Journal of Machine Learning Research 24 (2023) 1-62 Submitted 6/23; Revised 10/23; Published 12/23

On the Dynamics Under the Unhinged Loss and Beyond

Xiong Zhou CSZX@HIT.EDU.CN
Xianming Liu* CSXM@HIT.EDU.CN
Hanzhang Wang CSWHZ@HIT.EDU.CN
Deming Zhai ZHAIDEMING@HIT.EDU.CN
Junjun Jiang JIANGJUNJUN@HIT.EDU.CN
School of Computer Science and Technology
Harbin Institute of Technology
Harbin, 150001, China

Xiangyang Ji XYJI@TSINGHUA.EDU.CN
Department of Automation
Tsinghua University
Beijing, 100084, China

Editor: Samy Bengio

Abstract

Recent works have studied implicit biases in deep learning, especially the behavior of last-layer features and classifier weights. However, they usually need to simplify the intermediate dynamics under gradient flow or gradient descent due to the intractability of loss functions and model architectures. In this paper, we introduce the unhinged loss, a concise loss function, that offers more mathematical opportunities to analyze the closed-form dynamics while requiring as few simplifications or assumptions as possible. The unhinged loss allows for considering more practical techniques, such as time-vary learning rates and feature normalization. Based on the layer-peeled model that views last-layer features as free optimization variables, we conduct a thorough analysis in the unconstrained, regularized, and spherical constrained cases, as well as the case where the neural tangent kernel remains invariant. To bridge the performance of the unhinged loss to that of Cross-Entropy (CE), we investigate the scenario of fixing classifier weights with a specific structure, (e.g., a simplex equiangular tight frame). Our analysis shows that these dynamics converge exponentially fast to a solution depending on the initialization of features and classifier weights. These theoretical results not only offer valuable insights, including explicit feature regularization and rescaled learning rates for enhancing practical training with the unhinged loss, but also extend their applicability to other loss functions. Finally, we empirically demonstrate these theoretical results and insights through extensive experiments.

Keywords: implicit bias, neural collapse, gradient flow, gradient descent

1. Introduction

Deep learning with neural networks has achieved great success in a variety of tasks (LeCun et al., 2015), which, however, is not entirely understood in the interpolation and generalization of the learned models (Zhang et al., 2017; Neyshabur et al., 2017; Nakkiran et al.,

*. The corresponding author

2019; Bubeck and Sellke, 2021; Mei and Montanari, 2021; Zhou et al., 2023; Langer et al., 2021; Qi et al., 2021). Many modules, including loss functions (Lin et al., 2017; Hui and Belkin, 2021) and optimization algorithms (Auer et al., 2002; Duchi et al., 2011; Zeiler, 2012; Kingma and Ba, 2015), play a crucial role in the training of deep neural networks, but lack convincing explanations due to the complexity of multilayered architectures. Recent works are devoted to simplifying modeling to better understand the behavior of DNNs, (Papyan et al., 2020; Mixon et al., 2022; Fang et al., 2021b; Tian et al., 2021; Han et al., 2022) and then to gain insights for new algorithms, theoretical, and experimental investigations.

To better understand the implicit regularization that improves the generalization of trained models in deep learning, many studies have investigated the implicit bias of gradient descent (Hardt et al., 2016; Sekhari et al., 2021), with an emphasis on the behavior of linear predictors (or called classifiers) over linearly separable data (Soudry et al., 2018; Gunasekar et al., 2018; Nacson et al., 2019; Ji and Telgarsky, 2019; Ji et al., 2020; Shamir, 2021). In particular, Soudry et al. (2018) demonstrated that gradient descent iterates under exponentially-tailed loss minimization on separable data are biased toward ℓ_2-maximum-margin solutions and that continuing to optimize can still lead to performance improvements, even if the validation loss increases. Further, Ji et al. (2020) showed that the gradient descent path and the algorithm-independent regularization path converge to the same direction for general losses. Shamir (2021) formally proved that standard gradient methods never overfit on separable data. These works impressively expose the implicit regularization induced by optimization algorithms and help to understand the generalization of the learned models, but they *mainly focus on the behavior of linear classifiers* that is only the last layer of neural networks, while the classifier actually interacts strongly with the features produced by many nonlinear layers and parameterized layers. Thus, the relevant conclusions do not always apply to deep learning. For example, in (Soudry et al., 2018), the convergence rate of gradient descent is rather slow, wherein for almost all datasets, the distance to the maximum-margin solution decreases only as $O(1/\log t)$, and in some degenerate datasets, the rate further slows down to $O(\log\log t/\log t)$. However, the training of DNNs typically takes only a few hundred epochs. In this paper, we show that exponential convergence is more realistic.

This paper is closely related to another research line which emerged after the empirical discovery of Neural Collapse by Papyan et al. (2020). This phenomenon precisely characterizes a pervasive inductive bias of both features and linear classifiers at the terminal phase of training, and has opened a rich area of exploring this area with simplified mathematical frameworks (Mixon et al., 2022; Fang et al., 2021b; Lu and Steinerberger, 2022; Galanti et al., 2021; Fang et al., 2021a; Zhu et al., 2021; Hui et al., 2022; Tirer and Bruna, 2022; Lu and Steinerberger, 2022; Kothapalli et al., 2022; Zhou et al., 2022b). Neural collapse provides a clear view of how the last-layer features and linear classifiers behave after interpolation and enables us to understand the benefit of achieving zero training error in terms of generalization and robustness (Poggio and Liao, 2020; Wang et al., 2021; Kornblith et al., 2021; Thrampoulidis et al., 2022; Gao et al., 2023), but the intermediate dynamics have remained challenging to analyze due to the intractability of cross entropy (CE). To alleviate this issue, some studies (Mixon et al., 2022; Han et al., 2022; Zhou et al., 2022a; Tirer and Bruna, 2022; Kothapalli et al., 2022; Xu et al., 2023) have explored the more tractable mean squared error (MSE) loss, which performs comparably to those trained with CE (Demirkaya

Reference	Contribution	Extra Simplification or Assumption	Loss Function
Papyan et al. (2020)	Empirical results and theoretical formulation	✗	The CE loss
Fang et al. (2021b)	Global optimum in **regularized cases**	✗	The CE loss
Zhu et al. (2021)	Landscape analysis in **regularized** cases	✗	The CE loss
Ji et al. (2022)	Landscape analysis in **unconstrained** cases	✗	The CE loss
Tirer and Bruna (2022)	Global optimum with **extended unconstrained feature model**	✗	The MSE loss
Mixon et al. (2022)	Intermediate dynamics in **unconstrained** cases	✓	The MSE loss
Han et al. (2022)	Intermediate dynamics in **weight-regularized** cases	✓	The MSE loss
Tirer et al. (2023)	Intermediate dynamics by perturbation analysis in **regularized** cases	✓	The MSE loss
This paper	Intermediate dynamics and convergence analysis in **unconstrained**, **regularized**, **spherical constrained** cases with **time-varying learning rates**	✗	The unhinged loss

Table 1: Comparison of recent analysis for investigating the behavior of last-layer features and prototypes. Compared with prior work, this paper considers the time-varying learning rate, which is often used in practice, and provides intermediate dynamics and convergence analysis in unconstrained, regularized and spherical constrained cases while requiring fewer simplifications or assumptions.

et al., 2020; Hui and Belkin, 2021). However, these studies *still need to make additional simplifications or assumptions for intermediate dynamics.* For instance, Mixon et al. (2022) formulate the gradient flow of the unconstrained feature model as a nonlinear ordinary differential equation and then linearize the equation by claiming that nonlinear terms are negligible for models initialized near the origin. Han et al. (2022) assume that the gradient flow is along the central path which requires the linear classifier to stay MSE-optimal for features throughout the dynamics. Furthermore, Tirer et al. (2023) delve into a more practical model, demonstrating the exponential decay of within-class variability. They assume that features remain in close proximity to predefined features under MSE loss function, but note that the gradient flow is still essential along the "central path". Therefore, MSE is still not simple enough to derive exact dynamics in certain mathematical frameworks, making it difficult to grasp and bridge the gap between the modeling and practical optimization.

In this paper, our objective is to analyze the closed-form dynamics under gradient descent within the layer-peeled model (Fang et al., 2021b) (also known as the unconstrained

features model (Mixon et al., 2022)) with minimal simplifications and assumptions. To achieve this, we introduce the unhinged loss, which possesses a concise form and intuitively expresses the classification objective. To further contextualize the performance of the unhinged loss in relation to the CE loss, we investigate scenarios involving fixing classifier weights with specific structures. Compared to previous works (as depicted in Table 1), the unhinged loss provides more mathematical opportunities for gaining insights into deep learning with closed-form dynamics, while demanding fewer simplifications or assumptions for intermediate dynamics. This equips us for more practical considerations and more rational designs. Our primary contributions are outlined as follows:

- We introduce the unhinged loss for analyzing closed-form dynamics in deep learning with as few simplifications or assumptions as possible. The unhinged loss with some auxiliary techniques can achieve comparative performance to the CE loss.

- We derive exact dynamics of last-layer features and prototypes in unconstrained, regularized, prototype-anchored and spherical constrained cases as well as the NTK-invariant case. For spherical constrained cases that do not exhibit convexity, Lipschitzness, and β-smoothness, we also prove that gradient descent biases the normalized features towards a global minimizer.

- We provide the corresponding convergence analysis, which shows that the features and classifier weights converge to a solution depending on the initialization rather than induce the neural collapse solution that forms a simplex equiangular tight frame, suggesting that not all losses under gradient descent would lead to neural collapse (as verified in Section 3).

- We prove that the rate of convergence is exponential as a function of $\zeta(t) = \int_0^t \eta(\tau)\mathrm{d}\tau$, where $\eta(\tau)$ denotes the learning rate over time. This exponential convergence rate highlights the impact of the interaction of features and classifier weights.

- Moreover, we provide some insights and extensive verification for improvements in practical training with the unhinged loss and other losses (cf. Section 4).

2. The Unhinged Loss

In this paper, we primarily investigate the behavior of last-layer features and classifier weights in DNNs for classification tasks. We conduct our study to datasets comprising inputs from C different classes, each with N examples. The last-layer features $\mathbf{h}_{i,c} = \mathbf{f}_\Theta(\mathbf{x}_{i,c}) \in \mathbb{R}^p$ [1] are obtained from the i-th example $\mathbf{x}_{i,c}$ through a series of parameterized layers $\mathbf{f}_\Theta : \mathcal{X} \to \mathbb{R}^p$, commonly treated as free optimization variables (Mixon et al., 2022; Fang et al., 2021b; Han et al., 2022; Ji et al., 2022). The final layer of the network, known

1. For the closed-form solution of neural collapse, it is worth noting that Zhou et al. (2022d) assume $p \geq C-1$, while Han et al. (2022) assume $p > C$, as last-layer features typically have a higher dimension than the number of classes. This work will explicitly address the relationship between p and C in some scenarios, covering all potential choices of feature dimension in others.

as the linear classifier, is equipped with a class prototype $\mathbf{w}_c \in \mathbb{R}^p$ and a bias $b_c \in \mathbb{R}$ for each class $c \in [C]$. It predicts a label using the rule $\arg\max_{c'}(\langle \mathbf{w}_{c'}, \mathbf{h} \rangle + b_{c'})$ [2].

To better understand the dynamics of last-layer features and prototypes based on gradient flow or gradient descent, we consider a concise loss that offers more mathematical opportunities than the hard-to-analyze CE loss and the MSE loss [3]. Specifically, we investigate a generalized form of the unhinged loss (Van Rooyen et al., 2015) as follows:

$$L_\gamma(\boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}, y) = -\boldsymbol{w}_y^\top \boldsymbol{h} - b_y + \gamma \sum_{j \neq y} (\boldsymbol{w}_j^\top \boldsymbol{h} + b_j), \tag{1}$$

where $\gamma > 0$ is the trade-off parameter and y denotes the class label of the feature $\boldsymbol{h}$. When $\gamma = \frac{1}{C-1}$, the loss $\frac{1}{C-1}\sum_{j\neq y}[\boldsymbol{w}_y^\top \boldsymbol{h} + b_y - (\boldsymbol{w}_j^\top \boldsymbol{h} + b_j)]$ can be regarded as the unhinged version of the hinge loss[4] that removes the maximum operator and the margin term. We also note that the sample margin $m(\boldsymbol{h}, y) = \boldsymbol{w}_y^\top \boldsymbol{h} + b_y - \max_{j\neq y}(\boldsymbol{w}_j^\top \boldsymbol{h} + b_j)$ (Koltchinskii and Panchenko, 2002; Cao et al., 2019; Zhou et al., 2022d) is defined to measure the discriminativeness for a sample, which satisfies $m(\boldsymbol{h}, y) \leq \frac{1}{C-1}\sum_{j\neq y}[\boldsymbol{w}_y^\top \boldsymbol{h} + b_y - (\boldsymbol{w}_j^\top \boldsymbol{h} + b_j)]$, *i.e.*, $L_\gamma(\boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}, y)$ with $\gamma = \frac{1}{C-1}$ averaging the margins over all non-target classes is the lower bound of $-m(\boldsymbol{h}, y)$. Here, we replace $\frac{1}{C-1}$ with an additional parameter γ that balances positive and negative logits to draw general conclusions. More clarification about the unhinged loss can be found in Appendix A.

Intuitively, the unhinged loss in Equation (1) promotes the learned feature $\boldsymbol{h}$ to increase the logit of the target class while decreasing the logits of the other classes. If we follow up the layer-peeled model (Fang et al., 2021b) to restrict the norms of both features and prototypes, the global minimizer of $\frac{1}{CN}\sum_{i=1}^{N}\sum_{c=1}^{C} L_\gamma(\boldsymbol{W}\boldsymbol{h}_{i,c}, y_{i,c})$ (the bias term $\boldsymbol{b}$ is omitted) will lead to *Neural Collapse* (Papyan et al., 2020; Han et al., 2022):

Lemma 1 (**Neural Collapse under the Unhinged Loss**). *For norm-bounded prototypes and features, i.e., $\|\boldsymbol{w}_c\|_2 \leq E_1$ and $\|\boldsymbol{h}_{i,c}\|_2 \leq E_2$, $\forall i \in [N], \forall c \in [C]$, the global minimizer of $\frac{1}{CN}\sum_{i=1}^{N}\sum_{c=1}^{C} L_\gamma(\boldsymbol{W}\boldsymbol{h}_{i,c}, y_{i,c})$ implies neural collapse when $p \geq C-1$. More specifically, the global minimizer is uniquely obtained at $\frac{\boldsymbol{w}_i^\top \boldsymbol{w}_j}{\|\boldsymbol{w}_i\|_2\|\boldsymbol{w}_j\|_2} = -\frac{1}{C-1}$, $\forall i \neq j$, $\frac{\boldsymbol{w}_{y_{i,c}}^\top \boldsymbol{h}_{i,c}}{\|\boldsymbol{w}_{y_{i,c}}\|_2\|\boldsymbol{h}_{i,c}\|_2} = 1$, $\|\boldsymbol{w}_c\|_2 = E_1$, and $\|\boldsymbol{h}_{i,c}\|_2 = E_2$, $\forall i \in [N]$, $\forall c \in [C]$.*

This lemma shows that the Neural Collapse solution[5]. is the only global optimal solution to minimize $\frac{1}{CN}\sum_{i,c} L_\gamma(\boldsymbol{W}\boldsymbol{h}_{k,c}, y_{i,c})$ in the norm-bounded case. However, there exists an

2. Unless specified otherwise, we denote vectors using boldface italicized letters, and elements within the vector are denoted with italicized letters and subscripts.
3. The CE loss is known to have non-analytical intermediate dynamics, which makes it difficult to investigate the precise nature of the dynamics. In addition, the intermediate dynamics under MSE can be derived into closed-form, which however is quadratic and cannot be easily analyzed. Therefore, some additional simplifications or assumptions on CE and MSE are required to simplify the intermediate dynamics.
4. The multi-class hinge loss is formulated as $L_{\text{hinge}}(\boldsymbol{s}, y; m) = \sum_{i\neq y}\max\{0, s_i - s_y + m\}$, where m is the margin term.
5. The Neural Collapse solution exhibits three critical properties (Papyan et al., 2020): (i) **Within-class variability collapse**, the within-class variation of features becomes negligible as these features collapse to their class means; (ii) **Convergence to self-duality**, the class means and classifier weights converge to each other, up to rescaling; (iii) **Convergence to a simplex equiangular tight frame (ETF)**, the vectors of class means converge to having equal length, forming equal-sized angles between any given pair, and being the maximally pairwise-distanced configuration. For instance, prototypes $\boldsymbol{W}$ that forms a simplex ETF satisfies $\boldsymbol{W}^\top\boldsymbol{W} = \frac{C}{C-1}\boldsymbol{I} - \frac{1}{C-1}\boldsymbol{1}\boldsymbol{1}^\top$ when $C < p$

undesired direction to minimize the unhinged loss in unconstrained cases, since the norm of features and prototypes tends to grow to infinity. For example, we can directly scale up $\boldsymbol{W}$ and $\boldsymbol{b}$ to obtain a smaller loss if $L_\gamma(\boldsymbol{W}\boldsymbol{h}+\boldsymbol{b}, y) < 0$, which will happen analogously to CE (Liu et al., 2016; Wang et al., 2017; Zhou et al., 2022d). In this paper, we will analytically characterize the direction in which features $\boldsymbol{H}$ and prototypes $\boldsymbol{W}$ diverge. Specifically, we show that the gradient flow or gradient descent with the unhinged loss will exhibit an implicit bias associated with the initialization of features and prototypes. We further investigate the prototype-anchored case, wherein class prototypes are anchored with some specific structures, and the unhinged loss can perform comparative to CE.

3. Main Theoretical Results

In this section, we begin with the commonly used assumption that treats last-layer features as free optimization variables. We then conduct a comprehensive analysis of the closed-form dynamics of last-layer features and prototypes under the unhinged loss in various scenarios, including unconstrained, regularized, prototype-anchored, spherical constrained, and NTK-invariant cases. Additionally, we provide convergence analyses for each case, revealing a surprising result: all cases exhibit exponential convergence. **All proofs can be found in Appendix B.**

3.1 The Unconstrained Case

We first consider the unconstrained case (Mixon et al., 2022; Ji et al., 2022) in which there is no constraint or regularization on features and prototypes, *i.e.*, learning with the following objective

$$\mathcal{L} = \frac{1}{CN}\sum_{i,c}\big[-\boldsymbol{w}_{y_{i,c}}^\top \boldsymbol{h}_{i,c} - b_{y_{i,c}} + \gamma \sum_{j\neq y_{i,c}} (\boldsymbol{w}_j^\top \boldsymbol{h}_{i,c} + b_j)\big], \tag{2}$$

which can be reformulated as

$$\mathcal{L} = \mathrm{Tr}(\boldsymbol{Y}^\top \boldsymbol{W}^\top \boldsymbol{H}) + \tfrac{\gamma C-\gamma-1}{C}\mathbf{1}_C^\top \boldsymbol{b}, \tag{3}$$

where $\boldsymbol{H} = [\boldsymbol{h}_{1,1},\ldots,\boldsymbol{h}_{1,C},\boldsymbol{h}_{2,1},\ldots,\boldsymbol{h}_{N,C}] \in \mathbb{R}^{p\times CN}$ is the matrix resulting from stacking together the feature vectors as columns, $\boldsymbol{Y} = \frac{1}{CN}(\gamma \mathbf{1}_C \mathbf{1}_{CN}^\top - (1+\gamma)(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top))$, $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_C] \in \mathbb{R}^{p\times C}$ is the matrix resulting from stacking class prototypes as columns, $\otimes$ denotes the Kronecker product, $\boldsymbol{I}_C$ is the identity matrix, $\mathbf{1}_C$, $\mathbf{1}_N$, and $\mathbf{1}_{CN}$ are the length-C, -N, and -CN vectors of ones, respectively. For brevity, we represent the label set $\{y_{i,c}\}_{1\leq i\leq N, 1\leq c\leq C}$ as the columns of the matrix $\boldsymbol{I}_C \otimes \mathbf{1}_N^\top$.

Remark We follow the unconstrained features modeling perspective (Mixon et al., 2022; Ji et al., 2022) or the layer-peeled model (Fang et al., 2021b) that treats $\boldsymbol{H}$ as a free optimization variable. Within this model, we analyze the continuous dynamics of features $\boldsymbol{H}$, prototypes $\boldsymbol{W}$ and biases $\boldsymbol{b}$ with gradient flow where *time-of-training* is denoted by the variable t[6].

6. Intuitively, we interpret $t = 0$ as the initial state, that is $\boldsymbol{H}(0) = \boldsymbol{H}_0$, $\boldsymbol{W}(0) = \boldsymbol{W}_0$, and $\boldsymbol{b}(0) = \boldsymbol{b}_0$.

