbol{w}_{\bar{y}} - \sum_i p_i \boldsymbol{w}_i)^T(\hat{\boldsymbol{w}}_{\bar{y}} - a\boldsymbol{u}) \tag{39}$$

In the TPT, classification confidence is high, so $p_{\bar{y}} \to 1$ and $p_{i\neq\bar{y}} \to 0$. We can thus approximate the softmax-weighted sum of classifiers as $\sum_i p_i \boldsymbol{w}_i \approx p_{\bar{y}}\boldsymbol{w}_{\bar{y}}$. This leads to:

$$\boldsymbol{w}_{\bar{y}} - \sum_i p_i \boldsymbol{w}_i \approx \boldsymbol{w}_{\bar{y}} - p_{\bar{y}}\boldsymbol{w}_{\bar{y}} = (1 - p_{\bar{y}})\boldsymbol{w}_{\bar{y}} \tag{40}$$

Substituting this back into the expression for $\dot{a}$ yields:

$$\dot{a} \approx \frac{1}{r}\left((1 - p_{\bar{y}})\boldsymbol{w}_{\bar{y}}\right)^T(\hat{\boldsymbol{w}}_{\bar{y}} - a\boldsymbol{u}) \tag{41}$$

$$= \frac{1 - p_{\bar{y}}}{r}\left(\boldsymbol{w}_{\bar{y}}^T\hat{\boldsymbol{w}}_{\bar{y}} - a\boldsymbol{w}_{\bar{y}}^T\boldsymbol{u}\right) \tag{42}$$

Using the definitions $\hat{\boldsymbol{w}}_{\bar{y}} = \boldsymbol{w}_{\bar{y}}/\|\boldsymbol{w}_{\bar{y}}\|$ and $a = \boldsymbol{u}^T\hat{\boldsymbol{w}}_{\bar{y}} = \boldsymbol{u}^T\boldsymbol{w}_{\bar{y}}/\|\boldsymbol{w}_{\bar{y}}\|$, we have $\boldsymbol{w}_{\bar{y}}^T\hat{\boldsymbol{w}}_{\bar{y}} = \|\boldsymbol{w}_{\bar{y}}\|$ and $\boldsymbol{w}_{\bar{y}}^T\boldsymbol{u} = a\|\boldsymbol{w}_{\bar{y}}\|$. Substituting these gives:

$$\dot{a} \approx \frac{1 - p_{\bar{y}}}{r}\left(\|\boldsymbol{w}_{\bar{y}}\| - a \cdot a\|\boldsymbol{w}_{\bar{y}}\|\right) \tag{43}$$

$$= \frac{(1 - p_{\bar{y}})\|\boldsymbol{w}_{\bar{y}}\|}{r}(1 - a^2) \tag{44}$$

Since $(1 - a^2) \geq 0$ for $|a| \leq 1$ and $(1 - p_{\bar{y}}) \geq 0$, the alignment is monotonically non-decreasing. It remains to show convergence to $a = 1$ rather than to some $a^* < 1$. Under L_2 regularization, Theorem 1 confines representations to $\|\boldsymbol{z}\| \leq M_W/\gamma$, which implies bounded logits: $|\ell_i| \leq M_W^2/\gamma + \|\boldsymbol{b}\|_\infty$ for all i. Since softmax with finite logits cannot saturate, the correct-class probability is uniformly bounded away from 1:

$$1 - p_{\bar{y}}(\boldsymbol{z}) \geq \frac{(K-1)\exp\big(-2M_W^2/\gamma - 2\|\boldsymbol{b}\|_\infty\big)}{1 + (K-1)\exp\big(-2M_W^2/\gamma - 2\|\boldsymbol{b}\|_\infty\big)} =: \epsilon_0 > 0 \tag{45}$$

for all $\boldsymbol{z}$ in the confinement region. Similarly, the radius satisfies $r(t) \leq M_W/\gamma =: r_{\max}$ after transients. Define $\beta(t) = 1 - a(t) \geq 0$. From the alignment dynamics and the bounds above:

$$\dot{a} \geq \frac{\epsilon_0 \|\boldsymbol{w}_{\bar{y}}\|}{r_{\max}}(1 - a^2) \geq \frac{\epsilon_0 \|\boldsymbol{w}_{\bar{y}}\|}{r_{\max}}(1 - a) = c\,(1 - a) \tag{46}$$