Taking the partial derivative with respect to $\boldsymbol{H}$, $\boldsymbol{W}$, and $\boldsymbol{b}$, respectively, we have the following:

$$\begin{aligned}\nabla_{\boldsymbol{H}}\mathcal{L} &= \boldsymbol{W}\boldsymbol{Y},\\ \nabla_{\boldsymbol{W}}\mathcal{L} &= \boldsymbol{H}\boldsymbol{Y}^\top,\\ \nabla_{\boldsymbol{b}}\mathcal{L} &= \tfrac{\gamma C-\gamma-1}{C}\mathbf{1}_C,\end{aligned} \tag{4}$$

and the corresponding learning dynamics with respect to the gradient flow is

$$\begin{aligned}\boldsymbol{H}'(t) &= \eta_1(t)\boldsymbol{W}(t)\boldsymbol{M},\\ \boldsymbol{W}'(t) &= \eta_2(t)\boldsymbol{H}(t)\boldsymbol{M}^\top,\\ \boldsymbol{b}'(t) &= -\eta_2(t)\tfrac{\gamma C-\gamma-1}{C}\mathbf{1}_C,\end{aligned} \tag{5}$$

where $\boldsymbol{M} = -\boldsymbol{Y} = \frac{1}{CN}((1+\gamma)(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top) - \gamma \mathbf{1}_C\mathbf{1}_{CN}^\top)$, $\eta_1(t)$ and $\eta_2(t)$ are the time-varying part[7] of the learning rate of the features $\boldsymbol{H}$ and weights $(\boldsymbol{W}, \boldsymbol{b})$, respectively. The reason for introducing different learning rates is that the representation $\boldsymbol{H}$ is a result of the interaction between a number of nonlinear layers rather than a completely free variable like network parameters. This implies that even if we use the same learning rate to optimize all network parameters, the feature $\boldsymbol{H}$ assumed to be a free optimization variable is almost impossible to be optimized at this learning rate[8]. Moreover, we consider time-varying rates $\eta_1 = \eta_1(t)$ and $\eta_2 = \eta_2(t)$ that are usually adopted in practical implementations, such as the cosine annealing decay (Loshchilov and Hutter, 2017). As can be seen, the dynamics of features $\boldsymbol{H}$ and prototypes $\boldsymbol{W}$ are independent of the bias term $\boldsymbol{b}$, thus we can analyze the dynamics of $\boldsymbol{H}$ and $\boldsymbol{W}$ jointly, and analyze $\boldsymbol{b}$ independently:

Theorem 2 (**Dynamics of Features, Prototypes and Biases without Constraints**). *Consider the continual gradient flow in Equation 5, let $\boldsymbol{Z}(t) = (\boldsymbol{H}(t), \boldsymbol{W}(t))$, if $\eta_1(t_1)\eta_2(t_2) = \eta_1(t_2)\eta_2(t_1)$ for any $t_1, t_2 \geq 0$, we have the following closed-form dynamics*

$$\begin{aligned}\boldsymbol{Z}(t) =& \Pi_1^+ \boldsymbol{Z}_0 \left(\alpha_1^+(t)\boldsymbol{C}(t) + \beta_1^+(t)\boldsymbol{I}_{C(N+1)}\right)\\ &+\Pi_1^- \boldsymbol{Z}_0 \left(\alpha_1^-(t)\boldsymbol{C}(t) + \beta_1^-(t)\boldsymbol{I}_{C(N+1)}\right)\\ &+\Pi_2^+ \boldsymbol{Z}_0 \left(\alpha_2^+(t)\boldsymbol{C}(t) + \beta_2^+(t)\boldsymbol{I}_{C(N+1)}\right)\\ &+\Pi_2^- \boldsymbol{Z}_0 \left(\alpha_2^-(t)\boldsymbol{C}(t) + \beta_2^-(t)\boldsymbol{I}_{C(N+1)}\right) + \Pi_3 \boldsymbol{Z}_0,\end{aligned} \tag{6}$$

and

$$\boldsymbol{b}(t) = \boldsymbol{b}_0 + \frac{(1+\gamma-\gamma C)\zeta_2(t)}{C}\mathbf{1}_C, \tag{7}$$

where α_1^ϵ, α_2^ϵ, β_1^ϵ and β_2^ϵ for $\epsilon \in \{\pm\}$ are the scalars that only depend on C, N, γ, $\eta_1(0)$ and $\eta_2(0)$ (where the detailed forms of these scalars can be seen in the appendix), $\boldsymbol{Z}_0 = (\boldsymbol{H}_0, \boldsymbol{W}_0)$,

7. In this paper, we are interested in investigating the effect of changing the learning rate over time. Thus, $\eta(t)$ is not the complete rate, but rather the time varying part of the learning rate. Specifically, for the example gradient flow $x_t' = \frac{\mathrm{d}x_t}{\mathrm{d}t} = \eta(t)\nabla f(x_t)$ considered in this paper is a continuous-time approximation of gradient descent in the limit of an infinitesimally small time step Δ_t. Specifically, the corresponding of gradient descent in discrete time can be formulated as $x_{t+\Delta_t} = x_t - \Delta_t \cdot \eta(t)\nabla f(x_t)$, where $\Delta_t \cdot \eta(t)$ denotes the complete learning rate, and Δ_t is the time step.
8. In this paper, we mainly assume that $\eta_1(t_1)\eta_2(t_2) = \eta_1(t_2)\eta_2(t_1)$ for any pair of values of t, t_1 and t_2. This condition will be satisfied if and only if $\eta_1(t)$ is a scaled version of $\eta_2(t)$, *i.e.*, $\eta_1(t) = s \cdot \eta_2(t)$

$\boldsymbol{C}(t) = \begin{pmatrix} \zeta_1(t)\boldsymbol{I}_{CN} & 0 \\ 0 & \zeta_2(t)\boldsymbol{I}_C \end{pmatrix}$, $\zeta_1(t) = \int_0^t \eta_1(\tau)\mathrm{d}\tau$, $\zeta_2(t) = \int_0^t \eta_2(\tau)\mathrm{d}\tau$, Π_1^ϵ, Π_2^ϵ *and* Π_3 *for* $\epsilon \in \{\pm\}$ *are orthogonal projection operators onto the following respective eigenspaces:*

$$\begin{aligned} \mathcal{E}_1^\epsilon &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = \epsilon \cdot \tfrac{1}{\sqrt{N}}(\boldsymbol{W} \otimes \mathbf{1}_N^\top), \boldsymbol{W}\mathbf{1}_C = 0\}, \\ \mathcal{E}_2^\epsilon &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = \epsilon \cdot \tfrac{1}{\sqrt{N}}\boldsymbol{h}\mathbf{1}_{CN}^\top, \boldsymbol{W} = \boldsymbol{h}\mathbf{1}_C^\top, \boldsymbol{h} \in \mathbb{R}^p\}, \\ \mathcal{E}_3 &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H}(\boldsymbol{I}_C \otimes \mathbf{1}_N) = 0, \boldsymbol{W} = 0\}. \end{aligned} \tag{8}$$

Remark Note that $\mathcal{E}_1^\epsilon$, $\mathcal{E}_2^\epsilon$ and $\mathcal{E}_3$ are orthogonal to each other, $\mathcal{E}_1^+$ (or $\mathcal{E}_1^-$) denotes the subspace where all features are in the same (or opposite) direction of their corresponding prototypes while the mean of prototypes is zero, $\mathcal{E}_2^\epsilon$ denotes the subspace where all features and prototypes collapse respectively into two scaled versions of the same unit vector, and $\mathcal{E}_3$ denotes the subspace where the mean of all features from the same class is zero with all prototypes being zero. For classification tasks, we expect the features align to their corresponding prototypes with a cosine similarity of 1, *i.e.*, the solution in $\mathcal{E} = \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = k\boldsymbol{W} \otimes \mathbf{1}_N^\top, \boldsymbol{W} \in \mathbb{R}^{p\times C}, k \in \mathbb{R}^+\}$ that implies two manifestations of Neural Collapse: *within-class variability collapse* and *convergence to self-duality* (Papyan et al., 2020). In the following, we prove that the normalized dynamics $\frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|}$ under the unhinged loss will converge to a solution in $\mathcal{E}$:

Corollary 3 (**Convergence in the Unconstrained Case**). *Under the conditions and notation of Theorem 2, let* $s = \frac{\eta_1(0)}{\eta_2(0)}$. *If* $0 < \gamma < \frac{2}{C-2}$ *(where* $C > 2$*) or* $C = 2$, *and* $\lim_{t\to\infty} \zeta_1(t) = \infty$, *the gradient flow (as in Equation* (5)*) will behave as:*

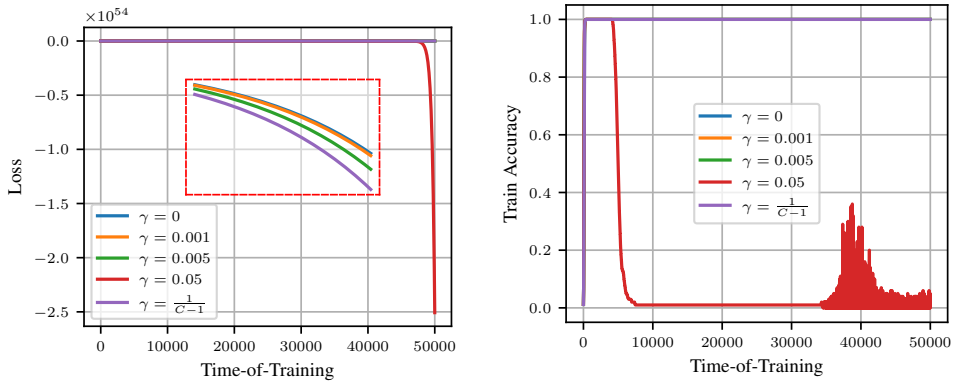
$$e^{-\frac{(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}}\boldsymbol{Z}(t) = \overline{\boldsymbol{Z}} + \boldsymbol{\Delta}(t), \tag{9}$$

where $\overline{\boldsymbol{Z}} = \left(\frac{1+\sqrt{s}}{2}\boldsymbol{H}_1^+ + \frac{1-\sqrt{s}}{2}\boldsymbol{H}_1^-, \frac{1+\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^+ - \frac{1-\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^-\right)$, $(\boldsymbol{H}_1^+, \boldsymbol{W}_1^+) = \Pi_1^+\boldsymbol{Z}_0$, $(\boldsymbol{H}_1^-, \boldsymbol{W}_1^-) = \Pi_1^-\boldsymbol{Z}_0$, *and the residual term* $\boldsymbol{\Delta}(t)$ *decreases as* $\|\boldsymbol{\Delta}(t)\| = O\left(e^{\frac{\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\cdot\max\{-\gamma C,(C-2)\gamma-2\}}\right)$, *and so the normalized* $\boldsymbol{Z}(t)$ *converges to* $\frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|}$ *in*

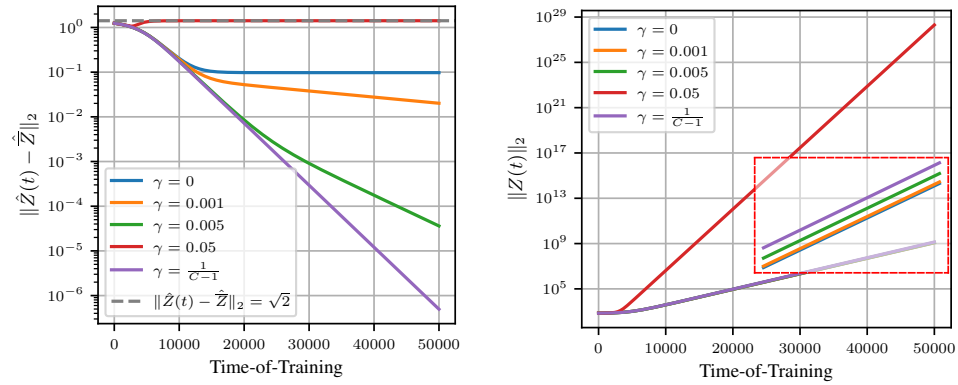
$$\left\|\frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|} - \frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|}\right\| = O\left(e^{\frac{\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\cdot\max\{-\gamma C,(C-2)\gamma-2\}}\right), \tag{10}$$

which indicates $\lim_{t\to\infty}\frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|} = \frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|} \in \mathcal{E}$. *Moreover, if* $\gamma \neq \frac{1}{C-1}$, $\lim_{t\to\infty}\frac{\max_i b_i(t)}{\min_i b_i(t)} = 1$.

The corollary above shows that even without any mandatory constraints, the gradient flow under the unhinged loss will converge to a solution $\frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|}$ that belongs to $\mathcal{E}$ and can be determined by the initialization of $\boldsymbol{Z}_0$ and the ratio s. It is worth noting, however, that this solution does not conform to the geometric structure of Neural Collapse. Neural collapse typically entails the formation of a simplex equiangular tight frame, as exemplified in the case of CE (Papyan et al., 2020), MSE (Han et al., 2022), margin-based losses (Zhou et al., 2022d), and other losses (Zhou et al., 2022b). This result suggests that *not all losses will lead to Neural Collapse solutions.*



(a) Train Loss (b) Train Accuracy



(c) $\|\hat{\boldsymbol{Z}}(t) - \hat{\overline{\boldsymbol{Z}}}\|_2$ (d) $\|\boldsymbol{Z}(t)\|_2$

Figure 1: Verification of the behavior of gradient descent iterates in Equation (5) with $\gamma \in \{0, 0.05, 0.005, 0.001, \frac{1}{C-1}\}$, where we set $p = 512$, $C = 100$, $N = 10$, $\eta_1(t) = \eta_2(t) = 0.1$ (*i.e.*, $s = \frac{\eta_1(0)}{\eta_2(0)} = 1$, thus $\overline{\boldsymbol{Z}} = \Pi_1^+ \boldsymbol{Z}_0$ according to Corollary 3), and then randomly initialize $\boldsymbol{H}_0$ and $\boldsymbol{W}_0$. The red box in the figures represents the zoomed-in view of the last 2,000 epochs. (a) The training loss. (b) The training accuracy with the prediction rule $\arg\max_c \boldsymbol{w}_c^\top \boldsymbol{h}$. As expected, the features align to their corresponding prototypes when $\gamma < \frac{2}{C-2}$. (c) The distance between $\hat{\boldsymbol{Z}}(t) = \boldsymbol{Z}(t)/\|\boldsymbol{Z}(t)\|_2$ and $\hat{\overline{\boldsymbol{Z}}} = \overline{\boldsymbol{Z}}/\|\overline{\boldsymbol{Z}}\|_2$. As expected in Equation (10), the convergence rate is exponential when $0 < \gamma < \frac{2}{C-2}$, and will be fastest if $\gamma = \frac{1}{C-1}$. (d) The norm of $\boldsymbol{Z}(t)$, which increases exponentially. As can be noticed, $\hat{\boldsymbol{Z}}(t)$ does not converge to $\hat{\overline{\boldsymbol{Z}}}$ but tend to be orthogonal to $\hat{\overline{\boldsymbol{Z}}}$ when $\gamma = 0.05 > \frac{2}{C-2}$, that is, $\lim_{t\to\infty} \|\hat{\boldsymbol{Z}}(t) - \hat{\overline{\boldsymbol{Z}}}\|_2 = \sqrt{2}$. Moreover, the exponential curve of $\|\boldsymbol{Z}(t)\|_2$ will be $O(e^{3.95\times\sqrt{10}\times 10^{-4}t})$ when $\gamma = 0.05$, which is faster than those curves of the rate roughly around $O(e^{\sqrt{10}\times 10^{-4}t})$ when $\gamma = 0, 0.001, 0.005$ and $\frac{1}{C-1} = \frac{1}{99}$. This key distinction elucidates why the curves overlap when $\gamma = 0, 0.001, 0.005$, and $\frac{1}{C-1}$.

In addition, the rate of convergence is exponential as a function of the cumulative learning rates, *i.e.*, $O\left(e^{\frac{\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\cdot\max\{-\gamma C,(C-2)\gamma-2\}}\right)$, which indicates that the convergence of updating both features and prototypes by gradient descent is much faster than $O(1/\log t)$ that only updates prototypes (linear predictors) on linearly separable data (Soudry et al., 2018). In a sense, this convergence rate may help explain why training deep neural networks usually takes only several hundred or thousand epochs. Moreover, if $\gamma = \frac{1}{C-1}$, we can obtain the fastest convergence of Equation (10), that is, $\left\| \frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|} - \frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|} \right\| = O\left(e^{-\frac{\sqrt{\zeta_1(t)\zeta_2(t)}}{(C-1)\sqrt{N}}}\right)$. As shown in Figure 1, if we set $\eta_1(t) = \eta$ as a constant learning rate, then $\zeta_1(t) = \eta t \to \infty$ as $t \to \infty$, and the gradient flow in Equation (5) exhibits an exponential convergence rate of $\frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|}$ to $\frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|}$. However, this will lead to an exponential increase in the norm of features and prototypes, with the rate $e^{\frac{(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}}$. Such growth is almost unbearable for the practical training of DNNs. Therefore, in what follows, we consider to limit the excessive growth of these norms.

3.2 The Regularized Case

In this subsection, we focus on a regularized optimization problem that introduces ℓ_2 regularization (also known as "weight decay") on features, prototypes, and biases. This regularization term helps to prevent the norms of these variables from growing too large during training. The optimization problem is defined as follows:

$$\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}} \operatorname{Tr}(\boldsymbol{Y}^\top \boldsymbol{W}^\top \boldsymbol{H}) + \tfrac{\gamma C-\gamma-1}{C}\mathbf{1}_C^\top \boldsymbol{b} + \tfrac{\lambda_1}{2}\|\boldsymbol{H}\|_F^2 + \tfrac{\lambda_2}{2}\|\boldsymbol{W}\|_F^2 + \tfrac{\lambda_2}{2}\|\boldsymbol{b}\|_2^2, \tag{11}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, λ_1 and λ_2 are trade-off parameters[9] that control the regularization strength.

Taking the partial derivative with respect to $\boldsymbol{H}$, $\boldsymbol{W}$, and $\boldsymbol{b}$, we have

$$\begin{aligned}
\nabla_{\boldsymbol{H}}\mathcal{L} &= \boldsymbol{W}\boldsymbol{Y} + \lambda_1\boldsymbol{H},\\
\nabla_{\boldsymbol{W}}\mathcal{L} &= \boldsymbol{H}\boldsymbol{Y}^\top + \lambda_2\boldsymbol{W},\\
\nabla_{\boldsymbol{b}}\mathcal{L} &= \tfrac{\gamma C-\gamma-1}{C}\mathbf{1}_C + \lambda_2\boldsymbol{b},
\end{aligned}$$

and the corresponding learning dynamics following the gradient flow with time-vary learning rates, ℓ_2 regularization can be formulated as

$$\begin{aligned}
\boldsymbol{H}'(t) &= \eta_1(t)\boldsymbol{W}(t)\boldsymbol{M} - \lambda_1\eta_1(t)\boldsymbol{H}(t),\\
\boldsymbol{W}'(t) &= \eta_2(t)\boldsymbol{H}(t)\boldsymbol{M}^\top - \lambda_2\eta_2(t)\boldsymbol{W}(t),\\
\boldsymbol{b}'(t) &= -\eta_2(t)\tfrac{\gamma C-\gamma-1}{C}\mathbf{1}_C - \lambda_2\eta_2(t)\boldsymbol{b}(t).
\end{aligned} \tag{12}$$

The dynamics of this regularized gradient flow can be proved as follows

9. To draw general conclusions, we consider a different regularization strength for each component.

Theorem 4 (**Dynamics of Features, Prototypes, and Biases under Weight Decay**). *Consider the continual gradient flow in Equation 12, let* $\boldsymbol{Z}(t) = (\boldsymbol{H}(t), \boldsymbol{W}(t))$. *If* $\eta_1(t_1)\eta_2(t_2) = \eta_1(t_2)\eta_2(t_1)$ *for any* $t_1, t_2 \geq 0$, *we have the following closed-form dynamics:*

$$\begin{aligned}\boldsymbol{Z}(t) =& \Pi_1^+ \boldsymbol{Z}_0 \begin{pmatrix} a_1^+(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_1^+(t)\boldsymbol{I}_C \end{pmatrix} + \Pi_1^- \boldsymbol{Z}_0 \begin{pmatrix} a_1^-(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_1^-(t)\boldsymbol{I}_C \end{pmatrix} \\ &+ \Pi_2^+ \boldsymbol{Z}_0 \begin{pmatrix} a_2^+(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_2^+(t)\boldsymbol{I}_C \end{pmatrix} + \Pi_2^- \boldsymbol{Z}_0 \begin{pmatrix} a_2^-(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_2^-(t)\boldsymbol{I}_C \end{pmatrix} \\ &+ \Pi_3 \boldsymbol{Z}_0 \begin{pmatrix} a_3(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_3(t)\boldsymbol{I}_C \end{pmatrix},\end{aligned} \tag{13}$$

and

$$\boldsymbol{b}(t) = \phi(t)\left(\boldsymbol{b}_0 + \tfrac{1+\gamma-\gamma C}{C}\psi(t)\mathbf{1}_C\right), \tag{14}$$

where $\Pi_1^+\boldsymbol{Z}_0$, $\Pi_1^-\boldsymbol{Z}_0$, $\Pi_1^+\boldsymbol{Z}_0$, $\Pi_1^-\boldsymbol{Z}_0$, *and* $\Pi_3\boldsymbol{Z}_0$ *follow the definition in Theorem 2,* a_1^ϵ, a_2^ϵ, b_1^ϵ, b_2^ϵ, a_3, *and* b_3 *for* $\epsilon \in \{\pm\}$ *are the scalars that depend only on* C, N, γ, λ_1, λ_2, η_1, *and* η_2 *(where the detailed forms can be seen in B),* $\phi(t) = \exp(-\lambda_2 \int_0^t \eta_2(\tau)\mathrm{d}\tau)$, *and* $\psi(t) = \int_0^t \zeta_2(\tau)\exp(\lambda_2 \int_0^\tau \eta_2(s)\mathrm{d}s)\mathrm{d}\tau$.

The convergence under the regularized case can also be derived as:

Corollary 5 (**Convergence in the ℓ_2 Regularized Case**). *Under the conditions and notation of Theorem 4, let* $s = \frac{\eta_1(0)}{\eta_2(0)}$. *If* $0 < \gamma < \frac{2}{C-2}$ *(where* $C > 2$*) or* $C = 2$, $\lambda_1 = \lambda_2 = \lambda$, *and* $\lim_{t\to\infty}\zeta_1(t) = \infty$, *then there exist constants* $\pi_h^+, \pi_h^-, \pi_w^+, \pi_w^-$, *and* ω *only depending on* λ, γ, s, C, *and* N, *such that the gradient flow (as in Equation* (12)*) behaves as:*

$$\left\| \frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|} - \frac{\boldsymbol{Z}_\pi}{\|\boldsymbol{Z}_\pi\|} \right\| = O\left(e^{-\omega\zeta_2(t)}\right), \tag{15}$$

where $(\boldsymbol{H}_1^+, \boldsymbol{W}_1^+) = \Pi_1^+\boldsymbol{Z}_0$, $(\boldsymbol{H}_1^-, \boldsymbol{W}_1^-) = \Pi_1^-\boldsymbol{Z}_0$, *and* $\boldsymbol{Z}_\pi = (\pi_h^+\boldsymbol{H}_1^+ + \pi_h^-\boldsymbol{H}_1^-, \pi_w^+\boldsymbol{W}_1^+ + \pi_w^-\boldsymbol{W}_1^-)$.

Furthermore, we have the following convergence results for $\boldsymbol{Z}(t)$*:*

- *If* $\lambda > \frac{1+\gamma}{C\sqrt{N}}$, *then* $\lim_{t\to\infty}\|\boldsymbol{Z}(t)\| = 0$;
- *If* $\lambda = \frac{1+\gamma}{C\sqrt{N}}$, *then* $\lim_{t\to\infty}\boldsymbol{Z}(t) = \left(\boldsymbol{H}_1^+ + \frac{1-s}{1+s}\boldsymbol{H}_1^-, \boldsymbol{W}_1^+ - \frac{1-s}{1+s}\boldsymbol{W}_1^-\right)$;
- *If* $\lambda < \frac{1+\gamma}{C\sqrt{N}}$, *then* $\lim_{t\to\infty}\|\boldsymbol{Z}(t)\| = \infty$.

As can be seen, the features and classifier weights under the unhinged loss converge to the solution depending on the initialization with the form $\boldsymbol{Z}_\pi(t) = (\pi_h^+\boldsymbol{H}_1^+ + \pi_h^-\boldsymbol{H}_1^-, \pi_w^+\boldsymbol{W}_1^+ + \pi_w^-\boldsymbol{W}_1^-)$, where $\boldsymbol{H}_1^+ = \frac{1}{\sqrt{N}}(\boldsymbol{W}^+ \otimes \mathbf{1}_N^\top)$ and $\boldsymbol{H}_1^- = -\frac{1}{\sqrt{N}}(\boldsymbol{W}^- \otimes \mathbf{1}_N^\top)$. Thus, the solution implies the property of within-class variability collapse. To full encompass all properties of Neural Collapse (when the class number and the feature dimensionality p satisfies $C \leq p-1$), it is essential for $\pi_w^+\boldsymbol{W}_1^+ + \pi_w^-\boldsymbol{W}_1^-$ to form a Simplex ETF, and for $\frac{\pi_h^+}{\pi_w^+} = -\frac{\pi_h^-}{\pi_w^-}$ to hold, inducing the property of convergence to self-duality (i.e., the class means and linear

classifiers converge to each other, up to rescaling). For example, when $\lambda = \frac{1+\gamma}{C\sqrt{N}}$, Corollary demonstrates that the solution satisfies the property of convergence to self-duality.

Moreover, the results in Corollary 5 suggest that adding an appropriate weight decay on both features and prototypes can avoid impractical effects, since the norm of $\boldsymbol{Z}(t)$ shrinking to 0 or diverging toward infinity will significantly affect the training of DNNs. Several recent works (Zhu et al., 2021; Zhou et al., 2022a) described that the features are implicitly penalized, but this implicit penalization may be fragile when using the unhinged loss (as depicted in Figure 19). Consequently, we emphasize the importance of *adding explicit regularization to features, rather than relying solely on the implicit penalization attached by other components like batch normalization and weight decay*. Explicit feature regularization also plays a role in mitigating minority collapse (Fang et al., 2021b). Firstly, minority collapse leads to features of the minority classes approaching 0, since the minimization of the objective pays too much emphasis on enlarging the feature norms of the majority classes. In this context, explicit feature regularization can effectively restrain the excessive growth of feature norms of the majority classes. Secondly, explicit feature regularization in some sense intuitively reduces the energy of features in the optimization program (Fang et al., 2021b), thereby shrinking the feasible domain of features. This further mitigates the imbalance of feature norms between majority and minority classes. In a nutshell, we need to limit the growth of feature norms.

3.3 The Prototype-anchored Case

We further consider the prototype-anchored case in which the class prototypes $\boldsymbol{W}$ are fixed[10] into some specific structures (*e.g.*, a simplex ETF) during the training process and the features $\boldsymbol{H}$ are with ℓ_2 regularization. Then, the dynamics of $\boldsymbol{H}$ in Equation (11) will be formulated as the first-order non-homogeneous linear difference equation:

$$\boldsymbol{H}'(t) = \eta(t)\boldsymbol{W}\boldsymbol{M} - \lambda\eta(t)\boldsymbol{H}(t), \tag{16}$$

and the solution to the linear difference equation can be easily derived as follows

Theorem 6. *Consider the continual gradient flow (Equation* (16)*) in which the prototypes* $\boldsymbol{W}$ *are fixed, we have the closed-form dynamics:*

$$\boldsymbol{H}(t) = e^{-\lambda\int_0^t \eta(\tau)\mathrm{d}\tau}\boldsymbol{H}(0) + \frac{1 - e^{-\lambda\int_0^t \eta(\tau)\mathrm{d}\tau}}{\lambda}\boldsymbol{W}\boldsymbol{M}, \tag{17}$$

which further indicates that $\left\|\boldsymbol{H}(t) - \frac{1}{\lambda}\boldsymbol{W}\boldsymbol{M}\right\| = O\left(e^{-\lambda\int_0^t \eta(\tau)\mathrm{d}\tau}\right)$ *when* $\lim_{t\to\infty} e^{-\lambda\int_0^t \eta(\tau)\mathrm{d}\tau} = 0$.

When the time-varying learning rate satisfies that $\lim_{t\to\infty} e^{-\lambda\int_0^t \eta(\tau)\mathrm{d}\tau} = 0$, then $\boldsymbol{H}(t)$ converges to $\frac{1}{\lambda}\boldsymbol{W}\boldsymbol{M}$. That is, the unhinged loss in the prototype-anchored case coincides with a feature alignment task such that each feature $\boldsymbol{h}_{i,c}$ in class c aligns to $\frac{1}{\lambda CN}(\boldsymbol{w}_c - \gamma\sum_{j\neq c}\boldsymbol{w}_j)$. It's worth noting that when $\lambda = 0$, the behavior of $\boldsymbol{H}(t)$ can be described as $\boldsymbol{H}(t) = \boldsymbol{H}(0) + \int_0^t \eta(\tau)\mathrm{d}\tau\boldsymbol{W}\boldsymbol{M}$, which is predominantly influenced by the term $\boldsymbol{W}\boldsymbol{M}$ when the integral $\int_0^t \eta(\tau)d\tau$ becomes significantly large.

10. That is, $\boldsymbol{W}$ is not updated, which can be done by simply setting W.require_grad = $False$ in PyTorch

As depicted in Figure 2, prototype-anchored learning demonstrates its effectiveness in mitigating training instability by transforming the classification problem into a feature alignment task. In this context, the unhinged loss with PAL yields results that are not only comparable but, in some cases, even superior to those achieved with the Cross-Entropy (CE) loss. This underscores the practicality of the unhinged loss as a valuable loss function for training classification models.

3.4 The Spherical Constrained Case

We consider another constrained case in which features are restricted on the p-sphere $\mathbb{S}^{p-1} = \{\boldsymbol{x} : \|\boldsymbol{x}\|_2 = 1, \boldsymbol{x} \in \mathbb{R}^p\}$ by explicitly performing ℓ_2 normalization[11], and we fix the prototypes[12] to satisfy $\boldsymbol{W}\mathbf{1}_C = 0$[13], then the optimization problem in Equation (2) can be reformulated as

$$\min_{\boldsymbol{H}} -\frac{1+\gamma}{CN}\operatorname{Tr}((\boldsymbol{I}_C \otimes \mathbf{1}_N)\boldsymbol{W}^\top \hat{\boldsymbol{H}}), \tag{18}$$

where $\hat{\boldsymbol{H}} = (\hat{\boldsymbol{h}}_{1,1}, ..., \hat{\boldsymbol{h}}_{c,N})$ and $\hat{\boldsymbol{h}} = \frac{\boldsymbol{h}}{\|\boldsymbol{h}\|_2}$ denotes the ℓ_2-normalized vector.

Take the partial derivative with respect to $\boldsymbol{H}$, then the discrete dynamical system based on gradient descent is formulated as

$$\boldsymbol{H}(t+1) = \boldsymbol{H}(t) + \tfrac{(1+\gamma)\eta(t)}{CN}\left(\tfrac{\partial \hat{\boldsymbol{H}}}{\partial \boldsymbol{H}}\big|_{\boldsymbol{H}=\boldsymbol{H}(t)}\right)^\top \boldsymbol{W}(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top), \tag{19}$$

where $(\frac{\partial \hat{\boldsymbol{H}}}{\partial \boldsymbol{H}})^\top$ is a vector-wise operator, and for any vector $\boldsymbol{h}_{i,c}$ in $\boldsymbol{H}' \in \mathbb{R}^{p\times CN}$, we have $(\frac{\partial \hat{\boldsymbol{H}}}{\partial \boldsymbol{H}})^\top \boldsymbol{h}_{i,c} = \left(\frac{1}{\|\boldsymbol{h}_{i,c}\|_2}(\boldsymbol{I}_p - \hat{\boldsymbol{h}}_{i,c}\hat{\boldsymbol{h}}_{i,c}^\top)\boldsymbol{h}'_{i,c}\right)$.

Despite the fact that the optimization objective in Equation (18) does not show convexity, Lipschitzness, and β-smoothness on $\boldsymbol{H}$ due to the ℓ_2 normalization operator, we prove that the normalized features which obey the gradient descent iterates in Equation (19) can still converge to their corresponding normalized prototypes, *i.e.*, achieve the global minimum of Equation (18):

Theorem 7 (**Convergence in the Spherical Constrained Case**). *Considering the discrete dynamics in Equation* (19)*, if* $\hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(0) > -1$ *for any* $i \in [N]$ *and* $c \in [C]$*, the learning rate* $\eta(t)$ *satisfies that* $\frac{\eta(t)}{\|\boldsymbol{h}_{i,c}(t)\|_2}$ *is non-increasing,* $\frac{\eta(0)(1+\gamma)}{CN\|\boldsymbol{h}_{i,c}(0)\|_2} \le \frac{1}{\|\boldsymbol{w}_c\|_2}$, $\lim_{t\to\infty}\frac{\eta(t+1)}{\eta(t)} = 1$*, and there exists a constant* $\varepsilon > 0$*, s.t.,* $\eta(t) > \varepsilon$*, then we have*

$$\lim_{t\to\infty}\left\|\hat{\boldsymbol{H}}(t) - \hat{\boldsymbol{W}}(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top)\right\| = 0, \tag{20}$$

and further if $\lim_{t\to\infty}\|\boldsymbol{H}(t)\| < \infty$*, then there exists a constant* $\mu > 0$*, such that the error above shows exponential decrease:*

$$\left\|\hat{\boldsymbol{H}}(t) - \hat{\boldsymbol{W}}(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top)\right\| = O(e^{-\mu t}). \tag{21}$$

Moreover, if $\hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(0) = -1$*, then* $\boldsymbol{h}_{i,c}(t) = \boldsymbol{h}_{i,c}(0)$.

11. ℓ_2 normalization can also prevent arithmetic overflow or underflow occurring in the training of DNNs.
12. The relevant studies are still few and often require some strict assumptions since the learning dynamics is very complicated when $\boldsymbol{w}$ participates the optimization process with both feature and prototypes normalization. In this paper, we are going to try a more concise theoretical analysis with fixed prototypes.
13. This aims to simplify Equation (1) as the objective of feature alignment, that is, $L_\gamma(\boldsymbol{W}\hat{\boldsymbol{h}}, y) = -\boldsymbol{w}_y^\top \hat{\boldsymbol{h}} + \gamma\sum_{j\neq y}\boldsymbol{w}_j^\top \hat{\boldsymbol{h}} = \frac{(1+\gamma)\|\boldsymbol{w}\|_2}{2}(\|\hat{\boldsymbol{h}} - \hat{\boldsymbol{w}}_y\|_2^2 - 2)$, and the global minimum will be obtained at $\hat{\boldsymbol{h}} = \hat{\boldsymbol{w}}_y$.

Remark As shown in the theorem above, we prove that gradient descent exerts a bias that steers the normalized features toward the global minimizer $\hat{\boldsymbol{W}}(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top)$, achieving exponential convergence under certain favorable conditions. This global solution shows two properties of Neural Collapse: within-class variability collapse and convergence to self-duality. If $\hat{W}$ also forms a simplex ETF, the global solution will exhibit all properties of Neural Collapse. Additionally, we establish that if the inner product between $\hat{\boldsymbol{w}}_c$ and $\hat{\boldsymbol{h}}_{i,c}(0)$ (i.e., the cosine similarity between $\boldsymbol{w}_c$ and $\boldsymbol{h}_{i,c}(0)$) is -1, then $\boldsymbol{h}_{i,c}(t) = \boldsymbol{h}_{i,c}(0)$, as the gradient induced by ℓ_2 normalization will be 0. More details can be found in the proof of Theorem 7. Therefore, we just analyze the case where the inner product $\hat{\boldsymbol{w}}_c$ and $\hat{\boldsymbol{h}}_{i,c}(0)$ is strictly greater than -1.

3.5 The NTK-Invariant Case

We further perform some analysis in the case where the neural tangent kernel $\nabla_{\boldsymbol{\Theta}} \boldsymbol{H}^\top \nabla_{\boldsymbol{\Theta}} \boldsymbol{H}$ (Jacot et al., 2018; Chizat et al., 2019; Yang and Hu, 2020, 2021) is assumed to be invariant during training. We prove that the unhinged loss has more potential to derive exact dynamics. Specifically, consider the last-layer feature $\boldsymbol{h}_{i,c} = \boldsymbol{f}_{\boldsymbol{\Theta}}(\boldsymbol{x}_{i,c})$ extracted from the example $\boldsymbol{x}_{i,c}$. When DNNs are trained using gradient descent to minimize the composition $\mathcal{L} \circ \boldsymbol{W} \circ \boldsymbol{f}_{\boldsymbol{\Theta}}$, we have

$$\begin{aligned} \boldsymbol{\Theta}(t+1) - \boldsymbol{\Theta}(t) &= -\eta \nabla_{\boldsymbol{\Theta}} \mathcal{L}\big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}(t)}, \\ \boldsymbol{W}(t+1) - \boldsymbol{W}(t) &= -\eta \nabla_{\boldsymbol{W}} \mathcal{L}\big|_{\boldsymbol{W}=\boldsymbol{W}(t)}. \end{aligned} \tag{22}$$

According to the first-order Taylor expansion, we obtain

$$\boldsymbol{H}(t+1) - \boldsymbol{H}(t) \approx \nabla_{\boldsymbol{\Theta}} \boldsymbol{H}^\top [\boldsymbol{\Theta}(t+1) - \boldsymbol{\Theta}(t)]. \tag{23}$$

For the unhinged loss, we can obtain the following closed-form dynamics

Theorem 8. *Assume that the neural tangent kernel $\nabla_{\boldsymbol{\Theta}} \boldsymbol{H}^\top \nabla_{\boldsymbol{\Theta}} \boldsymbol{H}$ remains invariant during iterations. Let $\boldsymbol{z}(t)$ denote the row-first vectorization of $\begin{pmatrix} \boldsymbol{H}(t) & 0 \\ 0 & \boldsymbol{W}(t) \end{pmatrix}$, $\boldsymbol{B} = \begin{pmatrix} 0 & \boldsymbol{M}^\top \\ \boldsymbol{M} & 0 \end{pmatrix}$, and $\boldsymbol{A} = \begin{pmatrix} 0 & \nabla_{\boldsymbol{\Theta}} \boldsymbol{H}^\top \nabla_{\boldsymbol{\Theta}} \boldsymbol{H} \\ \mathbf{I}_C & 0 \end{pmatrix}$. Considering the eigendecomposition $\boldsymbol{A} = \boldsymbol{U}_{\boldsymbol{A}} \boldsymbol{\Lambda}_{\boldsymbol{A}} \boldsymbol{U}_{\boldsymbol{A}}^{-1}$ and $\boldsymbol{B} = \boldsymbol{U}_{\boldsymbol{B}} \boldsymbol{\Lambda}_{\boldsymbol{B}} \boldsymbol{U}_{\boldsymbol{B}}^{-1}$, we have*

$$\boldsymbol{C}\boldsymbol{z}(t) = \exp[(\boldsymbol{\Lambda}_{\boldsymbol{A}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{B}})t] \boldsymbol{C}\boldsymbol{z}(0), \tag{24}$$

where $\boldsymbol{C} = \boldsymbol{U}_{\boldsymbol{B}}^{-1} \boldsymbol{U}_{\boldsymbol{A}}^{-1} \otimes \mathbf{I}$.

Though Theorem 8 does not imply properties of Neural Collapse, it reveals a specific dynamics under the unhinged loss, which exhibits an exponential convergence rate, contingent upon the assumption that the neural tangent kernel remains constant. To demonstrate the mathematical simplicity of the unhinged loss, we consider the dynamics of the MSE loss $\mathcal{L} = \frac{1}{CN} \|\boldsymbol{W}^\top \boldsymbol{H} + \boldsymbol{b}\mathbf{1}_{CN}^\top - \mathbf{I} \otimes \mathbf{1}_N^\top\|_F^2$ for comparison, and similarly derive that

$$\boldsymbol{H}'(t) = \nabla_{\boldsymbol{\Theta}} \boldsymbol{H}^\top \nabla_{\boldsymbol{\Theta}} \boldsymbol{H} \boldsymbol{W}(t) \boldsymbol{C}(t), \quad \boldsymbol{W}'(t) = \boldsymbol{H}(t) \boldsymbol{C}(t)^\top, \quad \boldsymbol{b}'(t) = \boldsymbol{C}(t) \mathbf{1}_{CN}, \tag{25}$$

where $\boldsymbol{C}(t) = \frac{2\eta}{CN}(\boldsymbol{W}(t)^\top \boldsymbol{H}(t) + \boldsymbol{b}(t)\mathbf{1}_{CN}^\top - \mathbf{I}_C \otimes \mathbf{1}_N^\top)$. As can be seen, the above gradient flow is nonlinear and then intractable to analyze the exact dynamics though $\nabla_{\boldsymbol{\Theta}} \boldsymbol{H}^\top \nabla_{\boldsymbol{\Theta}} \boldsymbol{H}$ stays constant.

Through the above analysis, it is clear to see that the unhinged loss offers more mathematical opportunities than the MSE loss to analyze the closed-form dynamics under the assumption of an invariant neural tangent kernel.

4. Insights and Experiments

In this section, we provide some insights into better training DNNs according to the conclusions in Section 3. We then corroborate our theoretical results and insights with extensive experiments. More specifically, in Section 4.1, we propose to use prototype-anchored learning (PAL) as a means of resolving the instability issues that arise during training with the unhinged loss. In Section 4.2, we conduct experiments to highlight the benefits of explicit feature regularization on imbalanced learning and out-of-distribution (OOD) detection. In these experiments, we employ the unhinged loss with PAL and the CE loss as training objective. In Section 4.3, we propose the rescaling learning rates (RLR) with feature norms for the spherical case to address the problem of slow convergence resulting from feature normalization, which may also have implications for improving other methods of performing feature normalization. **More details and results can be found in Appendix C.**

4.1 The Unhinged Loss with Prototype-Anchored Learning

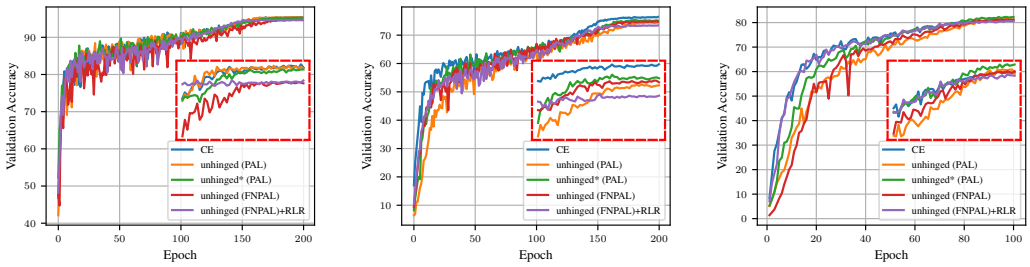
Since directly using the unhinged loss will lead to volatile effects, which is mainly reflected in the rapid increase of feature norms and the imbalance between class prototypes when training DNNs with the stochastic gradient method, as shown in Figure 19. Inspired by recent works (Zhou et al., 2022c; Kasarla et al., 2022; Yang et al., 2022) that use the Neural Collapse structure as an inductive bias (also called prototyping-anchored learning, PAL), we fix prototypes $\boldsymbol{W}$ as a simplex ETF during training, *i.e.*, $\boldsymbol{W}^\top \boldsymbol{W} = \frac{C}{C-1}\boldsymbol{I} - \frac{1}{C-1}\mathbf{1}\mathbf{1}^\top$[14].

Experimental Results We conduct experiments on widely-used classification datasets including CIFAR-10, CIFAR-100, and ImageNet-100. To mitigate training instability under the unhinged loss, we employ the Prototype-Anchored Learning (PAL) and Feature-Normalized PAL (FNPAL). As depicted in Figure 2, the results obtained by the unhinged loss with PAL and FNPAL variants demonstrate comparable or even better performance in comparison to CE. This substantiates the feasibility of utilizing the unhinged loss as a practical training objective for standard classification tasks.

4.2 Explicit Feature Regularization

In this paper, we directly consider explicit feature regularization to avoid excessive growth of feature norms, *i.e.*, adding the regularization term $\lambda \sum_{\boldsymbol{x}\in\mathcal{B}} \|\boldsymbol{f}_{\boldsymbol{\Theta}}(\boldsymbol{x})\|_2^2$ in the objective.

14. These prototypes $\boldsymbol{W}$ can be obtained using one of two methods: (i) By minimizing the objective $\sum_{i=1}^{C} \log \frac{\exp\left(s\hat{\boldsymbol{m}}_i^\top \hat{\boldsymbol{h}}_i\right)}{\sum_{j=1}^{C} \exp\left(s\hat{\boldsymbol{m}}_j^\top \hat{\boldsymbol{h}}_i\right)}$ and setting $\boldsymbol{W} = \hat{\boldsymbol{M}}$, where s is a scale factor (Zhou et al., 2022d); or (ii) By deriving them from the standard simplex ETF $\boldsymbol{M} = \sqrt{\frac{C}{C-1}}(\boldsymbol{I}_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top)$, *i.e.*, $\boldsymbol{W} = s\boldsymbol{U}\boldsymbol{M} \in \mathbb{R}^{p\times C}$, where $\boldsymbol{U} \in \mathbb{R}^{p\times C}$ $(p \geq C)$ is a partial orthogonal matrix (Papyan et al., 2020).