where $c = \frac{\epsilon_0 \gamma \|\boldsymbol{w}_{\bar{y}}\|}{M_W} > 0$ and we used $1 + a \geq 1$, which holds since $a(t) \geq a(0) \geq 0$ by monotonicity of the alignment. This gives $\dot{\beta} \leq -c\beta$, yielding:

$$1 - a(t) \leq (1 - a(0))\, e^{-ct} \tag{47}$$

Therefore $a(t) \to 1$ exponentially with rate at least $c = \frac{\epsilon_0 \gamma \|\boldsymbol{w}_{\bar{y}}\|}{M_W}$. □

This analysis demonstrates that L_2 regularization not only confines representations to a bounded region but actively drives them toward geometric alignment with their corresponding classifier weights. The regularization parameter γ controls the speed and tightness of this alignment by determining the equilibrium radius $r^* = O(M_W/\gamma)$ and the exponential convergence rate $c = \epsilon_0 \gamma \|\boldsymbol{w}_{\bar{y}}\|/M_W$, at which representations stabilize and align. Importantly, this alignment emerges naturally from the optimization dynamics without explicit geometric constraints, revealing how strong regularization implicitly promotes convergence to the NC geometry.

A.5 Unbounded Growth under Pure Cross-Entropy Minimization

Having established the beneficial effects of strong L_2 regularization, we now demonstrate its necessity by analyzing the pathological behavior that emerges in its absence. We show that without L_2 regularization, cross-entropy minimization alone leads to unbounded growth of representation norms, preventing convergence to the theoretical equilibrium predicted by UFM theory. This explains why standard networks without explicit feature regularization fail to reach the ideal NC configuration.

Theorem 3 (Unbounded Norm Growth without Regularization)**.** *In the TPT with $\gamma = 0$, cross-entropy minimization drives representations to grow unboundedly: $\|\boldsymbol{z}(t)\| \to \infty$ as $t \to \infty$.*

Proof. For pure cross-entropy loss with $\gamma = 0$, the gradient is:

$$\nabla_{\boldsymbol{z}} \mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}) = \sum_{i=1}^{K} p_i(\boldsymbol{z}) \boldsymbol{w}_i - \boldsymbol{w}_{\bar{y}} \tag{48}$$

Consider the rate of change of the squared norm:

$$\frac{d}{dt}\|\boldsymbol{z}\|^2 = 2\boldsymbol{z}^T \dot{\boldsymbol{z}} = -2\boldsymbol{z}^T \nabla_{\boldsymbol{z}} \mathcal{L}_{\mathrm{CE}}(\boldsymbol{z}) \tag{49}$$

$$= 2\boldsymbol{z}^T (\boldsymbol{w}_{\bar{y}} - \sum_{i=1}^{K} p_i(\boldsymbol{z}) \boldsymbol{w}_i) \tag{50}$$

$$= 2(\ell_{\bar{y}} - \sum_{i=1}^{K} p_i(\boldsymbol{z}) \ell_i) \tag{51}$$

where $\ell_i = \boldsymbol{z}^T \boldsymbol{w}_i$ are the logits. The term $\ell_{\bar{y}} - \sum_i p_i \ell_i$ represents the difference between the correct class logit and the expected logit over the softmax distribution. This quantity is equivalent to $\sum_i p_i(\ell_{\bar{y}} - \ell_i)$ and is strictly positive as long as perfect classification ($p_{\bar{y}} = 1$) has not been achieved, since $\ell_{\bar{y}}$ will be greater than other logits ℓ_i for which $p_i > 0$. Therefore, $\frac{d}{dt}\|\boldsymbol{z}\|^2 > 0$, indicating that under gradient flow in the terminal phase, the steepest descent direction for CE loss has an outward radial component, causing representation norms to grow continuously.

Crucially, cross-entropy can be decreased by simply scaling up $\boldsymbol{z}$ without requiring alignment. If $\boldsymbol{z}$ makes an angle θ with $\boldsymbol{w}_{\bar{y}}$ where $\cos\theta < 1$, scaling $\boldsymbol{z} \mapsto \alpha\boldsymbol{z}$ with $\alpha > 1$ yields:

$$\mathcal{L}_{\text{CE}}(\alpha\boldsymbol{z}) = -\log \frac{\exp(\alpha\ell_{\bar{y}})}{\sum_i \exp(\alpha\ell_i)} \to 0 \text{ as } \alpha \to \infty \tag{52}$$

provided $\ell_{\bar{y}} > \max_{i\neq\bar{y}} \ell_i$, which only requires $\cos\theta > \cos\theta_{\text{critical}}$ for some critical angle $\theta_{\text{critical}} < \pi/2$.