(a) CIFAR-10 (b) CIFAR-100 (c) ImageNet-100

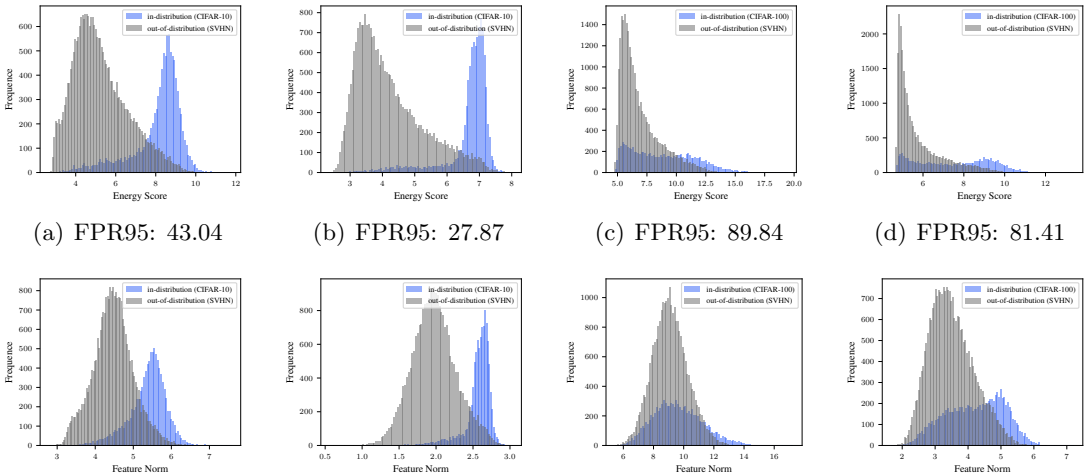
Figure 2: Validation accuracy of different loss functions on CIFAR-10, CIFAR-100, and ImageNet, where $*$ denotes training with explicit feature regularization. PAL and FNPAL denote the model trained with prototype-anchored learning (PAL) and feature-normalized and prototype-anchored learning (FNPAL) (Zhou et al., 2022c). The curve in the red box represents the zoomed-in curve of the last 50 epochs. As can be seen, DNNs trained with the unhinged loss can achieve comparative or even better performance than those of CE.

Table 2: Validation accuracies on long-tailed CIFAR-10/-100 with CE and different explicit feature regularization under Cross-Entropy (CE) and the unhinged loss (UL) with PAL. The imbalance ratio $\rho = \frac{\max_i n_i}{\min_i n_i}$ is the ratio between sample sizes of the most frequent and least frequent classes, and $\rho = 1$ denotes the original CIFAR-10/-100. $\lambda = 0$ denotes the model training with CE. All values are percentages. **Bold** numbers indicate the results that are better than the baseline vanilla CE or UL. The best results are underlined.

Dataset	Long-tailed CIFAR-10					Long-tailed CIFAR-100				
Imbalance Ratio	100	50	20	10	1	100	50	20	10	1
Vanilla CE	67.81	72.93	83.97	88.37	95.28	33.37	39.40	42.96	56.38	75.42
CE ($\lambda = 5e-6$)	**67.84**	72.85	83.17	**89.06**	95.27	**36.00**	**41.92**	**50.75**	**60.13**	**76.48**
CE ($\lambda = 1e-5$)	67.74	**76.14**	**84.17**	**89.19**	95.23	**36.61**	**42.36**	**49.21**	**58.91**	**77.34**
CE ($\lambda = 5e-5$)	**69.74**	**77.29**	**84.92**	**88.64**	**95.39**	**34.88**	**42.74**	**54.72**	**60.84**	**76.19**
Vanilla UL	72.60	77.20	85.46	89.29	95.23	39.20	45.81	54.66	59.92	75.37
UL ($\lambda = 1e-7$)	71.85	**78.80**	**85.76**	**89.68**	**95.33**	**41.39**	**46.13**	**55.25**	**61.15**	75.18
UL ($\lambda = 5e-7$)	**73.47**	**79.54**	**86.28**	**89.36**	95.14	**41.85**	**47.58**	54.00	**61.31**	**75.51**
UL ($\lambda = 1e-6$)	**73.71**	**79.40**	**86.12**	**89.43**	**95.50**	**40.85**	**47.07**	**55.71**	**61.47**	**75.84**
UL ($\lambda = 5e-6$)	71.13	**78.90**	85.33	89.20	95.10	**41.87**	**45.83**	**55.92**	**60.87**	**77.20**

Explicit regularization on features can significantly remedy over-confidence and even improve generalization. Moreover, adding explicit feature regularization can speed up the convergence of $\hat{\boldsymbol{H}}(t)$ to $\boldsymbol{W}\hat{\boldsymbol{M}}$ according to Theorem 6, as verified in Figure 15.

Experimental Results To validate the role of explicit feature regularization, we conduct experiments on two tasks: (i) Long-tailed recognition on benchmarks CIFAR-10-LT and CIFAR-100-LT with artificially created long-tailed settings; (ii) Out-of-distribution (OOD) detection between SVHN and CIFAR-10/-100. For long-tailed classification, we follow the



(a) FPR95: 43.04 (b) FPR95: 27.87 (c) FPR95: 89.84 (d) FPR95: 81.41

(e) FPR95: 52.10 (f) FPR95: 24.94 (g) FPR95: 95.54 (h) FPR95: 87.83

Figure 3: Distribution of energy scores (a-d) (Liu et al., 2020) and feature norms (e-h) from classification models trained without (a & c & e & g) or with (b & d & f & h) explicit feature regularization (EFR) ($\lambda = 1e-5$). (a & b & e & f) and (c & d & g & h) are from ResNet-18 (He et al., 2016) trained on CIFAR-10 and from ResNet-34 trained on CIFAR-100, respectively. As can be seen, EFR can improve the performance of OOD detection by alleviating the over-confidence of OOD samples and making the energy scores of ID samples more concentrated. More intuitively, comparing (f) to (e) and (h) to (g), EFR effectively limits the growth of feature norms and significantly improves the distinction between ID samples and OOD samples in feature norm.

controllable class imbalance strategy in (Cao et al., 2019) by reducing the number of training examples per class and keeping the validation set unchanged. As shown in Table 2, explicit feature regularization effectively improves the performance on long-tailed classification in most cases, even for normal classification. For ODD detection, we train ResNet-18 and ResNet-34 on in-distribution datasets CIFAR-10 and CIFAR-100, respectively, and then use SVHN as the OOD dataset to evaluate the performance.

4.3 Rescaled Learning Rate for the Spherical Case

In this subsection, we propose a strategy of rescaling the learning rate for the spherical case in which the features are performed ℓ_2 normalization.

As shown in Theorem 7 and its proof, the convergence behavior in the spherical constrained case is characterized by the term $\frac{\eta(t)}{\|\boldsymbol{h}_{i,c}(t)\|_2}$, when $\|\boldsymbol{h}_{i,c}\|_2$ is monotonically increasing and thus leads to a slower convergence rate. This result underscores the influence of the feature norm on the gradient dynamics in spherical case. Also, we observe that the training plots in the spherical case exhibit a slower rate in comparison to the CE loss in Figure 2. To bridge the performance of the unhinged loss to that of CE, we propose a straightforward approach—rescaling the learning rates based on feature norms, that is, scaling

Table 3: OOD detection performance using softmax-based (Hendrycks and Gimpel, 2016), energy-based (Liu et al., 2020), and feature norm-based approaches while model training with explicit feature regularization (EFL) (where $\lambda = 1e-5$ for CE and $\lambda = 5e-7$ for the unhinged loss (UL)). We use ResNet-18 and ResNet-34 to train on the in-distribution datasets CIFAR-10 and CIFAR-100, respectively. We then use SVHN (Netzer et al., 2011) as the OOD dataset to evaluate the performance of OOD detection. All values are percentages. ↑ indicates large values are better, and ↓ indicates smaller values are better. The best results are underlined.

Dataset $\mathcal{D}_{\text{in}}^{\text{test}}$	**Method**	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑
		Softmax-based / Energy-based / Feature Norm-based		
CIFAR-10	CE	52.09 / 43.04 / 52.10	91.67 / 91.94 / 89.54	84.11 / 82.80 / 77.06
	CE + EFL	**37.39** / **27.87** / **24.94**	**93.90** / **94.60** / **94.17**	**85.48** / **85.34** / **83.15**
	UL	53.29 / 53.33 / 52.34	87.72 / 87.73 / 89.45	74.86 / 74.83 / 79.05
	UL+EFL	**27.27** / **27.24** / **26.62**	**94.27** / **94.46** / **95.12**	**86.51** / **86.60** / **88.46**
		Softmax-based / Energy-based / Feature Norm-based		
CIFAR-100	CE	87.75 / 89.84 / 95.54	71.01 / 71.94 / 59.54	55.42 / 56.69 / 43.21
	CE+EFL	**81.48** / **81.41** / **87.83**	**77.02** / **78.03** / **73.91**	**62.92** / **63.66** / **58.81**
	UL	80.43 / 76.86 / 76.23	66.97 / 74.83 / 76.19	37.83 / 45.07 / 47.51
	UL+EFL	**75.74** / **74.64** / **73.59**	**75.21** / **75.36** / **76.80**	**45.16** / **45.27** / **48.34**

up the learning rate $\eta(t)$ with the feature norm $\|\boldsymbol{h}_{i,c}\|_2$ for each example[15]. Though we change the learning rates, we still guarantee the convergence analysis in Theorem 7 if $\eta(t)$ is non-increasing and satisfies $\frac{\eta(0)(1+\gamma)}{CN} \leq \frac{1}{\|\boldsymbol{w}_c\|_2}$.

Experimental Results To confirm the efficacy of the rescaled learning rate in accelerating convergence for the spherical case, we experiment with the unhinged loss (FNPAL). As illustrated in Figure 2, the utilization of the unhinged loss (FNPAL) with the rescaled learning rate (RLR) clearly demonstrates accelerated convergence in comparison to its counterpart without RLR.

5. Conclusion

In this paper, we introduced the unhinged loss as a surrogate to analyze the behavior of last-layer features and prototypes. Due to the conciseness of the unhinged loss, we derived exact dynamics under gradient descent in various scenarios, including unconstrained, regularized, and spherical constrained cases, as well as the case with an invariant neural tangent kernel. Furthermore, we demonstrated that these dynamics converge exponentially to a specific solution depending on the initialization. Inspired by these results, we proposed additional insights for improvements, such as using prototype-anchored learning with the unhinged loss to bridge its performance to that of cross-entropy loss, employing explicit feature regularization to address over-confidence, and implementing a rescaled learning rate to accelerate convergence in the spherical case. Finally, we validated these theoretical results and insights through extensive experiments, covering numerical analysis,

15. We can also implement this strategy by rescaling the loss, but the multiplicative feature norm stops gradients.

visual classification, imbalanced learning, and out-of-distribution detection. We anticipate that the unhinged loss will serve as a valuable tool for the community to gain a deeper understanding of the behavior of deep neural networks, extending beyond the scope of this paper.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China under Grants 92270116 and 62071155, and in part by the Fundamental Research Funds for the Central Universities (Grant No.HIT.OCEF.2023043 and HIT.DZJJ.2023075). Xiong Zhou thanks the support from MEGVII Technology and Qiyuan Lab.

Appendix for "On the Dynamics under the Unhinged Loss and Beyond"

Appendix A. Clarification for the Unhinged Loss

In this section, we provide some clarification for the unhinged loss as follows.

- **About the contribution.** As stated in the body of this paper, the unhinged loss serves as an alternative object of study, which has the advantage of being linear in the outputs of the model. Thus, it can provide more opportunities to analyze more concrete behaviors and explore the shortcomings of the existing modeling. The main contribution of this work is to glimpse into deep learning through a concise form under the unhinged loss, which enjoys the great advantage to simplify theoretical research very much. Moreover, it is not just a tool for theoretical analysis. To make the unhinged loss practical, we further provide several insights theoretically and empirically, including using prototype-anchored learning with the unhinged loss (Sec. 4.1), applying explicit regularization on features (Sec. 4.2), rescaling the learning rates for the spherical constrained case (Sec. 4.3), and other potential insights (Sec. D).

- **About the criterion of correct classification.** The value of the unhinged loss in Equation (1) cannot be directly used as the criterion to judge whether a sample is classified correctly since the unhinged loss can be arbitrary negative in unconstrained cases, since the gradient of the unhinged loss if $\frac{\partial L_\gamma}{\partial \boldsymbol{h}} = -(\boldsymbol{w}_y - \gamma \sum_{i \neq y} \boldsymbol{w}_i)$ cannot converge to zero unless $\boldsymbol{w}_y - \gamma \sum_{i \neq y} \boldsymbol{w}_i = 0$ for all y. This is also the main reason why the unhinged loss cannot be used directly for practical training. To overcome this drawback, we turn to limit the rapid growth of norms. For instance, in the spherical constrained case, the features are constrained to the unit sphere, and class prototypes are fixed as a simplex equiangular tight frame (ETF). Moreover, under this case, the unhinged loss would be lower bounded by $-(1+\gamma)$, and thus can be used in practical training (as shown in Figure 2). We can also derive that $L_\gamma < -(1+\gamma)\sqrt{\frac{C-2}{2(C-1}} \Rightarrow \boldsymbol{w}_y^\top \boldsymbol{h} > \max_{i \neq y} \boldsymbol{w}_i^\top \boldsymbol{h}$ when these prototypes are fixed as a simplex ETF, so the sample can be correctly classified if the value of the unhinged loss is less than $(1+\gamma)\sqrt{\frac{C-2}{2(C-1}}$.

- **About other losses.**

  1) For the CE and multi-binary CE loss, we have the following inequality.

$$
\begin{aligned}
& L_\gamma(\boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}, y) \\
\leq & \min \Bigg\{ \log(1 + \exp(-\boldsymbol{w}_y^\top \boldsymbol{h} - b_y)) + \gamma \sum_{j \neq y} \log(1 + \exp(\boldsymbol{w}_j^\top \boldsymbol{h} + b_j)), \\
& (\gamma(C-1) - 1)(\boldsymbol{w}_y^\top + b_y)\boldsymbol{h} - \gamma(C-1) \log \frac{\exp(\boldsymbol{w}_y^\top \boldsymbol{h} + b_y)}{\sum_{i=1}^C \exp(\boldsymbol{w}_i^\top \boldsymbol{h} + b_i)} \Bigg\}.
\end{aligned}
$$

  2) For the Wasserstein loss in Arjovsky et al. (2017), we can find that the empirical form of the unhinged loss $\frac{1}{N} \sum_{i=1}^N (\boldsymbol{w}_y^\top \boldsymbol{h} + b_y - \gamma \sum_{i \neq y} (\boldsymbol{w}_i^\top \boldsymbol{h} + b_i))$ resembles the

Wasserstein loss $\mathbb{E}_{x\sim P_r}[f_w(x)] - \mathbb{E}_{x\sim P_\theta}[f_w(x)]$ if we regard positive logits $\boldsymbol{w}_y^\top \boldsymbol{h} + b_y$ as scores from P_r and negative logits $\boldsymbol{w}_i^\top \boldsymbol{h} + b_i$ $(i \neq y)$ as scores from P_θ. The main difference is that in the unhinged loss the two logits $\boldsymbol{w}_y^\top \boldsymbol{h} + b_y$ and $\boldsymbol{w}_i^\top \boldsymbol{h} + b_i$ $(i \neq y)$ are computed from $\boldsymbol{h}$ (*i.e.*, the same input), rather than being sampled separably from two distributions as done in Wasserstein loss. Moreover, the unhinged loss does not constrain the values of these scores, in contrast the Wasserstein loss requires f_w to be Lipschitz continuous.

Appendix B. Proofs for Lemmas, Theorems, Propositions and Corollaries

B.1 Proof of Lemma 1

Lemma 1 (Neural Collapse of The Unhinged Loss). *For norm-bounded prototypes and features, i.e., $\|\boldsymbol{w}_c\|_2 \leq E_1$ and $\|\boldsymbol{h}_{i,c}\|_2 \leq E_2$, $\forall i \in [N], \forall c \in [C]$, the global minimizer of $\frac{1}{CN}\sum_{i=1}^{N}\sum_{c=1}^{C} L_\gamma(\boldsymbol{W}\boldsymbol{h}_{i,c}, y_{i,c})$ implies neural collapse when $p \geq C-1$. More specifically, the global minimizer is uniquely obtained at $\frac{\boldsymbol{w}_i^\top \boldsymbol{w}_j}{\|\boldsymbol{w}_i\|_2\|\boldsymbol{w}_j\|_2} = -\frac{1}{C-1}$, $\forall i \neq j$, $\frac{\boldsymbol{w}_{y_{i,c}}^\top \boldsymbol{h}_{i,c}}{\|\boldsymbol{w}_{y_{i,c}}\|_2\|\boldsymbol{h}_{i,c}\|_2} = 1$, $\|\boldsymbol{w}_c\|_2 = E_1$, and $\|\boldsymbol{h}_{i,c}\|_2 = E_2$, $\forall i \in [N]$, $\forall c \in [C]$.*

Proof The proof is based on lower bounding the objective $\frac{1}{CN}\sum_{i=1}^{N}\sum_{c=1}^{C} L_\gamma(\boldsymbol{W}\boldsymbol{h}_{i,c}, y_{i,c})$ by a sequence of inequalities that holds if and only if the solution forms Neural Collapse (Papyan et al., 2020). Let $\hat{\boldsymbol{w}} = \frac{1}{C}\sum_{c=1}^{C}\boldsymbol{w}_c$, according to the definition of L_{SM}, we have

$$
\begin{aligned}
&\frac{1}{CN}\sum_{i=1}^{N}\sum_{c=1}^{C} L_{SM}(\boldsymbol{W}\boldsymbol{h}_{i,c}, y_{i,c}) \\
=&\frac{1}{CN}\sum_{i=1}^{N}\sum_{c=1}^{C}(-\boldsymbol{w}_{y_{i,c}}^\top \boldsymbol{h}_{i,c} + \gamma \sum_{j\neq y_{i,c}} \boldsymbol{w}_j^\top \boldsymbol{h}_{i,c}) \\
\geq&\frac{1}{CN}\sum_{i=1}^{N}\sum_{c=1}^{C}(-E_1E_2 + \gamma(C\hat{\boldsymbol{w}} - \boldsymbol{w}_{y_{i,c}})^\top \boldsymbol{h}_{i,c}) \\
\geq&-\frac{\gamma E_2}{CN}\sum_{i=1}^{N}\sum_{c=1}^{C}\|C\hat{\boldsymbol{w}} - \boldsymbol{w}_{y_{i,c}}\|_2 - E_1E_2 \\
\geq&-\gamma E_2\sqrt{\frac{1}{CN}\sum_{i=1}^{N}\sum_{c=1}^{C}\|C\hat{\boldsymbol{w}} - \boldsymbol{w}_{y_{i,c}}\|_2^2} - E_1E_2 \\
=&-\gamma E_2\sqrt{\frac{1}{C}\sum_{c=1}^{C}\|\boldsymbol{w}_c\|_2^2 - C^2\|\hat{\boldsymbol{w}}\|_2^2} - E_1E_2 \\
\geq&-(1+\gamma)E_1E_2
\end{aligned}
,
$$

where the first and second inequalities are based on the facts that $\boldsymbol{w}_{y_{i,c}}^\top \boldsymbol{h}_{i,c} \leq E_1E_2$ and $(C\hat{\boldsymbol{w}} - \boldsymbol{w}_{y_{i,c}})^\top \boldsymbol{h}_{i,c} \geq -E_2\|C\hat{\boldsymbol{w}} - \boldsymbol{w}_{y_{i,c}}\|_2$, respectively. In the third equality, we used the Cauchy-Schwarz inequality, and the last inequality we use the facts that $\|\boldsymbol{w}_c\|_2 \leq E_1$ and $\|\hat{\boldsymbol{w}}\|_2 \geq 0$.

According the above derivation, the equality holds if and only if $\forall i \in [N]$, $\forall c \in [C]$, $\boldsymbol{w}_{y_{i,c}}^\top \boldsymbol{h}_{i,c} = E_1 E_2$, $(C\hat{\boldsymbol{w}} - \boldsymbol{w}_{y_{i,c}})^\top \boldsymbol{h}_{i,c} = -E_2 \|C\hat{\boldsymbol{w}} - \boldsymbol{w}_{y_{i,c}}\|_2$, $\|C\hat{\boldsymbol{w}} - \boldsymbol{w}_c\|_2 = \|C\hat{\boldsymbol{w}} - \boldsymbol{w}_C\|_2$, $\|\boldsymbol{w}_c\|_2 = E_1$, and $\|\hat{\boldsymbol{w}}\|_2 = 0$. These equations can be simplified as $\frac{\boldsymbol{w}_i^\top \boldsymbol{w}_j}{\|\boldsymbol{w}_i\|_2 \|\boldsymbol{w}_j\|_2} = -\frac{1}{C-1}$, $\forall i \neq j$, $\frac{\boldsymbol{w}_{y_{i,c}}^\top \boldsymbol{h}_{i,c}}{\|\boldsymbol{w}_{y_{i,c}}\|_2 \|\boldsymbol{h}_{i,c}\|_2} = 1$, $\|\boldsymbol{w}_c\|_2 = E_1$, and $\|\boldsymbol{h}_{i,c}\|_2 = E_2$, $\forall i \in [N]$, $\forall c \in [C]$, which also implies neural collapse. ■

B.2 Proof of Theorem 2

In this section, we will provide the proof of Theorem 2. Our analysis will actually rely on the eigenvalues and eigenspaces of five subspaces $\mathcal{E}_1^+$, $\mathcal{E}_1^-$, $\mathcal{E}_2^+$, $\mathcal{E}_2^-$ and $\mathcal{E}_3$ in Theorem 2. Their concrete projection operator can be found in Appendix B.9. In the following, we show that these five subspaces are orthogonal:

Lemma 9. *The following five subspaces are orthogonal to each other and satisfy* $\mathbb{R}^{p\times C(N+1)} = \mathcal{E}_1^+ \oplus \mathcal{E}_1^- \oplus \mathcal{E}_2^+ \oplus \mathcal{E}_2^- \oplus \mathcal{E}_3$*:*

$$\begin{aligned}
\mathcal{E}_1^\epsilon &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = \epsilon \cdot \tfrac{1}{\sqrt{N}}(\boldsymbol{W} \otimes \boldsymbol{1}_N^\top), \boldsymbol{W}\boldsymbol{1}_C = 0, \boldsymbol{W} \in \mathbb{R}^{p\times C}\},\\
\mathcal{E}_2^\epsilon &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = \epsilon \cdot \tfrac{1}{\sqrt{N}}\boldsymbol{h}\boldsymbol{1}_{CN}^\top, \boldsymbol{W} = \boldsymbol{h}\boldsymbol{1}_C^\top, \boldsymbol{h} \in \mathbb{R}^p\},\\
\mathcal{E}_3 &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H}(\boldsymbol{I}_C \otimes \boldsymbol{1}_N) = 0, \boldsymbol{W} = 0, \boldsymbol{H} \in \mathbb{R}^{p\times CN}\}.
\end{aligned} \tag{26}$$

where $\epsilon \in \{\pm 1\}$*, and* $k \neq 0$*.*

Proof *For* $(\boldsymbol{H}_1, \boldsymbol{W}_1) = (\frac{1}{\sqrt{N}}(\boldsymbol{W}_1 \otimes \boldsymbol{1}_N^\top), \boldsymbol{W}_1) \in \mathcal{E}_1^+$ *and* $(\boldsymbol{H}_2, \boldsymbol{W}_2) = (-\frac{1}{\sqrt{N}}(\boldsymbol{W}_2 \otimes \boldsymbol{1}_N^\top), \boldsymbol{W}_2) \in \mathcal{E}_1^-$*, we have*

$$\boldsymbol{H}_1\boldsymbol{H}_2^\top + \boldsymbol{W}_1\boldsymbol{W}_2^\top = -\frac{1}{N}(\boldsymbol{W}_1 \otimes \boldsymbol{1}_N^\top)(\boldsymbol{W}_2^\top \otimes \boldsymbol{1}_N) + \boldsymbol{W}_1\boldsymbol{W}_2^\top = 0.$$

For $(\boldsymbol{H}_1, \boldsymbol{W}_1) = (\frac{1}{\sqrt{N}}\boldsymbol{h}_1\boldsymbol{1}_{CN}^\top, \boldsymbol{h}_1\boldsymbol{1}_C^\top) \in \mathcal{E}_2^+$ *and* $(\boldsymbol{H}_2, \boldsymbol{W}_2) = (-\frac{1}{\sqrt{N}}\boldsymbol{h}_2\boldsymbol{1}_{CN}^\top, \boldsymbol{h}_2\boldsymbol{1}_C^\top) \in \mathcal{E}_2^-$*, we have*

$$\boldsymbol{H}_1\boldsymbol{H}_2^\top + \boldsymbol{W}_1\boldsymbol{W}_2^\top = -\frac{1}{N}\boldsymbol{h}_1\boldsymbol{1}_{CN}^\top\boldsymbol{1}_{CN}\boldsymbol{h}_2^\top + \boldsymbol{h}_1\boldsymbol{1}_C^\top\boldsymbol{1}_C\boldsymbol{h}_2^\top = 0.$$

For $(\boldsymbol{H}_1, \boldsymbol{W}_1) = (\epsilon_1\frac{1}{\sqrt{N}}(\boldsymbol{W}_1 \otimes \boldsymbol{1}_N^\top), \boldsymbol{W}_1) \in \mathcal{E}_1^\epsilon$ *and* $(\boldsymbol{H}_2, \boldsymbol{W}_2) = (\epsilon_2\frac{1}{\sqrt{N}}\boldsymbol{h}_2\boldsymbol{1}_{CN}^\top, \boldsymbol{h}_2\boldsymbol{1}_C^\top) \in \mathcal{E}_2^\epsilon$*, since* $\boldsymbol{W}_1\boldsymbol{1}_C = 0$ *and* $(\boldsymbol{W}_1 \otimes \boldsymbol{1}_N^\top)\boldsymbol{1}_{CN} = N\boldsymbol{W}_1\boldsymbol{1}_C = 0$*, we have*

$$\boldsymbol{H}_1\boldsymbol{H}_2^\top + \boldsymbol{W}_1\boldsymbol{W}_2^\top = \frac{\epsilon_1\epsilon_2}{N}(\boldsymbol{W}_1 \otimes \boldsymbol{1}_N^\top)\boldsymbol{1}_{CN}\boldsymbol{h}_2^\top + \boldsymbol{W}_1\boldsymbol{1}_C\boldsymbol{h}_2^\top = 0.$$

For $(\boldsymbol{H}_1, \boldsymbol{W}_1) = (\epsilon\frac{1}{\sqrt{N}}(\boldsymbol{W}_1 \otimes \boldsymbol{1}_N^\top), \boldsymbol{W}_1) \in \mathcal{E}_1^\epsilon$ *and* $(\boldsymbol{H}_2, \boldsymbol{W}_2) = (\boldsymbol{H}_2, 0) \in \mathcal{E}_3$*, we have*

$$\boldsymbol{H}_1\boldsymbol{H}_2^\top + \boldsymbol{W}_1\boldsymbol{W}_2^\top = \frac{\epsilon}{\sqrt{N}}(\boldsymbol{W}_1 \otimes \boldsymbol{1}_N^\top)\boldsymbol{H}_2^\top = \frac{\epsilon}{\sqrt{N}}\boldsymbol{W}_1(\boldsymbol{H}_2(\boldsymbol{I}_C \otimes \boldsymbol{1}_N))^\top = 0.$$

For $(\boldsymbol{H}_1, \boldsymbol{W}_1) = (\epsilon\frac{1}{\sqrt{N}}\boldsymbol{h}_1\boldsymbol{1}_{CN}^\top, \boldsymbol{h}_1\boldsymbol{1}_C^\top) \in \mathcal{E}_2^\epsilon$ *and* $(\boldsymbol{H}_2, \boldsymbol{W}_2) = (\boldsymbol{H}_2, 0) \in \mathcal{E}_3$*, since* $\boldsymbol{H}_2(\boldsymbol{I}_C \otimes \boldsymbol{1}_N) = 0$*, we have*

$$\boldsymbol{H}_1\boldsymbol{H}_2^\top + \boldsymbol{W}_1\boldsymbol{W}_2^\top = \frac{\epsilon}{\sqrt{N}}\boldsymbol{h}_1\boldsymbol{1}_{CN}^\top\boldsymbol{H}_2^\top = \frac{\epsilon}{\sqrt{N}}\boldsymbol{h}_1(\boldsymbol{H}_2(\boldsymbol{I}_C \otimes \boldsymbol{1}_N)\boldsymbol{1}_C^\top)^\top = 0$$

To sum up, we prove that the five subspaces $\mathcal{E}_1^+, \mathcal{E}_1^-, \mathcal{E}_2^+, \mathcal{E}_2^-, \mathcal{E}_3$ *are orthogonal to each other. Moreover, we have*

$$dim\,\mathcal{E}_1^+ = dim\,\mathcal{E}_1^- = p(C-1), \quad dim\,\mathcal{E}_2^+ = dim\,\mathcal{E}_2^- = p, \quad dim\,\mathcal{E}_3 = pC(N-1).$$

Since these dimensions sum to $pC(N+1) = dim(\mathbb{R}^{p\times CN} \oplus \mathbb{R}^{p\times C})$*, then* $\mathbb{R}^{p\times C(N+1)} = \mathcal{E}_1^+ \oplus \mathcal{E}_1^- \oplus \mathcal{E}_2^+ \oplus \mathcal{E}_2^- \oplus \mathcal{E}_3$. ■

Theorem 2 (Dynamics of Features, Class Prototypes and Biases without Constraints). *Consider the continual gradient flow in Equation 5, let* $\boldsymbol{Z}(t) = (\boldsymbol{H}(t), \boldsymbol{W}(t))$*, if* $\eta_1(t_1)\eta_2(t_2) = \eta_1(t_2)\eta_2(t_1)$ *for any* $t_1, t_2 \geq 0$*, we have the following closed-form dynamics*

$$\begin{aligned}\boldsymbol{Z}(t) =& \Pi_1^+ \boldsymbol{Z}_0 \left(\alpha_1^+(t)\boldsymbol{C}(t) + \beta_1^+(t)\boldsymbol{I}_{C(N+1)}\right) \\ &+\Pi_1^- \boldsymbol{Z}_0 \left(\alpha_1^-(t)\boldsymbol{C}(t) + \beta_1^-(t)\boldsymbol{I}_{C(N+1)}\right) \\ &+\Pi_2^+ \boldsymbol{Z}_0 \left(\alpha_2^+(t)\boldsymbol{C}(t) + \beta_2^+(t)\boldsymbol{I}_{C(N+1)}\right) \\ &+\Pi_2^- \boldsymbol{Z}_0 \left(\alpha_2^-(t)\boldsymbol{C}(t) + \beta_2^-(t)\boldsymbol{I}_{C(N+1)}\right) + \Pi_3 \boldsymbol{Z}_0,\end{aligned} \tag{27}$$

and

$$\boldsymbol{b}(t) = \boldsymbol{b}_0 + \frac{(1+\gamma-\gamma C)\zeta_2(t)}{C}\mathbf{1}_C, \tag{28}$$

where α_1^ϵ*,* α_2^ϵ*,* β_1^ϵ *and* β_2^ϵ *for* $\epsilon \in \{\pm\}$ *are the scalars that only depend on* C*,* N*,* γ*,* η_1 *and* η_2 *(where the detailed forms of these scalars can be seen in the appendix),* $\boldsymbol{Z}_0 = (\boldsymbol{H}_0, \boldsymbol{W}_0)$*,* $\boldsymbol{C}(t) = \begin{pmatrix} \zeta_1(t)\boldsymbol{I}_{CN} & 0 \\ 0 & \zeta_2(t)\boldsymbol{I}_C \end{pmatrix}$*,* $\zeta_1(t) = \int_0^t \eta_1(\tau)\mathrm{d}\tau$*,* $\zeta_2(t) = \int_0^t \eta_2(\tau)\mathrm{d}\tau$*,* Π_1^ϵ*,* Π_2^ϵ *and* Π_3 *for* $\epsilon \in \{\pm\}$ *are orthogonal projection operators onto the following respective eigenspaces:*

$$\begin{aligned}\mathcal{E}_1^\epsilon &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = \epsilon \cdot \tfrac{1}{\sqrt{N}}(\boldsymbol{W} \otimes \mathbf{1}_N^\top), \boldsymbol{W}\mathbf{1}_C = 0\}, \\ \mathcal{E}_2^\epsilon &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = \epsilon \cdot \tfrac{1}{\sqrt{N}}\boldsymbol{h}\mathbf{1}_{CN}^\top, \boldsymbol{W} = \boldsymbol{h}\mathbf{1}_C^\top, \boldsymbol{h} \in \mathbb{R}^p\}, \\ \mathcal{E}_3 &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H}(\boldsymbol{I}_C \otimes \mathbf{1}_N) = 0, \boldsymbol{W} = 0\}.\end{aligned} \tag{29}$$

Proof Writing $\boldsymbol{Z}(t) = (\boldsymbol{H}(t), \boldsymbol{W}(t))$, then the unsolved portion of the system is given by

$$\boldsymbol{Z}'(t) = \boldsymbol{Z}(t)\begin{pmatrix} 0 & \boldsymbol{M}^\top \\ \boldsymbol{M} & 0 \end{pmatrix}\begin{pmatrix} \eta_1(t)\boldsymbol{I}_{CN} & 0 \\ 0 & \eta_2(t)\boldsymbol{I}_C \end{pmatrix}, \tag{30}$$

where $\boldsymbol{M} = \frac{1}{CN}((1+\gamma)(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top) - \gamma\mathbf{1}_C\mathbf{1}_{CN}^\top)$.

Let $\boldsymbol{A}(t) = \begin{pmatrix} 0 & \boldsymbol{M}^\top \\ \boldsymbol{M} & 0 \end{pmatrix}\begin{pmatrix} \eta_1(t)\boldsymbol{I}_{CN} & 0 \\ 0 & \eta_2(t)\boldsymbol{I}_C \end{pmatrix} = \begin{pmatrix} 0 & \eta_2(t)\boldsymbol{M}^\top \\ \eta_1(t)\boldsymbol{M} & 0 \end{pmatrix}$, then the equation above can be reformulated as the initial-value problem associated with the linear ordinary differential equation:

$$\boldsymbol{Z}'(t) = \boldsymbol{Z}(t)\boldsymbol{A}(t), \quad \boldsymbol{Z}(0) = \boldsymbol{Z}_0. \tag{31}$$

For any t_1, t_2, we have the matrix commutator of $\boldsymbol{A}(t_1)$ and $\boldsymbol{A}(t_2)$

$$\begin{aligned}
&[\boldsymbol{A}(t_1), \boldsymbol{A}(t_2)] \\
=&\boldsymbol{A}(t_1)\boldsymbol{A}(t_2) - \boldsymbol{A}(t_2)\boldsymbol{A}(t_1) \\
=&\begin{pmatrix} 0 & \eta_2(t_1)\boldsymbol{M}^\top \\ \eta_1(t_1)\boldsymbol{M} & 0 \end{pmatrix}\begin{pmatrix} 0 & \eta_2(t_2)\boldsymbol{M}^\top \\ \eta_1(t_2)\boldsymbol{M} & 0 \end{pmatrix} - \boldsymbol{A}(t_2)\boldsymbol{A}(t_1) \\
=&\begin{pmatrix} \eta_2(t_1)\eta_1(t_2)\boldsymbol{M}^\top\boldsymbol{M} & 0 \\ 0 & \eta_1(t_1)\eta_2(t_2)\boldsymbol{M}\boldsymbol{M}^\top \end{pmatrix} - \boldsymbol{A}(t_2)\boldsymbol{A}(t_1) \\
=&\begin{pmatrix} (\eta_2(t_1)\eta_1(t_2) - \eta_2(t_2)\eta_1(t_1))\boldsymbol{M}^\top\boldsymbol{M} & 0 \\ 0 & (\eta_1(t_1)\eta_2(t_2) - \eta_2(t_1)\eta_1(t_2))\boldsymbol{M}\boldsymbol{M}^\top \end{pmatrix} \\
=&0
\end{aligned}$$

where the last equality is based on the fact that $\eta_2(t_1)\eta_1(t_2) = \eta_2(t_2)\eta_1(t_1)$. Therefore, according to Magnus approach, we have

$$\boldsymbol{Z}(t) = \boldsymbol{Z}_0 \exp\left(\int_0^t \boldsymbol{A}(\tau)\mathrm{d}\tau\right) = \boldsymbol{Z}_0 \exp\begin{pmatrix} 0 & \int_0^t \eta_2(\tau)\mathrm{d}\tau\boldsymbol{M}^\top \\ \int_0^t \eta_1(\tau)\mathrm{d}\tau\boldsymbol{M} & 0 \end{pmatrix}. \tag{32}$$

Let $\zeta_1(t) = \int_0^t \eta_1(\tau)\mathrm{d}\tau$, $\zeta_2(t) = \int_0^t \eta_2(\tau)\mathrm{d}\tau$, $\boldsymbol{B} = \begin{pmatrix} 0 & \boldsymbol{M}^\top \\ \boldsymbol{M} & 0 \end{pmatrix}$, $\boldsymbol{C}(t) = \begin{pmatrix} \zeta_1(t)\boldsymbol{I}_{CN} & 0 \\ 0 & \zeta_2(t)\boldsymbol{I}_C \end{pmatrix}$, and $\boldsymbol{L}(t) = \boldsymbol{B}\boldsymbol{C}(t)$, we have

$$\boldsymbol{Z}(t) = \boldsymbol{Z}_0 \exp(\boldsymbol{L}(t)) = \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\boldsymbol{L}(t))^k}{k!}. \tag{33}$$

Moreover, we have

$$\begin{aligned}
(\boldsymbol{L}(t))^2 &= \begin{pmatrix} 0 & \zeta_2(t)\boldsymbol{M}^\top \\ \zeta_1(t)\boldsymbol{M} & 0 \end{pmatrix}\begin{pmatrix} 0 & \zeta_2(t)\boldsymbol{M}^\top \\ \zeta_1(t)\boldsymbol{M} & 0 \end{pmatrix} \\
&= \begin{pmatrix} \zeta_1(t)\zeta_2(t)\boldsymbol{M}^\top\boldsymbol{M} & 0 \\ 0 & \zeta_1(t)\zeta_2(t)\boldsymbol{M}\boldsymbol{M}^\top \end{pmatrix} \\
&= \zeta_1(t)\zeta_2(t)\boldsymbol{B}^2,
\end{aligned} \tag{34}$$

thus we obtain

$$\begin{aligned}
\boldsymbol{Z}(t) =&\boldsymbol{Z}_0 \left(\sum_{k=0}^{\infty} \frac{(\boldsymbol{L}(t))^{2k+1}}{(2k+1)!} + \sum_{k=0}^{\infty} \frac{(\boldsymbol{L}(t))^{2k}}{(2k)!}\right) \\
=&\boldsymbol{Z}_0 \left(\sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k+1}\boldsymbol{C}(t)}{(2k+1)!} + \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k}}{(2k)!}\right) \\
=&\boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k+1}\boldsymbol{C}(t)}{(2k+1)!} + \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k}}{(2k)!}
\end{aligned} \tag{35}$$

Looking at the above equation, we just need to analyze the eigenspaces and eigenvalues of $\boldsymbol{B}$.

Considering the following five subspaces:

$$\begin{aligned}
\mathcal{E}_1^\epsilon &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = \epsilon \cdot \tfrac{1}{\sqrt{N}}(\boldsymbol{W} \otimes \mathbf{1}_N^\top), \boldsymbol{W}\mathbf{1}_C = 0\}, \\
\mathcal{E}_2^\epsilon &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = \epsilon \cdot \tfrac{1}{\sqrt{N}}\boldsymbol{h}\mathbf{1}_{CN}^\top, \boldsymbol{W} = \boldsymbol{h}\mathbf{1}_C^\top, \boldsymbol{h} \in \mathbb{R}^p\}, \\
\mathcal{E}_3 &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H}(\boldsymbol{I}_C \otimes \mathbf{1}_N) = 0, \boldsymbol{W} = 0\}.
\end{aligned} \tag{36}$$

where $\epsilon \in \{\pm\}$. According to Lemma 9, these five subspaces are orthogonal to each other and satisfy $\mathbb{R}^{p\times C(N+1)} = \mathcal{E}_1^+ \oplus \mathcal{E}_1^- \oplus \mathcal{E}_2^+ \oplus \mathcal{E}_2^- \oplus \mathcal{E}_3$.

In the following, we will prove that $\mathcal{E}_1^\epsilon$, $\mathcal{E}_2^\epsilon$ and $\mathcal{E}_3$ are five eigenspaces of $\boldsymbol{B}$. More specifically, each nonzero member of each claimed eigenspace is an eigenvector, and the claimed eigenspaces have distinct eigenvalues.

Note that for $(\boldsymbol{H}, \boldsymbol{W}) \in \mathbb{R}^{p\times CN} \oplus \mathbb{R}^{p\times C}$, we have $(\boldsymbol{H}, \boldsymbol{W})\boldsymbol{B} = (\boldsymbol{W}\boldsymbol{M}^\top, \boldsymbol{H}\boldsymbol{W})$.

For $(\boldsymbol{H}, \boldsymbol{W}) \in \mathcal{E}_1^\epsilon$, we have $\boldsymbol{H} = \frac{\epsilon}{\sqrt{N}}\boldsymbol{W} \otimes \mathbf{1}_N^\top$ and $\boldsymbol{W}\mathbf{1}_C = 0$, thus

$$\begin{aligned}
\boldsymbol{W}\boldsymbol{M} &= \tfrac{1}{CN}\boldsymbol{W}((1+\gamma)(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top) - \gamma\mathbf{1}_C\mathbf{1}_{CN}^\top) \\
&= \tfrac{(1+\gamma)}{CN}\boldsymbol{W} \otimes \mathbf{1}_N^\top \\
&= \tfrac{\epsilon(1+\gamma)}{C\sqrt{N}}\boldsymbol{H}, \\
\boldsymbol{H}\boldsymbol{M}^\top &= \tfrac{1}{CN}[\epsilon \cdot \tfrac{1}{\sqrt{N}}(\boldsymbol{W} \otimes \mathbf{1}_N^\top)][(1+\gamma)(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top) - \gamma\mathbf{1}_C\mathbf{1}_{CN}^\top]^\top \\
&= \tfrac{\epsilon}{CN\sqrt{N}}[(\boldsymbol{W} \otimes \mathbf{1}_N^\top)((1+\gamma)(\boldsymbol{I}_C \otimes \mathbf{1}_N) - \gamma(\mathbf{1}_C \otimes \mathbf{1}_N)\mathbf{1}_C^\top)] \\
&= \tfrac{\epsilon(1+\gamma)}{C\sqrt{N}}\boldsymbol{W},
\end{aligned}$$

i.e., $(\boldsymbol{H}, \boldsymbol{W})$ is an eigenvector of $\boldsymbol{B}$ with eigenvalue $\frac{\epsilon(1+\gamma)}{C\sqrt{N}}$.

For $(\boldsymbol{H}, \boldsymbol{W}) \in \mathcal{E}_2^\epsilon$, we have $\boldsymbol{H} = \frac{\epsilon}{\sqrt{N}}\boldsymbol{h}\mathbf{1}_{CN}^\top$ and $\boldsymbol{W} = \boldsymbol{h}\mathbf{1}_C^\top$, thus

$$\begin{aligned}
\boldsymbol{W}\boldsymbol{M} &= \tfrac{1}{CN}\boldsymbol{h}\mathbf{1}_C^\top((1+\gamma)(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top) - \gamma\mathbf{1}_C\mathbf{1}_{CN}^\top) \\
&= \tfrac{1}{CN}\boldsymbol{h}((1+\gamma)\mathbf{1}_C(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top) - \gamma C\mathbf{1}_{CN}^\top) \\
&= \tfrac{(1+\gamma-\gamma C)}{CN}\boldsymbol{h}\mathbf{1}_{CN}^\top \\
&= \tfrac{\epsilon(1+\gamma-\gamma C)}{C\sqrt{N}}\boldsymbol{H}, \\
\boldsymbol{H}\boldsymbol{M}^\top &= \tfrac{1}{CN\sqrt{N}}(\epsilon \cdot \boldsymbol{h}\mathbf{1}_{CN}^\top)((1+\gamma)(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top) - \gamma\mathbf{1}_C\mathbf{1}_{CN}^\top)^\top \\
&= \tfrac{\epsilon}{CN}\boldsymbol{h}((1+\gamma)\mathbf{1}_{CN}^\top(\boldsymbol{I}_C \otimes \mathbf{1}_N) - \gamma\mathbf{1}_{CN\sqrt{N}}^\top\mathbf{1}_{CN}\mathbf{1}_C^\top) \\
&= \tfrac{\epsilon}{CN\sqrt{N}}\boldsymbol{h}((1+\gamma)N\mathbf{1}_C^\top - \gamma CN\mathbf{1}_C^\top) \\
&= \tfrac{\epsilon(1+\gamma-\gamma C)}{C\sqrt{N}}\boldsymbol{h}\mathbf{1}_C^\top \\
&= \tfrac{\epsilon(1+\gamma-\gamma C)}{C\sqrt{N}}\boldsymbol{W},
\end{aligned}$$

i.e., $(\boldsymbol{H}, \boldsymbol{W})$ is an eigenvector of $\boldsymbol{B}$ with eigenvalue $\frac{\epsilon(1+\gamma-\gamma C)}{C\sqrt{N}}$.