Finally, monotone growth alone does not preclude convergence to a finite radius; however, the radial growth rate $\sum_i p_i(\ell_{\bar{y}} - \ell_i)$ decays only exponentially in $\|\boldsymbol{z}\|$, and integrating $\frac{d}{dt}\|\boldsymbol{z}\|^2$ yields unbounded, logarithmic-in-time growth of $\|\boldsymbol{z}(t)\|$. □

Corollary 1 (Misalignment Compatible with Loss Decrease)**.** *Even with slight misalignment between $\boldsymbol{z}$ and $\boldsymbol{w}_{\bar{y}}$, cross-entropy loss can decrease through norm growth alone. Specifically, if:*

$$\boldsymbol{z}^T\boldsymbol{w}_{\bar{y}} > \max_{i\neq\bar{y}} \boldsymbol{z}^T\boldsymbol{w}_i \tag{53}$$

then increasing $\|\boldsymbol{z}\|$ while maintaining fixed direction decreases $\mathcal{L}_{CE}(\boldsymbol{z})$.

This unbounded growth has critical consequences:

Proposition 1 (Gradient Saturation and Stalled Convergence)**.** *As $\|\boldsymbol{z}\| \to \infty$ without regularization*

1. *The gradient norm vanishes: $\|\nabla_{\boldsymbol{z}}\mathcal{L}_{CE}(\boldsymbol{z})\| \to 0$*
2. *The loss plateaus: $\mathcal{L}_{CE}(\boldsymbol{z}) \to 0$ at rate $O(e^{-\|\boldsymbol{z}\|})$*
3. *Convergence to any finite equilibrium point becomes impossible*

Proof. As $\|\boldsymbol{z}\| \to \infty$ with $\ell_{\bar{y}} > \max_{i\neq\bar{y}} \ell_i$:

$$p_{\bar{y}}(\boldsymbol{z}) = \frac{\exp(\ell_{\bar{y}})}{\sum_i \exp(\ell_i)} \to 1 \tag{54}$$

The gradient becomes:

$$\|\nabla_{\boldsymbol{z}}\mathcal{L}_{\text{CE}}(\boldsymbol{z})\| = \left\|\sum_{i\neq\bar{y}} p_i(\boldsymbol{w}_i - \boldsymbol{w}_{\bar{y}})\right\| \leq 2M_W(1 - p_{\bar{y}}) \to 0 \tag{55}$$

With vanishing gradients, the dynamics essentially halt, preventing convergence to any finite equilibrium point. The representations are trapped in a regime of infinite growth with diminishing returns in loss reduction. □

This analysis shows that pure cross-entropy optimization, despite achieving low loss values, fails to reach the structured equilibria predicted by theory. The unbounded norm growth and gradient saturation prevent convergence to the exact NC configuration; within-class variability remains finite and only the relative form of collapse is approached. Strong L_2 regularization with large γ is therefore essential to constrain representations to a bounded region where meaningful equilibria can be reached.

A.6 Bounded Output Sensitivity

The confinement effects of strong L_2 regularization lead to a crucial property: bounded output sensitivity.

Theorem 4 (Output Sensitivity Bound)**.** *For a neural network with L_2-regularized penultimate representations trained with any optimizer satisfying Assumption 2, after sufficient training with learning rate $\eta < \frac{1}{2\gamma\bar{\lambda}}$, the expected network output difference satisfies:*

$$\mathbb{E}[\|\boldsymbol{f}(\boldsymbol{x}_1) - \boldsymbol{f}(\boldsymbol{x}_2)\|] \leq L(\gamma) \tag{56}$$

where the output sensitivity bound is:

$$L(\gamma) = 2\sqrt{K}M_W \left(\frac{M_W}{\gamma} + \sqrt{\frac{d\eta\bar{\lambda}\sigma^2}{4\gamma(1-\eta\gamma\bar{\lambda})}} \right) \tag{57}$$

This bound is independent of the input distance $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$. The guarantee applies to representations realized at the optimizer's steady state, i.e., to inputs processed during training; empirically, held-out inputs are equally confined (Table 3), so in practice the bound also curbs the unbounded high-confidence extrapolation of standard ReLU networks (Hein et al., 2019) and restores the bounded loss range assumed by stability-based generalization theory (Bousquet & Elisseeff, 2002).

Proof. For any two inputs $\boldsymbol{x}_1, \boldsymbol{x}_2$, their penultimate-layer representations are $\boldsymbol{z}_1, \boldsymbol{z}_2$. The distance between them is bounded by the triangle inequality: $\|\boldsymbol{z}_1 - \boldsymbol{z}_2\| \leq \|\boldsymbol{z}_1\| + \|\boldsymbol{z}_2\|$.

From Theorem 1, any equilibrium point $\boldsymbol{z}^*$ satisfies $\|\boldsymbol{z}^*\| \leq M_W/\gamma$. From the Hessian analysis, the steady-state deviation from equilibrium satisfies:

$$\mathbb{E}[\|\boldsymbol{d}_\infty\|^2] = \frac{d\eta\bar{\lambda}\sigma^2}{4\gamma(1-\eta\gamma\bar{\lambda})} \tag{58}$$

For the expected norm, we use the fact that for any random vector, $\mathbb{E}[\|\boldsymbol{v}\|] \leq \sqrt{\mathbb{E}[\|\boldsymbol{v}\|^2]}$ by Jensen's inequality:

$$\mathbb{E}[\|\boldsymbol{d}_\infty\|] \leq \sqrt{\mathbb{E}[\|\boldsymbol{d}_\infty\|^2]} = \sqrt{\frac{d\eta\bar{\lambda}\sigma^2}{4\gamma(1-\eta\gamma\bar{\lambda})}} \tag{59}$$

Therefore, the expected norm of any representation satisfies:

$$\mathbb{E}[\|\boldsymbol{z}\|] \leq \|\boldsymbol{z}^*\| + \mathbb{E}[\|\boldsymbol{d}\|] \leq \frac{M_W}{\gamma} + \sqrt{\frac{d\eta\bar{\lambda}\sigma^2}{4\gamma(1-\eta\gamma\bar{\lambda})}} \tag{60}$$

The expectation here is taken over the steady-state distribution of the stochastic optimizer at fixed inputs $\boldsymbol{x}_1, \boldsymbol{x}_2$; it characterizes the typical output sensitivity of the trained model. A high-probability statement for any fixed realization follows from concentration, as noted below. With this understanding, we bound the expected distance between the two representations:

$$\mathbb{E}[\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|] \leq \mathbb{E}[\|\boldsymbol{z}_1\|] + \mathbb{E}[\|\boldsymbol{z}_2\|] \leq 2\left(\frac{M_W}{\gamma} + \sqrt{\frac{d\eta\bar{\lambda}\sigma^2}{4\gamma(1-\eta\gamma\bar{\lambda})}} \right) \tag{61}$$

The network output difference is $\boldsymbol{f}(\boldsymbol{x}_1) - \boldsymbol{f}(\boldsymbol{x}_2) = \boldsymbol{W}(\boldsymbol{z}_1 - \boldsymbol{z}_2)$. Using the spectral norm bound $\|\boldsymbol{W}\|_2 \leq \sqrt{K}M_W$ from Lemma 1:

$$\|\boldsymbol{f}(\boldsymbol{x}_1) - \boldsymbol{f}(\boldsymbol{x}_2)\| \leq \|\boldsymbol{W}\|_2\|\boldsymbol{z}_1 - \boldsymbol{z}_2\| \leq \sqrt{K}M_W\|\boldsymbol{z}_1 - \boldsymbol{z}_2\| \tag{62}$$

Therefore, taking expectations over the steady-state distribution:

$$\mathbb{E}[\|\boldsymbol{f}(\boldsymbol{x}_1) - \boldsymbol{f}(\boldsymbol{x}_2)\|] \leq \sqrt{K}M_W \cdot \mathbb{E}[\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|] \leq L(\gamma) \tag{63}$$

which establishes the expected bound on output sensitivity. □

The structure of the bound shows that both terms decrease as γ increases:

$$L(\gamma) = 2\sqrt{K}M_W \underbrace{\left(\frac{M_W}{\gamma}\right)}_{\text{equilibrium term}} + 2\sqrt{K}M_W \underbrace{\left(\sqrt{\frac{d\eta\bar{\lambda}\sigma^2}{4\gamma(1-\eta\gamma\bar{\lambda})}}\right)}_{\text{stochastic fluctuation term}} \tag{64}$$