For $(\boldsymbol{H}, \boldsymbol{W}) \in \mathcal{E}_3$, we have $\boldsymbol{H}(\boldsymbol{I}_C \otimes \mathbf{1}_N) = 0$ and $\boldsymbol{W} = 0$, thus

$$\begin{aligned}
\boldsymbol{W}\boldsymbol{M} &= \tfrac{1}{CN} \cdot 0((1+\gamma)(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top) - \gamma \mathbf{1}_C \mathbf{1}_{CN}^\top) = 0, \\
\boldsymbol{H}\boldsymbol{M}^\top &= \tfrac{1}{CN}\boldsymbol{H}((1+\gamma)(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top) - \gamma \mathbf{1}_C \mathbf{1}_{CN}^\top)^\top \\
&= \tfrac{1}{CN}\boldsymbol{H}((1+\gamma)(\boldsymbol{I}_C \otimes \mathbf{1}_N) - \gamma \mathbf{1}_{CN} \mathbf{1}_C^\top) \\
&= -\tfrac{\gamma}{CN}\boldsymbol{H}(\boldsymbol{I}_C \otimes \mathbf{1}_N)\mathbf{1}_C \mathbf{1}_C^\top \\
&= 0
\end{aligned}$$

i.e., $(\boldsymbol{H}, \boldsymbol{W})$ is an eigenvector of $\boldsymbol{B}$ with eigenvalue 0.

Overall, letting Π_i^ϵ denote orthogonal projection onto $\mathcal{E}_i^\epsilon$, we have the spectral decomposition

$$\boldsymbol{B} = \tfrac{1}{C\sqrt{N}}\left[(1+\gamma)(\Pi_1^+ - \Pi_1^-) + (1+\gamma-\gamma C)(\Pi_2^+ - \Pi_2^-)\right]. \tag{37}$$

We then provide the concrete formulation of $\boldsymbol{Z}(t) = \boldsymbol{Z}_0 \exp(\boldsymbol{L}(t))$ by the orthogonal projection of $\boldsymbol{Z}_0$ onto each eigenspace of $\boldsymbol{B}$, *i.e.*,

$$\boldsymbol{Z}_0 = \Pi_1^+ \boldsymbol{Z}_0 + \Pi_1^- \boldsymbol{Z}_0 + \Pi_2^+ \boldsymbol{Z}_0 + \Pi_2^- \boldsymbol{Z}_0 + \Pi_3 \boldsymbol{Z}_0.$$

Decomposition along $\Pi_1^\epsilon \boldsymbol{Z}_0$. First, $\Pi_1^\epsilon \boldsymbol{Z}_0 \boldsymbol{B} = \frac{\epsilon(1+\gamma)}{C\sqrt{N}} \Pi_1^\epsilon \boldsymbol{Z}_0$, so $\Pi_1^\epsilon \boldsymbol{Z}_0 \boldsymbol{B}^k = \left(\frac{\epsilon(1+\gamma)}{C\sqrt{N}}\right)^k \Pi_1^\epsilon \boldsymbol{Z}_0$ for $k \geq 0$, then

$$\begin{aligned}
&\Pi_1^\epsilon \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k+1} \boldsymbol{C}(t)}{(2k+1)!} \\
=&\Pi_1^\epsilon \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \left(\frac{\epsilon(1+\gamma)}{C\sqrt{N}}\right)^{2k+1} \boldsymbol{C}(t)}{(2k+1)!} \\
=&\frac{\Pi_1^\epsilon \boldsymbol{Z}_0 \boldsymbol{C}(t)}{\sqrt{\zeta_1(t)\zeta_2(t)}} \sum_{k=0}^{\infty} \frac{\left(\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right)^{2k+1}}{(2k+1)!} \\
=&\frac{\Pi_1^\epsilon \boldsymbol{Z}_0 \boldsymbol{C}(t)}{2\sqrt{\zeta_1(t)\zeta_2(t)}} \left(e^{\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}} - e^{-\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}}\right),
\end{aligned} \tag{38}$$

and

$$\begin{aligned}
&\Pi_1^\epsilon \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k}}{(2k)!} \\
=&\Pi_1^\epsilon \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \left(\frac{\epsilon(1+\gamma)}{C\sqrt{N}}\right)^{2k}}{(2k)!} \\
=&\frac{\Pi_1^\epsilon \boldsymbol{Z}_0}{2} \left(e^{\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}} + e^{-\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}}\right),
\end{aligned} \tag{39}$$

which is based on the facts that $\frac{e^x - e^{-x}}{2} = \sum_{k=0}^{\infty} \frac{x^{2k+1}}{(2k+1)!}$ and $\frac{e^x + e^{-x}}{2} = \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k)!}$. Thus we have

$$\Pi_1^\epsilon \boldsymbol{Z}_0 \exp(\boldsymbol{L}(t)) = \Pi_1^\epsilon \boldsymbol{Z}_0 (\alpha_1^\epsilon \boldsymbol{C}(t) + \beta_1^\epsilon \boldsymbol{I}_{C(N+1)}), \tag{40}$$

with

$$\begin{aligned}\alpha_1^\epsilon(t) &= \frac{\exp\left(\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right) - \exp\left(-\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right)}{2\sqrt{\zeta_1(t)\zeta_2(t)}},\\ \beta_1^\epsilon(t) &= \frac{\exp\left(\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right) + \exp\left(-\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right)}{2}.\end{aligned} \tag{41}$$

Decomposition along $\Pi_2^\epsilon \boldsymbol{Z}_0$. Similarly, for $\Pi_2^\epsilon \boldsymbol{Z}_0 \boldsymbol{B} = \frac{\epsilon(1+\gamma-\gamma C)}{C\sqrt{N}} \Pi_2^\epsilon$, we have

$$\Pi_2^\epsilon \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k+1} \boldsymbol{C}(t)}{(2k+1)!} = \frac{\Pi_2^\epsilon \boldsymbol{Z}_0 \boldsymbol{C}(t)}{2\sqrt{\zeta_1(t)\zeta_2(t)}} \left(e^{\frac{\epsilon(1+\gamma-\gamma C)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}} - e^{-\frac{\epsilon(1+\gamma-\gamma C)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}} \right),$$

and

$$\Pi_2^\epsilon \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k}}{(2k)!} = \frac{\Pi_2^\epsilon \boldsymbol{Z}_0}{2} \left(e^{\frac{\epsilon(1+\gamma-\gamma C)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}} + e^{-\frac{\epsilon(1+\gamma-\gamma C)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}} \right).$$

Thus we have

$$\Pi_2^\epsilon \boldsymbol{Z}_0 \exp(\boldsymbol{L}(t)) = \Pi_2^\epsilon \boldsymbol{Z}_0 (\alpha_2^\epsilon \boldsymbol{C}(t) + \beta_2^\epsilon \boldsymbol{I}_{C(N+1)}), \tag{42}$$

with

$$\begin{aligned}\alpha_2^\epsilon(t) &= \frac{\exp\left(\frac{\epsilon(1+\gamma-\gamma C)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right) - \exp\left(-\frac{\epsilon(1+\gamma-\gamma C)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right)}{2\sqrt{\zeta_1(t)\zeta_2(t)}},\\ \beta_2^\epsilon(t) &= \frac{\exp\left(\frac{\epsilon(1+\gamma-\gamma C)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right) + \exp\left(-\frac{\epsilon(1+\gamma-\gamma C)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right)}{2}.\end{aligned} \tag{43}$$

Decomposition along $\Pi_3 \boldsymbol{Z}_0$. Since each vector in $\mathcal{E}_3$ is a eigenvector of $\boldsymbol{B}$ with eigenvalue 0, then we have

$$\Pi_3 \boldsymbol{Z}_0 \left(\sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k+1} \boldsymbol{C}(t)}{(2k+1)!} + \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k}}{(2k)!} \right) = \Pi_3 \boldsymbol{Z}_0 \tag{44}$$

Note that $\mathcal{E}_1^+$, $\mathcal{E}_1^-$, $\mathcal{E}_2^+$, $\mathcal{E}_2^-$ and $\mathcal{E}_3$ are orthogonal subspace of $\mathbb{R}^{p\times CN} \oplus \mathbb{R}^{C\times p}$, thus

$$\begin{aligned}\boldsymbol{Z}(t) =& \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k+1} \boldsymbol{C}(t)}{(2k+1)!} + \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k}}{(2k)!}\\ =& (\Pi_1^+ \boldsymbol{Z}_0 + \Pi_1^- \boldsymbol{Z}_0 + \Pi_2^+ \boldsymbol{Z}_0 + \Pi_2^- \boldsymbol{Z}_0 + \Pi_3 \boldsymbol{Z}_0) \exp(\boldsymbol{L}(t))\\ =& \Pi_1^+ \boldsymbol{Z}_0 \left(\alpha_1^+(t)\boldsymbol{C}(t) + \beta_1^+(t)\boldsymbol{I}_{C(N+1)}\right) + \Pi_1^- \boldsymbol{Z}_0 \left(\alpha_1^-(t)\boldsymbol{C}(t) + \beta_1^-(t)\boldsymbol{I}_{C(N+1)}\right) + \Pi_3 \boldsymbol{Z}_0\\ &+ \Pi_2^+ \boldsymbol{Z}_0 \left(\alpha_2^+(t)\boldsymbol{C}(t) + \beta_2^+(t)\boldsymbol{I}_{C(N+1)}\right) + \Pi_2^- \boldsymbol{Z}_0 \left(\alpha_2^-(t)\boldsymbol{C}(t) + \beta_2^-(t)\boldsymbol{I}_{C(N+1)}\right)\end{aligned} \tag{45}$$

Moreover, since $\boldsymbol{b}'(t) = -\eta_2(t)\frac{\gamma C-\gamma-1}{C}\mathbf{1}_C$, we obtain

$$\boldsymbol{b}(t) = \boldsymbol{b}(0) + \int_0^t -\eta_2(\tau)\frac{\gamma C-\gamma-1}{C}\mathrm{d}\tau \mathbf{1}_C = \boldsymbol{b}_0 + \frac{(1+\gamma-\gamma C)\zeta_2(t)}{C}\mathbf{1}_C, \tag{46}$$

with $\zeta_2(t) = \int_0^t \eta_2(\tau)\mathrm{d}\tau$. ■

B.3 Proof of Corollary 3

Corollary 3 *Under the conditions and notation of Theorem 2, let* $s = \frac{\eta_1(0)}{\eta_2(0)}$, *if* $0 < \gamma < \frac{2}{C-2}$ *(where* $C > 2$*) or* $C = 2$, *and* $\lim_{t\to\infty} \zeta_1(t) = \infty$, *then the gradient flow (as in Equation* (5)*) will behave as:*

$$e^{-\frac{(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}} \boldsymbol{Z}(t) = \overline{\boldsymbol{Z}} + \boldsymbol{\Delta}(t), \tag{47}$$

where $\overline{\boldsymbol{Z}} = \left(\frac{1+\sqrt{s}}{2}\boldsymbol{H}_1^+ + \frac{1-\sqrt{s}}{2}\boldsymbol{H}_1^-, \frac{1+\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^+ - \frac{1-\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^-\right)$, $(\boldsymbol{H}_1^+, \boldsymbol{W}_1^+) = \Pi_1^+ \boldsymbol{Z}_0$, $(\boldsymbol{H}_1^-, \boldsymbol{W}_1^-) = \Pi_1^- \boldsymbol{Z}_0$, *and the residual term* $\boldsymbol{\Delta}(t)$ *decreases as* $\|\boldsymbol{\Delta}(t)\| = O\left(e^{\frac{\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\cdot\max\{-\gamma C,(C-2)\gamma-2\}}\right)$, *and so the normalized* $\boldsymbol{Z}(t)$ *converges to* $\frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|}$ *in*

$$\left\| \frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|} - \frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|} \right\| = O\left(e^{\frac{\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\cdot\max\{-\gamma C,(C-2)\gamma-2\}}\right), \tag{48}$$

which further indicates $\lim_{t\to\infty} \frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|} \in \mathcal{E}$. *Moreover, if* $\gamma \neq \frac{1}{C-1}$, *then* $\lim_{t\to\infty} \frac{\max_i b_i(t)}{\min_i b_i(t)} = 1$.

Proof Let $(\boldsymbol{H}_1^\epsilon, \boldsymbol{W}_1^\epsilon) = \Pi_1^\epsilon \boldsymbol{Z}_0$, $(\boldsymbol{H}_2^\epsilon, \boldsymbol{W}_2^\epsilon) = \Pi_2^\epsilon \boldsymbol{Z}_0$ and $(\boldsymbol{H}_3, \boldsymbol{W}_3) = \Pi_3 \boldsymbol{Z}_0$ for $\epsilon \in \{\pm 1\}$, according to Theorem 2, we have

$$\begin{aligned} \boldsymbol{H}(t) &= \sum_{\substack{i\in\{1,2\}\\ \epsilon\in\{\pm\}}} (\alpha_i^\epsilon(t)\zeta_1(t) + \beta_i^\epsilon(t))\boldsymbol{H}_i^\epsilon + \boldsymbol{H}_3, \\ \boldsymbol{W}(t) &= \sum_{\substack{i\in\{1,2\}\\ \epsilon\in\{\pm\}}} (\alpha_i^\epsilon(t)\zeta_2(t) + \beta_i^\epsilon(t))\boldsymbol{W}_i^\epsilon + \boldsymbol{W}_3. \end{aligned} \tag{49}$$

Since $\eta_1(t_1)\eta_2(t_2) = \eta_1(t_2)\eta_2(t_1)$, then $\eta_1(t) = \frac{\eta_1(0)}{\eta_2(0)}\eta_2(t)$. Let $s = \frac{\eta_1(0)}{\eta_2(0)}$, $p(t) = \frac{(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}$, and $q(t) = \frac{(1+\gamma-\gamma C)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}$, we have $\zeta_1(t) = \int_0^t \eta_1(\tau)\mathrm{d}\tau = s\int_0^t \eta_2(\tau)\mathrm{d}\tau = s\zeta_2(t)$, and then

$$\begin{aligned} \alpha_1^\epsilon(t)\zeta_1(t) + \beta_1^\epsilon(t) &= \tfrac{1+\sqrt{s}}{2}e^{\epsilon p(t)} + \tfrac{1-\sqrt{s}}{2}e^{-\epsilon p(t)} = \tfrac{1+\epsilon\sqrt{s}}{2}e^{p(t)} + O(e^{-p(t)}), \\ \alpha_1^\epsilon(t)\zeta_2(t) + \beta_1^\epsilon(t) &= \tfrac{1+\sqrt{s}}{2\sqrt{s}}e^{\epsilon p(t)} - \tfrac{1-\sqrt{s}}{2\sqrt{s}}e^{-\epsilon p(t)} = \tfrac{\epsilon+\sqrt{s}}{2\sqrt{s}}e^{p(t)} + O(e^{-p(t)}), \\ \alpha_2^\epsilon(t)\zeta_1(t) + \beta_2^\epsilon(t) &= \tfrac{1+\sqrt{s}}{2}e^{\epsilon q(t)} + \tfrac{1-\sqrt{s}}{2}e^{-\epsilon q(t)} = \tfrac{1+\epsilon\sqrt{s}}{2}e^{q(t)} + O(e^{-q(t)}), \\ \alpha_2^\epsilon(t)\zeta_2(t) + \beta_2^\epsilon(t) &= \tfrac{1+\sqrt{s}}{2\sqrt{s}}e^{\epsilon q(t)} - \tfrac{1-\sqrt{s}}{2\sqrt{s}}e^{-\epsilon q(t)} = \tfrac{\epsilon+\sqrt{s}}{2\sqrt{s}}e^{q(t)} + O(e^{-q(t)}). \end{aligned}$$

Since $0 < \gamma < \frac{2}{C-2}$ (where $C > 2$) or $C = 2$, then we have $p(t) - q(t) = \frac{\gamma C\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}} > 0$, $p(t) + q(t) = \frac{(2+2\gamma-\gamma C)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}} > 0$, and substitute these results into Equation (49) to

obtain

$$\begin{aligned} e^{-p(t)}\boldsymbol{H}(t) &= \frac{1+\sqrt{s}}{2}\boldsymbol{H}_1^+ + \frac{1-\sqrt{s}}{2}\boldsymbol{H}_1^- + \boldsymbol{\Delta}_1(t), \\ e^{-p(t)}\boldsymbol{W}(t) &= \frac{1+\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^+ - \frac{1-\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^- + \boldsymbol{\Delta}_2(t), \end{aligned} \tag{50}$$

where $\|\boldsymbol{\Delta}_1(t)\| = O(e^{\max\{q(t)-p(t),-q(t)-p(t)\}}$ and $\|\boldsymbol{\Delta}_2(t)\| = O(e^{\max\{q(t)-p(t),-q(t)-p(t)\}}$. Therefore, we have

$$e^{-p(t)}\boldsymbol{Z}(t) = \left(\tfrac{1+\sqrt{s}}{2}\boldsymbol{H}_1^+ + \tfrac{1-\sqrt{s}}{2}\boldsymbol{H}_1^-, \tfrac{1+\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^+ - \tfrac{1-\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^-\right) + \boldsymbol{\Delta}(t), \tag{51}$$

where $\boldsymbol{\Delta}(t) = (\boldsymbol{\Delta}_1(t), \boldsymbol{\Delta}_2(t))$ and $\|\boldsymbol{\Delta}(t)\| \le \|\boldsymbol{\Delta}_1(t)\| + \|\boldsymbol{\Delta}_2(t)\| = O(e^{\max\{q(t)-p(t),-q(t)-p(t)\}}$.

Let $\overline{\boldsymbol{Z}} = \left(\tfrac{1+\sqrt{s}}{2}\boldsymbol{H}_1^+ + \tfrac{1-\sqrt{s}}{2}\boldsymbol{H}_1^-, \tfrac{1+\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^+ - \tfrac{1-\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^-\right)$, we have

$$\left\| \frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|} - \frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|} \right\| = \left\| \frac{\overline{\boldsymbol{Z}} + \boldsymbol{\Delta}(t)}{\|\overline{\boldsymbol{Z}} + \boldsymbol{\Delta}(t)\|} - \frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|} \right\| \le \frac{2\|\overline{\boldsymbol{Z}}\|\|\boldsymbol{\Delta}(t)\|}{\|\overline{\boldsymbol{Z}} + \boldsymbol{\Delta}(t)\|\|\overline{\boldsymbol{Z}}\|} = \frac{2\|\boldsymbol{\Delta}(t)\|}{\|\overline{\boldsymbol{Z}} + \boldsymbol{\Delta}(t)\|}, \tag{52}$$

thus $\left\| \frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|} - \frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|} \right\| = O(e^{\max\{q(t)-p(t),-q(t)-p(t)\}}$, and further $\lim_{t\to\infty} \frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|} = \frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|}$ when $\lim_{t\to\infty} \zeta_1(t) = \infty$.

According to the definition of $\mathcal{E}_1$ and $\mathcal{E}_2$, we have

$$\overline{\boldsymbol{Z}} = \left(\tfrac{\sqrt{s}}{\sqrt{N}}\left(\tfrac{1+\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^+ - \tfrac{1-\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^-\right) \otimes \mathbf{1}_N^\top, \tfrac{1+\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^+ - \tfrac{1-\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^-\right), \tag{53}$$

thus we have $\overline{\boldsymbol{Z}} \in \mathcal{E}$, then $\lim_{t\to\infty} \frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|} = \frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|} \in \mathcal{E}$.

Moreover, we have $\boldsymbol{b}(t) = \boldsymbol{b}_0 + \frac{(1+\gamma-\gamma C)\zeta_2(t)}{C}\mathbf{1}_C$, then $\forall i, j$,

$$\lim_{t\to\infty} \frac{b_i(t)}{b_j(t)} = \lim_{t\to\infty} \frac{b_i(0) + \frac{1+\gamma-\gamma C}{C}\zeta_2(t)}{b_j(0) + \frac{1+\gamma-\gamma C}{C}\zeta_2(t)} = 1,$$

thus $\lim_{t\to\infty} \frac{\max_i b_i(t)}{\max_i b_i(t)}$=1. ■

B.4 Proof of Theorem 4

Theorem 4 (Dynamics of Features and Prototypes Under Weight Decay) *Consider the continual gradient flow in Equation 12, let $\boldsymbol{Z}(t) = (\boldsymbol{H}(t), \boldsymbol{W}(t))$. If $\eta_1(t_1)\eta_2(t_2) = \eta_1(t_2)\eta_2(t_1)$ for any $t_1, t_2 \ge 0$, we have the following closed-form dynamics:*

$$\begin{aligned} \boldsymbol{Z}(t) =& \Pi_1^+ \boldsymbol{Z}_0 \begin{pmatrix} a_1^+(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_1^+(t)\boldsymbol{I}_C \end{pmatrix} + \Pi_1^- \boldsymbol{Z}_0 \begin{pmatrix} a_1^-(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_1^-(t)\boldsymbol{I}_C \end{pmatrix} + \\ & \Pi_2^+ \boldsymbol{Z}_0 \begin{pmatrix} a_2^+(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_2^+(t)\boldsymbol{I}_C \end{pmatrix} + \Pi_2^- \boldsymbol{Z}_0 \begin{pmatrix} a_2^-(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_2^-(t)\boldsymbol{I}_C \end{pmatrix} + \\ & \Pi_3 \boldsymbol{Z}_0 \begin{pmatrix} a_3(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_3(t)\boldsymbol{I}_C \end{pmatrix} \end{aligned} \tag{54}$$

and

$$\boldsymbol{b}(t) = \phi(t)\left(\boldsymbol{b}_0 + \tfrac{1+\gamma-\gamma C}{C}\psi(t)\mathbf{1}_C\right), \tag{55}$$

where $\Pi_1^+\boldsymbol{Z}_0$, $\Pi_1^-\boldsymbol{Z}_0$, $\Pi_1^+\boldsymbol{Z}_0$, $\Pi_1^-\boldsymbol{Z}_0$, *and* $\Pi_3\boldsymbol{Z}_0$ *follow the definition in Theorem 2,* a_1^ϵ, a_2^ϵ, b_1^ϵ, b_2^ϵ, a_3, *and* b_3 *for* $\epsilon \in \{\pm\}$ *are the scalars that depend only on* C, N, γ, λ_1, λ_2, η_1, *and* η_2 *(where the detailed forms can be seen in B),* $\phi(t) = \exp(-\lambda\int_0^t \eta_2(\tau)\mathrm{d}\tau)$, *and* $\psi(t) = \int_0^t \zeta_2(\tau)\exp(\lambda\int_0^\tau \eta_2(s)\mathrm{d}s)\mathrm{d}\tau$.