The first term decreases as $O(1/\gamma)$ while the second decreases as $O(1/\sqrt{\gamma})$ for small learning rates where $\eta\gamma\bar{\lambda} \ll 1$. Therefore, increasing the regularization strength γ monotonically tightens the uniform output bound, with the dominant improvement coming from the $O(1/\gamma)$ reduction in the equilibrium bound.

Remark 1 (High-Probability Bound). The bound on output differences is established here in expectation. A high-probability statement for any fixed realization of training follows from concentration inequalities; for instance, by Markov's inequality, for any $\delta > 0$,

$$\Pr[\|\boldsymbol{f}(\boldsymbol{x}_1) - \boldsymbol{f}(\boldsymbol{x}_2)\| > L(\gamma)/\delta] \leq \delta. \tag{65}$$

A.7 Main Result

Theorem 5 (Latent Point Collapse under Strong L_2 Regularization)**.** *For a neural network with bounded classifier weights $\|\boldsymbol{w}_i\| \leq M_W$ and L_2-regularized penultimate representations with parameter γ, trained with an optimizer satisfying Assumption 2 and learning rate $\eta < \frac{1}{2\gamma\bar{\lambda}}$:*

1. ***Confinement:*** *All equilibrium representations satisfy $\|\boldsymbol{z}^*\| \leq \frac{M_W}{\gamma}$*
2. ***Collapse:*** *Under stochastic optimization, intra-class representations concentrate within radius:*

$$r_{collapse} = \sqrt{\mathbb{E}[\|\boldsymbol{d}_\infty\|^2]} = \frac{\sigma}{2}\sqrt{\frac{d\eta\bar{\lambda}}{\gamma(1-\eta\gamma\bar{\lambda})}} = O\left(\sqrt{\frac{\sigma^2 d\eta}{\gamma}}\right) \tag{66}$$

   where the asymptotic form holds for small learning rates $\eta\gamma\bar{\lambda} \ll 1$.

3. ***Alignment:*** *In the terminal phase of training, representations align with their corresponding classifier weights, $\lim_{t\to\infty} \frac{\boldsymbol{z}(t)}{\|\boldsymbol{z}(t)\|} = \frac{\boldsymbol{w}_{\bar{y}}}{\|\boldsymbol{w}_{\bar{y}}\|}$, at exponential rate $c = \epsilon_0\gamma\|\boldsymbol{w}_{\bar{y}}\|/M_W$ (Theorem 2)*
4. ***Bounded Output Sensitivity:*** *For any inputs $\boldsymbol{x}_1, \boldsymbol{x}_2$, in expectation:*

$$\mathbb{E}[\|\boldsymbol{f}(\boldsymbol{x}_1) - \boldsymbol{f}(\boldsymbol{x}_2)\|] \leq \frac{2\sqrt{K}M_W^2}{\gamma} + O\left(\sqrt{\frac{\sigma^2 d\eta}{\gamma}}\right) \tag{67}$$

   This bound is independent of input distance $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$.

5. ***Necessity of Regularization:*** *Without L_2 regularization ($\gamma = 0$), cross-entropy minimization alone causes unbounded norm growth ($\|\boldsymbol{z}\| \to \infty$), gradient saturation, and failure to reach theoretical equilibria.*

A.8 Summary

Strong L_2 regularization with parameter γ is both necessary and sufficient to achieve LPC while ensuring a uniform bound on output sensitivity. Under the overdamped stability condition $\eta < \frac{1}{2\gamma\bar{\lambda}}$, which ensures both convergence and monotonic variance reduction, the regularization parameter provides a unified mechanism that:

1. **Prevents representation explosion:** Creates an inward force proportional to $2\gamma\boldsymbol{z}$ that counteracts the outward bias of gradient descent in the terminal phase. Without this regularization, gradient-based optimization of cross-entropy loss leads to unbounded norm growth, as the steepest descent direction has an outward radial component whenever $p_{\bar{y}} < 1$ (Theorem 3).

2. **Induces intra-class collapse:** The penalty makes the feature-space loss 2γ-strongly convex, and for $\gamma \gg KM_W^2$ its curvature is dominated by the isotropic term $2\gamma\boldsymbol{I}$. Under stochastic optimization, representations concentrate around class-specific equilibria with collapse radius $r_{\text{collapse}} = O(\sqrt{\sigma^2 d\eta/\gamma})$, where the primary control is through γ. The steady-state variance $\mathbb{E}[\|\boldsymbol{d}_\infty\|^2] = \frac{d\eta\bar{\lambda}\sigma^2}{4\gamma(1-\eta\gamma\bar{\lambda})}$ decreases monotonically as γ increases.

3. **Drives weight alignment:** Because the L_2 gradient is purely radial, it does not enter the angular dynamics of representations; angular evolution is driven entirely by the cross-entropy term. In the terminal phase, where classification confidence is high but confinement keeps $p_{\bar{y}}$ bounded away from one, this driving force aligns each representation with its corresponding classifier weight vector. The alignment rate $(1-a^2)(1-p_{\bar{y}})\|\boldsymbol{w}_{\bar{y}}\|/r$ depends on the current alignment a and the equilibrium radius $r = O(M_W/\gamma)$, so a smaller $r^* = O(M_W/\gamma)$ amplifies the angular response and accelerates alignment.

4. **Bounds output sensitivity:** By confining representations to a bounded region, yields an expected uniform bound on output differences:

$$L(\gamma) = 2\sqrt{K}M_W\left(\frac{M_W}{\gamma} + O\left(\sqrt{\frac{\sigma^2 d\eta}{\gamma}}\right)\right) \tag{68}$$

   Both terms decrease with increasing γ, with the dominant $O(1/\gamma)$ term providing the primary improvement. This bound is independent of input distance $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$, establishing uniform stability across the input space.

5. **Enables convergence to global optima:** Strong convexity of the feature-space loss yields a single-basin landscape with a unique, finite global minimum for any $\gamma > 0$. Combined with appropriate learning rate $\eta < \frac{1}{2\gamma\bar{\lambda}}$, guarantees convergence to a neighborhood of the global optimum, with neighborhood radius controlled by the regularization strength.

Without large γ, cross-entropy optimization alone produces unbounded representation growth without achieving the structured geometry predicted by UFM theory. In the terminal phase, gradient descent follows a path with an outward radial component, leading to gradient saturation as $\|\nabla_{\boldsymbol{z}}\mathcal{L}_{\text{CE}}\| \to 0$ and preventing convergence to finite equilibria. Strong L_2 regularization resolves this pathology by providing a countervailing inward force, transforming an ill-posed optimization problem into a well-conditioned one with provable convergence guarantees and explicit control over both geometric structure and robustness properties.

B Training details

Our ablation study systematically evaluates each component of the proposed method. All architectures employ a shared backbone network that produces the latent representation $\boldsymbol{h}(\boldsymbol{x})$, while differing in their approach to final classification.

We denote the architecture with L_2 regularization applied to a linear penultimate layer of intermediate dimensionality as LPC, with its lower-dimensional and higher-dimensional variants designated as LPC-Narrow and LPC-Wide, respectively. These three variants examine the effect of penultimate-layer dimensionality on LPC formation and network performance. To isolate the contribution of L_2 regularization, we include LinPen and NonlinPen controls, that is linear and nonlinear penultimate layers matching LPC's dimensionality but trained exclusively with cross-entropy loss.

The LPC-NoPen model tests whether the penultimate-layer structure is necessary by applying L_2 regularization directly to the backbone output $\boldsymbol{h}(\boldsymbol{x})$ and, unlike other models, excludes an intermediate layer. The NoPen model serves as our baseline, performing linear classification directly on $\boldsymbol{h}(\boldsymbol{x})$ using only cross-entropy loss. To demonstrate that LPC benefits cannot be attributed solely to stronger weight decay regularization, we include NoPenWD, which employs the same architecture as NoPen but with substantially increased weight decay.

We compare against SCL and ArcFace baselines, which implement their respective loss functions on the NoPen baseline architecture. The LPC-SCL architecture combines L_2 regularization on an intermediate penultimate linear layer with SupCon applied to the backbone's latent representations, testing compatibility with other metric learning approaches. Note that in LPC-SCL, these losses operate on distinct network layers.