Proof According to the gradient flow in Equation (12) and the notations in the proof of Theorem 2, we have:

$$\boldsymbol{Z}'(t) = \boldsymbol{Z}(t)\boldsymbol{A}(t) - \boldsymbol{Z}(t)\begin{pmatrix}\lambda_1\eta_1(t)\boldsymbol{I}_{CN} & 0\\ 0 & \lambda_2\eta_2(t)\boldsymbol{I}_C\end{pmatrix}, \tag{56}$$

i.e., $\boldsymbol{Z}'(t) = \boldsymbol{Z}(t)\boldsymbol{A}_\lambda(t)$, where $\boldsymbol{A}_\lambda(t) = \boldsymbol{A}(t) - \boldsymbol{\Lambda}(t)$ and $\boldsymbol{\Lambda}(t) = \left(\begin{smallmatrix}\lambda_1\eta_1(t)\boldsymbol{I}_{CN} & 0\\ 0 & \lambda_2\eta_2(t)\boldsymbol{I}_C\end{smallmatrix}\right)$.

For any t_1, t_2, we have the matrix commutator of $\boldsymbol{A}_\lambda(t_1)$ and $\boldsymbol{A}_\lambda(t_2)$

$$\begin{aligned}
&[\boldsymbol{A}_\lambda(t_1), \boldsymbol{A}_\lambda(t_2)]\\
=&\boldsymbol{A}_\lambda(t_1)\boldsymbol{A}_\lambda(t_2) - \boldsymbol{A}_\lambda(t_2)\boldsymbol{A}_\lambda(t_1)\\
=&[\boldsymbol{A}(t_1) - \boldsymbol{\Lambda}(t_1)][\boldsymbol{A}(t_2) - \boldsymbol{\Lambda}(t_2)] - [\boldsymbol{A}(t_2) - \boldsymbol{\Lambda}(t_2)][\boldsymbol{A}(t_1) - \boldsymbol{\Lambda}(t_1)]\\
=&\boldsymbol{\Lambda}(t_2)\boldsymbol{A}(t_1) - \boldsymbol{\Lambda}(t_1)\boldsymbol{A}(t_2) + \boldsymbol{A}(t_2)\boldsymbol{\Lambda}(t_1) - \boldsymbol{A}(t_1)\boldsymbol{\Lambda}(t_2)\\
=&\begin{pmatrix}0 & \lambda_1[\eta_1(t_2)\eta_2(t_1) - \eta_1(t_1)\eta_2(t_2)]\boldsymbol{M}^\top\\ \lambda_2[\eta_2(t_2)\eta_1(t_2) - \eta_2(t_1)\eta_1(t_2)]\boldsymbol{M} & 0\end{pmatrix}\\
=&0
\end{aligned} \tag{57}$$

where the last equality is based on the fact that $\eta_2(t_1)\eta_1(t_2) = \eta_2(t_2)\eta_1(t_1)$. Therefore, according to Magnus approach, we have

$$\boldsymbol{Z}(t) = \boldsymbol{Z}_0\exp\left(\int_0^t \boldsymbol{A}_\lambda(\tau)\mathrm{d}\tau\right) = \boldsymbol{Z}_0\exp\begin{pmatrix}-\lambda_1\zeta_1(t)\boldsymbol{I}_{CN} & \zeta_2(t)\boldsymbol{M}^\top\\ \zeta_1(t)\boldsymbol{M} & -\lambda_2\zeta_2(t)\boldsymbol{I}_C\end{pmatrix}, \tag{58}$$

where $\zeta_1(t) = \int_0^t \eta(\tau)\mathrm{d}\tau$ and $\zeta_2(t) = \int_0^t \eta(\tau)\mathrm{d}\tau$.

Let $\boldsymbol{B}_\lambda = \boldsymbol{B} - \left(\begin{smallmatrix}\lambda_1\boldsymbol{I}_{CN} & 0\\ 0 & \lambda_2\boldsymbol{I}_C\end{smallmatrix}\right)$, we have

$$\boldsymbol{Z}(t) = \boldsymbol{Z}_0\exp(\boldsymbol{B}_\lambda\boldsymbol{C}(t)) = \boldsymbol{Z}_0\sum_{k=0}^{\infty}\frac{(\boldsymbol{B}_\lambda\boldsymbol{C}(t))^k}{k!}. \tag{59}$$

We again consider the orthogonal decomposition of $\boldsymbol{Z}_0$, *i.e.*, $\boldsymbol{Z}_0 = (\Pi_1^+ + \Pi_1^- + \Pi_2^+ + \Pi_2^- + \Pi_3)\boldsymbol{Z}_0$. As mentioned in the proof of Theorem 2, we have

$$\begin{aligned}
&\Pi_1^\epsilon\boldsymbol{Z}_0\boldsymbol{B} = \tfrac{\epsilon(1+\gamma)}{C\sqrt{N}}\Pi_1^\epsilon\boldsymbol{Z}_0,\\
&\Pi_2^\epsilon\boldsymbol{Z}_0\boldsymbol{B} = \tfrac{\epsilon(1+\gamma-\gamma C)}{C\sqrt{N}}\Pi_2^\epsilon\boldsymbol{Z}_0, \text{ and}\\
&\Pi_3\boldsymbol{Z}_0\boldsymbol{B} = 0.
\end{aligned} \tag{60}$$

Therefore, for any $\boldsymbol{D} = (\boldsymbol{H}, \boldsymbol{W}) \in \{\Pi_1^\epsilon \boldsymbol{Z}_0, \Pi_2^\epsilon \boldsymbol{Z}_0, \Pi_3 \boldsymbol{Z}_0\}$ (where $\boldsymbol{H} \in \mathbb{R}^{p\times CN}$ and $\boldsymbol{W} \in \mathbb{R}^{p\times C}$) and the corresponding eigenvalue $\sigma \in \{\frac{\epsilon(1+\gamma)}{C\sqrt{N}}, \frac{\epsilon(1+\gamma-\gamma C)}{C\sqrt{N}}, 0\}$, we have

$$\boldsymbol{D}\boldsymbol{B} = \sigma \boldsymbol{D},\ \boldsymbol{H}\boldsymbol{M}^\top = \sigma \boldsymbol{W},\ \text{and}\ \boldsymbol{W}\boldsymbol{M} = \sigma \boldsymbol{H}. \tag{61}$$

In the following, we will prove that there exist two scalars $a(t)$ and $b(t)$, such that $\boldsymbol{D}\exp(\boldsymbol{B}_\lambda \boldsymbol{C}(t)) = \boldsymbol{D}\begin{pmatrix} a(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b(t)\boldsymbol{I}_C \end{pmatrix}$ for any $\boldsymbol{D} = (\boldsymbol{H}, \boldsymbol{W}) \in \{\Pi_1^\epsilon \boldsymbol{Z}_0, \Pi_2^\epsilon \boldsymbol{Z}_0, \Pi_3 \boldsymbol{Z}_0\}$.

First, we prove that $\boldsymbol{D}(\boldsymbol{B}_\lambda \boldsymbol{C}(t))^k$ can be represented as $\boldsymbol{D}(\boldsymbol{B}_\lambda \boldsymbol{C}(t))^k = (a_k(t)\boldsymbol{H}, b_k(t)\boldsymbol{W})$ by induction, where $a_k(t), b_k(t) \in \mathbb{R}$.

For $k = 0$, we have $\boldsymbol{D}(\boldsymbol{B}_\lambda \boldsymbol{C}(t))^0 = (\boldsymbol{H}, \boldsymbol{W})$, *i.e.*, $a_0 = b_0 = 1$. Assume that $\boldsymbol{D}(\boldsymbol{B}_\lambda \boldsymbol{C}(t))^n = (a_n(t)\boldsymbol{H}, b_n(t)\boldsymbol{W})$ for $k = n$. Then for $k = n+1$, we have

$$\begin{aligned}
&\boldsymbol{D}(\boldsymbol{B}_\lambda \boldsymbol{C}(t))^{n+1} \\
=&\boldsymbol{D}(\boldsymbol{B}_\lambda \boldsymbol{C}(t))^{n}(\boldsymbol{B}_\lambda \boldsymbol{C}(t)) \\
=&(a_n(t)\boldsymbol{H}, b_n(t)\boldsymbol{W})(\boldsymbol{B}_\lambda \boldsymbol{C}(t)) \\
=&(a_n(t)\boldsymbol{H}, b_n(t)\boldsymbol{W}) \begin{pmatrix} -\lambda_1 \boldsymbol{I}_{CN} & \boldsymbol{M}^\top \\ \boldsymbol{M} & -\lambda_2 \boldsymbol{I}_C \end{pmatrix} \begin{pmatrix} \zeta_1(t)\boldsymbol{I}_{CN} & 0 \\ 0 & \zeta_2(t)\boldsymbol{I}_C \end{pmatrix} \\
=&(b_n(t)\boldsymbol{W}\boldsymbol{M} - \lambda_1 a_n(t)\boldsymbol{H}, a_n(t)\boldsymbol{H}\boldsymbol{M}^\top - \lambda_2 b_n(t)\boldsymbol{W}) \begin{pmatrix} \zeta_1(t)\boldsymbol{I}_{CN} & 0 \\ 0 & \zeta_2(t)\boldsymbol{I}_C \end{pmatrix} \\
=&(\zeta_1(t)(\sigma b_n(t) - \lambda_1 a_n(t))\boldsymbol{H}, \zeta_2(t)(\sigma a_n(t) - \lambda_2 b_n(t))\boldsymbol{W})
\end{aligned}, \tag{62}$$

thus $a_{n+1}(t) = \zeta_1(t)(\sigma b_n(t) - \lambda_1 a_n(t))$ and $b_{n+1}(t) = \zeta_2(t)(\sigma a_n(t) - \lambda_2 b_n(t))$.

To sum up, we have shown by induction that $\boldsymbol{D}(\boldsymbol{B}_\lambda \boldsymbol{C}(t))^k$ can be represented as $\boldsymbol{D}(\boldsymbol{B}_\lambda \boldsymbol{C}(t))^k = (a_k(t)\boldsymbol{H}, b_k(t)\boldsymbol{W}) = \boldsymbol{D}\begin{pmatrix} a_k(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_k(t)\boldsymbol{I}_C \end{pmatrix}$, and $\begin{pmatrix} a_k(t) \\ b_k(t) \end{pmatrix}$ satisfies

$$\begin{pmatrix} a_k(t) \\ b_k(t) \end{pmatrix} = \begin{pmatrix} -\lambda_1\zeta_1(t) & \sigma\zeta_1(t) \\ \sigma\zeta_2(t) & -\lambda_2\zeta_2(t) \end{pmatrix} \begin{pmatrix} a_{k-1}(t) \\ b_{k-1}(t) \end{pmatrix} \quad \text{with} \quad \begin{pmatrix} a_0 \\ b_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{63}$$

i.e., $\begin{pmatrix} a_k(t) \\ b_k(t) \end{pmatrix} = (\boldsymbol{S}(\sigma, \lambda_1, \lambda_2, \zeta_1(t), \zeta_2(t)))^k \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, where $\boldsymbol{S}(\sigma, \lambda_1, \lambda_2, \zeta_1, \zeta_2) = \begin{pmatrix} -\lambda_1\zeta_1 & \sigma\zeta_1 \\ \sigma\zeta_2 & -\lambda_2\zeta_2 \end{pmatrix}$.

Therefore, we have

$$\boldsymbol{D}\exp(\boldsymbol{B}_\lambda \boldsymbol{C}(t)) = \boldsymbol{D}\sum_{k=0}^\infty \frac{(\boldsymbol{B}_\lambda \boldsymbol{C}(t))^k}{k!} = \boldsymbol{D}\begin{pmatrix} a(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b(t)\boldsymbol{I}_C \end{pmatrix}, \tag{64}$$

with $a(t) = \sum_{k=0}^\infty \frac{a_k(t)}{k!}$ and $b(t) = \sum_{k=0}^\infty \frac{b_k(t)}{k!}$, *i.e.*,

$$\begin{aligned}
\begin{pmatrix} a(t) \\ b(t) \end{pmatrix} &= \sum_{k=0}^\infty \frac{1}{k!} \begin{pmatrix} a_k(t) \\ b_k(t) \end{pmatrix} = \sum_{k=0}^\infty \frac{(\boldsymbol{S}(\sigma, \lambda_1, \lambda_2, \zeta_1(t), \zeta_2(t)))^k}{k!} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\
&= \exp(\boldsymbol{S}(\sigma, \lambda_1, \lambda_2, \zeta_1(t), \zeta_2(t))) \begin{pmatrix} 1 \\ 1 \end{pmatrix}.
\end{aligned} \tag{65}$$

Next, we are going to derive a concrete expression of $\exp(\boldsymbol{S}(\sigma, \lambda_1, \lambda_2, \zeta_1(t), \zeta_2(t)))$. Let the determinant $|\boldsymbol{S}(\sigma, \lambda_1, \lambda_2, \zeta_1(t), \zeta_2(t) - \theta \boldsymbol{I}| = 0$, we can derive that the eigenvalues of $\boldsymbol{S}(\sigma, \lambda_1, \lambda_2, \zeta_1(t), \zeta_2(t))$ are

$$\begin{aligned} \theta_1 &= \frac{-(\lambda_1\zeta_1(t)+\lambda_2\zeta_2(t))-\sqrt{(\lambda_1\zeta_1(t)-\lambda_2\zeta_2(t))^2+4\sigma^2\zeta_1(t)\zeta_2(t)}}{2} < 0 \text{ and} \\ \theta_2 &= \frac{-(\lambda_1\zeta_1(t)+\lambda_2\zeta_2(t))+\sqrt{(\lambda_1\zeta_1(t)-\lambda_2\zeta_2(t))^2+4\sigma^2\zeta_1(t)\zeta_2(t)}}{2}, \end{aligned} \tag{66}$$

and the corresponding eigenvectors are

$$\boldsymbol{v}_1 = \begin{pmatrix} 1 \\ \frac{\lambda_1\zeta_1(t)+\theta_1}{\sigma\zeta_1(t)} \end{pmatrix}, \text{ and } \boldsymbol{v}_2 = \begin{pmatrix} 1 \\ \frac{\lambda_1\zeta_1(t)+\theta_2}{\sigma\zeta_1(t)} \end{pmatrix}.$$

Let $\boldsymbol{P} = (\boldsymbol{v}_1, \boldsymbol{v}_2)$, we have

$$\boldsymbol{S}(\sigma, \lambda_1, \lambda_2, \zeta_1(t), \zeta_2(t)) = \boldsymbol{P}\begin{pmatrix} \theta_1 & 0 \\ 0 & \theta_2 \end{pmatrix}\boldsymbol{P}^{-1}, \ \boldsymbol{P}^{-1} = \frac{\sigma\zeta_1(t)}{\theta_2-\theta_1}\begin{pmatrix} \frac{\lambda_1\zeta_1(t)+\theta_2}{\sigma\zeta_1(t)} & -1 \\ -\frac{\lambda_1\zeta_1(t)+\theta_1}{\sigma\zeta_1(t)} & 1 \end{pmatrix}, \tag{67}$$

and

$$\begin{aligned} &\exp\left(\boldsymbol{S}(\sigma, \lambda, \zeta_1(t), \zeta_2(t))\right) = \boldsymbol{P}\begin{pmatrix} e^{\theta_1} & 0 \\ 0 & e^{\theta_2} \end{pmatrix}\boldsymbol{P}^{-1} \\ =&\frac{\sigma\zeta_1(t)}{\theta_2-\theta_1}\begin{pmatrix} 1 & 1 \\ \frac{\lambda_1\zeta_1(t)+\theta_1}{\sigma\zeta_1(t)} & \frac{\lambda_1\zeta_1(t)+\theta_2}{\sigma\zeta_1(t)} \end{pmatrix}\begin{pmatrix} e^{\theta_1} & 0 \\ 0 & e^{\theta_2} \end{pmatrix}\begin{pmatrix} \frac{\lambda_1\eta_1(t)+\theta_2}{\sigma\zeta_1(t)} & -1 \\ -\frac{\lambda_1\zeta_1(t)+\theta_1}{\sigma\zeta_1(t)} & 1 \end{pmatrix}, \\ =&\frac{1}{\theta_2-\theta_1}\begin{pmatrix} (\lambda_1\zeta_1(t)+\theta_2)e^{\theta_1} - (\lambda_1\zeta_1(t)+\theta_1)e^{\theta_2} & \sigma\zeta_1(t)e^{\theta_2} - \sigma\zeta_1(t)e^{\theta_1} \\ (\lambda_1\zeta_1(t)+\theta_2)e^{\theta_2} - (\lambda_1\zeta_1(t)+\theta_1)e^{\theta_1} & \sigma\zeta_2(t)e^{\theta_2} - \sigma\theta_2(t)e^{\theta_1} \end{pmatrix} \end{aligned} \tag{68}$$

thus

$$\begin{aligned} a(t) &= \frac{e^{\theta_2}}{\theta_2-\theta_1}\left[(\sigma\zeta_1(t) - \lambda_1\zeta_1(t) - \theta_1) - (\sigma\zeta_1(t) - \lambda_1\zeta_1(t) - \theta_2)e^{\theta_1-\theta_2}\right], \\ b(t) &= \frac{e^{\theta_2}}{\theta_2-\theta_1}\left[(\lambda_1\zeta_1(t) + \theta_2 + \sigma\zeta_2(t)) + (\lambda_1\zeta_1(t) + \theta_1 - \sigma\zeta_2(t))e^{\theta_1-\theta_2}\right]. \end{aligned} \tag{69}$$

Finally, since $\mathcal{E}_1^+$, $\mathcal{E}_1^-$, $\mathcal{E}_2^+$, $\mathcal{E}_2^-$ and $\mathcal{E}_3$ are orthogonal subspace of $\mathbb{R}^{p\times CN} \oplus \mathbb{R}^{C\times p}$, we obtain that

$$\begin{aligned} \boldsymbol{Z}(t) =&(\Pi_1^+\boldsymbol{Z}_0 + \Pi_1^-\boldsymbol{Z}_0 + \Pi_2^+\boldsymbol{Z}_0 + \Pi_2^-\boldsymbol{Z}_0 + \Pi_3\boldsymbol{Z}_0)\exp(\boldsymbol{B}_\lambda\boldsymbol{C}(t)) \\ =&\Pi_1^+\boldsymbol{Z}_0\begin{pmatrix} a_1^+(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_1^+(t)\boldsymbol{I}_C \end{pmatrix} + \Pi_1^-\boldsymbol{Z}_0\begin{pmatrix} a_1^-(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_1^-(t)\boldsymbol{I}_C \end{pmatrix} + \\ &\Pi_2^+\boldsymbol{Z}_0\begin{pmatrix} a_2^+(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_2^+(t)\boldsymbol{I}_C \end{pmatrix} + \Pi_2^-\boldsymbol{Z}_0\begin{pmatrix} a_2^-(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_2^-(t)\boldsymbol{I}_C \end{pmatrix} + \\ &\Pi_3\boldsymbol{Z}_0\begin{pmatrix} a_3(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_3(t)\boldsymbol{I}_C \end{pmatrix}, \end{aligned} \tag{70}$$

where $a_1^\epsilon(t)$, $b_1^\epsilon(t)$, $a_2^\epsilon(t)$, $b_2^\epsilon(t)$, $a_3(t)$ and $b_3(t)$ satisfy

$$\begin{aligned}
\begin{pmatrix} a_1^\epsilon(t) \\ b_1^\epsilon(t) \end{pmatrix} &= \exp\left(\boldsymbol{S}\left(\tfrac{\epsilon(1+\gamma)}{C\sqrt{N}}, \lambda_1, \lambda_2, \zeta_1(t), \zeta_2(t)\right)\right)\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \\
\begin{pmatrix} a_2^\epsilon(t) \\ b_2^\epsilon(t) \end{pmatrix} &= \exp\left(\boldsymbol{S}\left(\tfrac{\epsilon(1+\gamma-\gamma C)}{C\sqrt{N}}, \lambda_1, \lambda_2, \zeta_1(t), \zeta_2(t)\right)\right)\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \\
\begin{pmatrix} a_3(t) \\ b_3(t) \end{pmatrix} &= \exp\left(\boldsymbol{S}\left(0, \lambda_1, \lambda_2, \zeta_1(t), \zeta_2(t)\right)\right)\begin{pmatrix} 1 \\ 1 \end{pmatrix}.
\end{aligned} \tag{71}$$

Moreover, since $\boldsymbol{b}'(t) = \eta_2(t)\frac{1+\gamma-\gamma C}{C}\mathbf{1}_C - \lambda_2\eta_2(t)\boldsymbol{b}(t)$ is a first-order linear differential equation, then we have

$$\boldsymbol{b}(t) = \phi(t)\left(\boldsymbol{b}_0 + \tfrac{1+\gamma-\gamma C}{C}\psi(t)\mathbf{1}_C\right), \tag{72}$$

where $\phi(t) = \exp(-\lambda_2\int_0^t \eta_2(\tau)\mathrm{d}\tau)$ and $\psi(t) = \int_0^t \zeta_2(\tau)\exp(\lambda_2\int_0^\tau \eta_2(s)\mathrm{d}s)\mathrm{d}\tau$. ■