All experiments were conducted on CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), and ImageNet-1K (Deng et al., 2009) datasets. To generate latent representations $\boldsymbol{h}(\boldsymbol{x})$, we employed ResNet (He et al., 2016) backbones tailored to dataset complexity: ResNet-18 for CIFAR-10, ResNet-50 for CIFAR-100, and WideResNet-50 (Zagoruyko & Komodakis, 2017) for ImageNet. All architectures incorporated batch normalization and Swish activation functions (Ramachandran et al., 2017) throughout.

Architecture configurations varied by method: LPC, LPC-SCL, LinPen, and NonlinPen included a 64-dimensional fully connected penultimate layer. LPC-Wide expanded this to 128 dimensions, while LPC-Narrow reduced it to 32 dimensions for CIFAR datasets. For ImageNet, LPC employed 128 penultimate dimensions. The LPC-NoPen, NoPen, NoPenWD, SCL, and ArcFace architectures omitted the penultimate layer entirely (see Table 1).

We trained models using AdamW (Kingma & Ba, 2017; Loshchilov & Hutter, 2019) with default PyTorch parameters. All experiments used a batch size of 128 for CIFAR datasets and 64 for ImageNet. Weight decay was set to 5×10^{-4} for CIFAR datasets and 1×10^{-4} for ImageNet. In the architecture NoPenWD weight decay had a larger value set to 1×10^{-1}. Data augmentation comprised random horizontal flips and random crops (padding=4) for CIFAR datasets, while ImageNet used random horizontal flips and random cropping to 224×224 pixels. We evaluated four learning rates from a geometric sequence $(10^{-4}, 2{\times}10^{-4}, 4{\times}10^{-4}, 8{\times}10^{-4})$ for CIFAR experiments and three rates $(5 \times 10^{-5}, 1 \times 10^{-4}, 1 \times 10^{-3})$ for ImageNet, selecting the best-performing configuration based on final test accuracy. For ArcFace, we omitted the classifier bias b (Deng et al., 2019).

Training protocols differed by dataset: CIFAR models trained for 1 000 epochs with cosine annealing from epoch 200 (reducing learning rate to 1×10^{-8}), excluding the final classifier layer. ImageNet models trained for 200 epochs with annealing from epoch 70 (minimum 1×10^{-7}). Extended training ensured CIFAR experiments operated predominantly in the TPT, defined as achieving $> 99.9\%$ training accuracy (Papyan et al., 2020). ImageNet experiments did not reach TPT. We note that learning rate decay begins when the L_2 coefficient γ is already quite large (though still increasing), which we found provides additional training stability during the later stages of optimization.

For L_2-regularized architectures, we initialized the coefficient $\gamma = 10^{-4}$ and increased it geometrically: $\gamma \leftarrow \gamma \cdot \gamma_{\text{step}}$ every 5 epochs (CIFAR) or 1 epoch (ImageNet). We set $\gamma_{\text{step}} = 2$ until $\gamma = 10^3$, then $\gamma_{\text{step}} = 1.25$ until reaching $\gamma_{\max} = 10^6$, where it remained constant.

The supervised contrastive loss (SCL) (Khosla et al., 2020) was defined as:

$$\mathcal{L}^{\text{SCL}} = -\frac{1}{N} \sum_{i=1}^{N} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(s_{i,p}/\tau)}{\sum_{j \neq i} \exp(s_{i,j}/\tau)}, \tag{69}$$

where N denotes batch size, $P(i)$ represents positive samples for instance i, $s_{i,j}$ is the cosine similarity between samples i and j, and $\tau = 0.05$ is the temperature parameter. We jointly optimized $\mathcal{L}^{\text{SCL}}$ with cross-entropy loss for classification.

For ArcFace (Deng et al., 2019), we applied an angular margin to scaled cosine similarities before computing cross-entropy:

$$\mathcal{L}^{\text{ArcFace}} = \frac{1}{N} \sum_{i=1}^{N} \text{CE}\left(\text{Softmax}[s \cdot \cos(\theta_i + m)], y_i\right), \tag{70}$$

where the angular margin m increased from 0.1 to 0.5 and scale factor s from 16 to 64 during training.

All experiments were repeated five times with different random seeds. Results report mean ± standard deviation across trials.