B.5 Proof of Corollary 5

Corollary 5. *Under the conditions and notation of Theorem 4, let $s = \frac{\eta_1(0)}{\eta_2(0)}$. If $0 < \gamma < \frac{2}{C-2}$ (where $C > 2$) or $C = 2$, $\lambda_1 = \lambda_2 = \lambda$, and $\lim_{t\to\infty}\zeta_1(t) = \infty$, then there exist constants $\pi_h^+, \pi_h^-, \pi_w^+, \pi_w^-$, and ω only depending on λ, γ, s, C, and N, such that the gradient flow (as in Equation (12)) behaves as:*

$$\left\|\frac{\boldsymbol{H}(t)}{\|\boldsymbol{H}(t)\|} - \frac{\pi_h^+\boldsymbol{H}_1^+ + \pi_h^-\boldsymbol{H}_1^-}{\|\pi_h^+\boldsymbol{H}_1^+ + \pi_h^-\boldsymbol{H}_1^-\|}\right\| + \left\|\frac{\boldsymbol{W}(t)}{\|\boldsymbol{W}(t)\|} - \frac{\pi_w^+\boldsymbol{W}_1^+ + \pi_w^-\boldsymbol{W}_1^-}{\|\pi_w^+\boldsymbol{H}_1^+ + \pi_w^-\boldsymbol{H}_1^-\|}\right\| = O(e^{-\omega\zeta_2(t)}), \tag{73}$$

where $(\boldsymbol{H}_1^+, \boldsymbol{W}_1^+) = \Pi_1^+\boldsymbol{Z}_0$, $(\boldsymbol{H}_1^-, \boldsymbol{W}_1^-) = \Pi_1^-\boldsymbol{Z}_0$. Furthermore, we have the following results:

- *If $\lambda > \frac{1+\gamma}{C\sqrt{N}}$, then $\lim_{t\to\infty}\|\boldsymbol{Z}(t)\| = 0$;*
- *If $\lambda = \frac{1+\gamma}{C\sqrt{N}}$, then $\lim_{t\to\infty}\boldsymbol{H}(t) = \boldsymbol{H}_1^+ + \frac{1-s}{1+s}\boldsymbol{H}_1^-, \lim_{t\to\infty}\boldsymbol{W}(t) = \boldsymbol{W}_1^+ - \frac{1-s}{1+s}\boldsymbol{W}_1^-$;*
- *If $\lambda < \frac{1+\gamma}{C\sqrt{N}}$, then $\lim_{t\to\infty}\|\boldsymbol{Z}(t)\| = \infty$.*

Proof Since $\zeta_1(t) = s\zeta_2(t)$, then the eigenvalues of $\boldsymbol{S}(\sigma, \lambda, \zeta_1(t), \zeta_2(t))$ satisfy

$$\begin{aligned}
\theta_1 &= -\frac{\lambda(s+1) + \sqrt{\lambda^2(s-1)^2 + 4s\sigma^2}}{2}\zeta_2(t) \to -\infty \text{ as } t \to \infty, \text{ and} \\
\theta_2 &= \frac{(\sqrt{\lambda^2(s-1)^2 + 4s\sigma^2} - \lambda(s+1))}{2}\zeta_2(t).
\end{aligned} \tag{74}$$

Let $\omega_1(\sigma, \lambda, s) = -\frac{\lambda(s+1)+\sqrt{\lambda^2(s-1)^2+4s\sigma^2}}{2} < 0$ and $\omega_2(\sigma, \lambda, s) = \frac{(\sqrt{\lambda^2(s-1)^2+4s\sigma^2}-\lambda(s+1))}{2}$. For brevity, let ω_1 and ω_2 denote $\omega_1(\sigma, \lambda, s)$ and $\omega_2(\sigma, \lambda, s)$, respectively, and then we can

reformulate $a(t)$ and $b(t)$ as

$$\begin{aligned} a(t) &= \frac{e^{\omega_2\zeta_2(t)}}{\omega_2-\omega_1}\left[(s\sigma-s\lambda-\omega_1)-(s\sigma-s\lambda-\omega_2)e^{(\omega_1-\omega_2)\zeta_2(t)}\right] \\ &= \tfrac{s\sigma-s\lambda-\omega_1}{\omega_2-\omega_1}e^{\omega_2\zeta_2(t)}+O(e^{(\omega_1)\zeta_2(t)}), \\ b(t) &= \frac{e^{\omega_2\zeta_2(t)}}{\omega_2-\omega_1}\left[(s\lambda+\omega_2+\sigma)+(s\lambda+\omega_1-\sigma)e^{(\omega_1-\omega_2)\zeta_2(t)}\right] \\ &= \tfrac{s\lambda+\omega_2+\sigma}{\omega_2-\omega_1}e^{\omega_2\zeta_2(t)}+O(e^{\omega_1\zeta_2(t)}). \end{aligned} \tag{75}$$

Moreover, according to Theorem 4, we have

$$\boldsymbol{H}(t)=\sum_{\substack{\epsilon\in\{\pm\}\\ i\in\{1,2\}}} a_i^\epsilon(t)\boldsymbol{H}_i^\epsilon+a_3(t)\boldsymbol{H}_3,\quad \boldsymbol{W}(t)=\sum_{\substack{\epsilon\in\{\pm\}\\ i\in\{1,2\}}} b_i^\epsilon(t)\boldsymbol{W}_i^\epsilon+b_3(t)\boldsymbol{W}_3, \tag{76}$$

with

$$\begin{aligned} \begin{pmatrix} a_1^\epsilon(t)\\ b_1^\epsilon(t)\end{pmatrix} &= \exp\left(\boldsymbol{S}\left(\tfrac{\epsilon(1+\gamma)}{C\sqrt{N}},\lambda,\zeta_1(t),\zeta_2(t)\right)\right)\begin{pmatrix}1\\1\end{pmatrix}, \\ \begin{pmatrix} a_2^\epsilon(t)\\ b_2^\epsilon(t)\end{pmatrix} &= \exp\left(\boldsymbol{S}\left(\tfrac{\epsilon(1+\gamma-\gamma C)}{C\sqrt{N}},\lambda,\zeta_1(t),\zeta_2(t)\right)\right)\begin{pmatrix}1\\1\end{pmatrix}, \\ \begin{pmatrix} a_3(t)\\ b_3(t)\end{pmatrix} &= \exp\left(\boldsymbol{S}\left(0,\lambda,\zeta_1(t),\zeta_2(t)\right)\right)\begin{pmatrix}1\\1\end{pmatrix}. \end{aligned} \tag{77}$$

Since $0<\gamma<\frac{2}{C-2}$ (where $C>2$) or $C=2$, thus we have $\frac{1+\gamma}{C\sqrt{N}}>\frac{|1+\gamma-\gamma C|}{C\sqrt{N}}$, and then $\omega_2\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right)>\omega_2\left(\frac{|1+\gamma-\gamma C|}{C\sqrt{N}},\lambda,s\right)$. When $t\to\infty$, the dominant terms in $\boldsymbol{H}(t)$ and $\boldsymbol{W}(t)$ are the ones whose coefficient contains $\exp\left(\omega_2\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right)\zeta_2(t)\right)$, *i.e.*, $a_1^+(t)$, $a_1^-(t)$, $b_1^+(t)$ and $b_1^-(t)$. Let $(\boldsymbol{H}_1^\epsilon,\boldsymbol{W}_1^\epsilon)=\Pi_1^\epsilon\boldsymbol{Z}_0$, $(\boldsymbol{H}_2^\epsilon,\boldsymbol{W}_2^\epsilon)=\Pi_2^\epsilon\boldsymbol{Z}_0$ and $(\boldsymbol{H}_3,\boldsymbol{W}_3)=\Pi_3\boldsymbol{Z}_0$ for $\epsilon\in\{\pm1\}$, thus we have

$$\begin{aligned} e^{-\omega_2\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right)\zeta_2(t)}\boldsymbol{H}(t) &= \pi_h^+(\lambda,\gamma,s,C,N)\boldsymbol{H}_1^+ + \pi_h^-(\lambda,\gamma,s,C,N)\boldsymbol{H}_1^- + \boldsymbol{\Delta}_1, \\ e^{-\omega_2\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right)\zeta_2(t)}\boldsymbol{W}(t) &= \pi_w^+(\lambda,\gamma,s,C,N)\boldsymbol{W}_1^+ + \pi_w^-(\lambda,\gamma,s,C,N)\boldsymbol{W}_1^- + \boldsymbol{\Delta}_2, \end{aligned} \tag{78}$$

where $\boldsymbol{\Delta}_1$ and $\boldsymbol{\Delta}_2$ decrease to zero as least as $O\left(e^{\frac{\left(\omega_2\left(\frac{|1+\gamma-\gamma C|}{C\sqrt{N}},\lambda,s\right)-\omega_2\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right)\right)\zeta_2(t)}{C\sqrt{N}}}\right)$, and

$$\begin{aligned}
\pi_h^+(\lambda,\gamma,s,C,N) &= \frac{s\frac{1+\gamma}{C\sqrt{N}} - s\lambda - \omega_1\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right)}{\omega_2\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right) - \omega_1\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right)},\\
\pi_h^-(\lambda,\gamma,s,C,N) &= \frac{-s\frac{1+\gamma}{C\sqrt{N}} - s\lambda - \omega_1\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right)}{\omega_2\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right) - \omega_1\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right)},\\
\pi_w^+(\lambda,\gamma,s,C,N) &= \frac{s\lambda + \omega_2\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right) + \frac{1+\gamma}{C\sqrt{N}}}{\omega_2\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right) - \omega_1\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right)},\\
\pi_w^-(\lambda,\gamma,s,C,N) &= \frac{s\lambda + \omega_2\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right) - \frac{1+\gamma}{C\sqrt{N}}}{\omega_2\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right) - \omega_1\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right)}.
\end{aligned} \tag{79}$$

Therefore, we have

$$\begin{aligned}
\lim_{t\to\infty}\frac{\boldsymbol{H}(t)}{\|\boldsymbol{H}(t)\|} &= \frac{\pi_h^+(\lambda,\gamma,s,C,N)\boldsymbol{H}_1^+ + \pi_h^-(\lambda,\gamma,s,C,N)\boldsymbol{H}_1^-}{\|\pi_h^+(\lambda,\gamma,s,C,N)\boldsymbol{H}_1^+ + \pi_h^-(\lambda,\gamma,s,C,N)\boldsymbol{H}_1^-\|},\\
\lim_{t\to\infty}\frac{\boldsymbol{W}(t)}{\|\boldsymbol{W}(t)\|} &= \frac{\pi_w^+(\lambda,\gamma,s,C,N)\boldsymbol{W}_1^+ + \pi_w^-(\lambda,\gamma,s,C,N)\boldsymbol{W}_1^-}{\|\pi_w^+(\lambda,\gamma,s,C,N)\boldsymbol{W}_1^+ + \pi_w^-(\lambda,\gamma,s,C,N)\boldsymbol{W}_1^-\|},
\end{aligned} \tag{80}$$

and the rate of convergence is $O\left(e^{\frac{\left(\omega_2\left(\frac{|1+\gamma-\gamma C|}{C\sqrt{N}},\lambda,s\right)-\omega_2\left(\frac{1+\gamma}{C\sqrt{N}},\lambda,s\right)\right)\zeta_2(t)}{C\sqrt{N}}}\right)$.

Moreover, we have the following conclusions:

- If $\lambda = \frac{1+\gamma}{C\sqrt{N}}$, we have $\omega_2 = 0$, $\omega_1 = -\lambda(s+1)$, $\pi_h^+ = 1$, $\pi_h^- = \frac{1-s}{1+s}$, $\pi_w^+ = 1$ and $\pi_w^- = -\frac{1-s}{1+s}$. Since $\lim_{t\to\infty}\zeta_2(t) = \infty$, we have

$$\lim_{t\to\infty}\boldsymbol{H}(t) = \boldsymbol{H}_1^+ + \frac{1-s}{1+s}\boldsymbol{H}_1^-, \lim_{t\to\infty}\boldsymbol{W}(t) = \boldsymbol{W}_1^+ - \frac{1-s}{1+s}\boldsymbol{W}_1^-. \tag{81}$$

- If $\lambda > \frac{1+\gamma}{C\sqrt{N}}$, we have $\omega_2 > 0$, and then $\lim_{t\to\infty}\|\boldsymbol{Z}(t)\| = 0$ since $\lim_{t\to\infty}\zeta_2(t) = \infty$.
- If $\lambda < \frac{1+\gamma}{C\sqrt{N}}$, we have $\omega_2 < 0$, and then $\lim_{t\to\infty}\|\boldsymbol{Z}(t)\| = \infty$ since $\lim_{t\to\infty}\zeta_2(t) = \infty$.

So far the proof has been completed. ■

B.6 Proof of Theorem 7

Lemma 10. *For $\boldsymbol{h}(t), \boldsymbol{w} \in \mathbb{R}^p$, $\eta(t) > 0$, let $\hat{\boldsymbol{v}} = \frac{\hat{\boldsymbol{v}}}{\|\boldsymbol{v}\|_2}$ denote the ℓ_2-normalized vector of $\boldsymbol{v}$, considering the discrete dynamical system $\boldsymbol{h}(t+1) = \boldsymbol{h}(t) + \frac{\eta(t)}{\|\boldsymbol{h}(t)\|_2}\left(\boldsymbol{I}_p - \hat{\boldsymbol{h}}(t)\hat{\boldsymbol{h}}^\top(t)\right)\boldsymbol{w}$,*

if $\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{h}}(0) > -1$*, the learning rate* $\eta(t)$ *satisfies that* $\lim_{t\to\infty} \frac{\eta(t+1)}{\eta(t)} = 1$*,* $\frac{\eta(t)}{\|\boldsymbol{h}(t)\|_2}$ *is non-increasing with* $\frac{\eta(0)}{\|\boldsymbol{h}(0)\|_2} < \frac{1}{\|\boldsymbol{w}\|_2}$*, and there exists a constant* $\varepsilon > 0$*, s.t.,* $\eta(t) > \varepsilon$*, s.t.,* $\eta(t) > \varepsilon$*, then we have*

$$\lim_{t\to\infty} \left\| \frac{\boldsymbol{h}(t)}{\|\boldsymbol{h}(t)\|_2} - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \right\| = 0. \tag{82}$$

Proof For brevity, let $\alpha_t = \frac{\eta(t)\|\boldsymbol{w}\|_2}{\|\boldsymbol{h}(t)\|_2^2}$, $\xi_t = \frac{\eta(t+1)}{\eta(t)}$, and $\beta_t = \hat{\boldsymbol{w}}^\top \hat{\boldsymbol{h}}(t)$ denote the cosine similarity between $\hat{\boldsymbol{w}}$ and $\hat{\boldsymbol{h}}(t)$, then we can easily derive that $\alpha_t > 0$ and $\beta_t > -1$ for all $t \geq 0$.

We will show that α_t *is monotonically decreasing and* β_t *is monotonically increasing.* Note that $\boldsymbol{h}(t)$ is orthogonal with $\left(\boldsymbol{I}_p - \hat{\boldsymbol{h}}(t)\hat{\boldsymbol{h}}^\top(t)\right)\boldsymbol{w}$, thus we have

$$\|\boldsymbol{h}(t+1)\|_2^2 = \|\boldsymbol{h}(t)\|_2^2 + \frac{\eta^2(t)}{\|\boldsymbol{h}(t)\|_2^2} \left\| \left(\boldsymbol{I}_p - \hat{\boldsymbol{h}}(t)\hat{\boldsymbol{h}}^\top(t)\right)\boldsymbol{w} \right\|_2^2 \geq \|\boldsymbol{h}(t)\|_2^2,$$

which indicates $\|\boldsymbol{h}(t)\|_2$ is monotonically increasing as a function of t, and then α_t *is monotonically decreasing* since $\frac{\eta(t)\|\boldsymbol{w}\|_2}{\|\boldsymbol{h}(t)\|_2}$ is non-increasing.

Moreover, we can rearrange the discrete dynamics and formulate $\boldsymbol{h}(t+1)$ as a positive combination of $\boldsymbol{h}(t)$ and $\boldsymbol{w}$:

$$\boldsymbol{h}(t+1) = \left(1 - \frac{\eta(t)\boldsymbol{w}^\top \hat{\boldsymbol{h}}(t)}{\|\boldsymbol{h}(t)\|_2}\right)\boldsymbol{h}(t) + \frac{\eta(t)}{\|\boldsymbol{h}(t)\|_2}\boldsymbol{w}, \tag{83}$$

so $\hat{\boldsymbol{w}}^\top \hat{\boldsymbol{h}}(t+1) \geq \hat{\boldsymbol{w}}^\top \hat{\boldsymbol{h}}(t)$, *i.e.,* β_t *is monotonically increasing*, which is based on the facts that $1 - \frac{\eta(t)\boldsymbol{w}^\top \hat{\boldsymbol{h}}(t)}{\|\boldsymbol{h}(t)\|_2} \geq 1 - \frac{\eta(t)\|\boldsymbol{w}\|_2}{\|\boldsymbol{h}(t)\|_2} \geq 1 - \frac{\eta(0)\|\boldsymbol{w}\|_2}{\|\boldsymbol{h}(0)\|_2} > 0$, $\frac{\eta(t)}{\|\boldsymbol{h}(t)\|_2} > 0$, and $\frac{\boldsymbol{y}^\top(\boldsymbol{x}+k\boldsymbol{y})}{\|\boldsymbol{x}+k\boldsymbol{y}\|_2} \geq \frac{\boldsymbol{y}^\top \boldsymbol{x}}{\|\boldsymbol{x}\|_2}$ holds for all $k > 0$ and $\boldsymbol{x}, \boldsymbol{y} \neq 0$.

We can formulate the discrete iterations of α_t and β_t from the Equation (83) as follows:

$$\begin{aligned}
\beta_{t+1} &= \frac{\boldsymbol{w}^\top \boldsymbol{h}(t+1)}{\|\boldsymbol{w}\|_2 \|\boldsymbol{h}(t+1)\|_2} \\
&= \frac{\boldsymbol{w}^\top \left(\boldsymbol{h}(t) + \frac{\eta(t)}{\|\boldsymbol{h}(t)\|_2}\left(\boldsymbol{I}_p - \hat{\boldsymbol{h}}(t)\hat{\boldsymbol{h}}^\top(t)\right)\boldsymbol{w}\right)}{\|\boldsymbol{w}\|_2 \sqrt{\|\boldsymbol{h}(t)\|_2^2 + \frac{\eta^2(t)}{\|\boldsymbol{h}(t)\|_2^2}\left\|\left(\boldsymbol{I}_p - \hat{\boldsymbol{h}}(t)\hat{\boldsymbol{h}}^\top(t)\right)\boldsymbol{w}\right\|_2^2}} \\
&= \frac{\beta_t + \frac{\eta(t)\|\boldsymbol{w}\|_2}{\|\boldsymbol{h}(t)\|_2^2}(1-\beta_t^2)}{\sqrt{1 + \frac{\eta^2(t)\|\boldsymbol{w}\|_2^2}{\|\boldsymbol{h}(t)\|_2^4}(1-\beta_t^2)}} \\
&= \frac{\beta_t + \alpha_t(1-\beta_t^2)}{\sqrt{1+\alpha_t^2(1-\beta_t^2)}}, \\
\alpha_{t+1} &= \frac{\eta(t+1)\|\boldsymbol{w}\|_2}{\|\boldsymbol{h}(t+1)\|_2^2} \\
&= \frac{\xi_t \eta(t)\|\boldsymbol{w}\|_2}{\|\boldsymbol{h}(t)\|_2^2 + \frac{\eta^2(t)}{\|\boldsymbol{h}(t)\|_2^2}\left\|\left(\boldsymbol{I}_p - \hat{\boldsymbol{h}}(t)\hat{\boldsymbol{h}}^\top(t)\right)\boldsymbol{w}\right\|_2^2} \\
&= \frac{\xi_t \alpha_t}{1+\alpha_t^2(1-\beta_t^2)},
\end{aligned} \tag{84}$$

with $\beta_0 = \hat{\boldsymbol{w}}^\top \hat{\boldsymbol{h}}(0) > -1$, $\alpha_0 = \frac{\eta(0)\|\boldsymbol{w}\|_2}{\|\boldsymbol{h}(0)\|_2^2} > 0$, $\xi_t \leq 1$ and $\lim_{t\to\infty} \xi_t = 1$.

To prove $\lim_{t\to\infty} \left\| \frac{\boldsymbol{h}(t)}{\|\boldsymbol{h}(t)\|_2} - \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2} \right\| = 0$, we just need prove $\lim_{t\to\infty} \beta_t = 1$. Note that α_t is monotonic decreasing and lower bounded by 0, then the sequence (α_t) is convergent. Similarly, the sequence (β_t) is convergent. Let $a = \lim_{t\to} \alpha_t$ and $b = \lim_{t\to\beta_t}$, we obtain

$$\lim_{t\to\infty} \alpha_{t+1} = \lim_{t\to\infty} \frac{\xi_t \alpha_t}{1+\alpha_t^2(1-\beta_t^2)} \quad \Rightarrow \quad a = \frac{a}{1+\lim_{t\to\infty}\alpha_t^2(1-\beta_t^2)}, \tag{85}$$

thus $a = 0$ or $\lim_{t\to\infty} \alpha_t^2(1-\beta_t^2) = 0$, *i.e.*, $a = 0$ or $b = 1$. Therefore, the limits of α_t and β_t exist if and only if $\lim_{t\to\infty} \alpha_t = 0$ or $\lim_{t\to\infty} \beta_t = 1$. In the following, we will prove that the limit of β_t must be equal to 1.

Firstly, we prove a simpler result when $\beta_0 > 0$:

Lemma 11. *For the discrete dynamics in Equation* (84)*, if* $\beta_0 \geq 0$*, then* $\lim_{t\to\infty} \beta_t = 1$*.*

Proof As aforementioned, due to the existence of the limit of α_t, we have $\lim_{t\to\infty} \alpha_t = 0$ or $\lim_{t\to\infty} \beta_t = 1$. Thus, we just need to prove that $\lim_{t\to\infty} \beta_t = 1$ as $\lim_{t\to\infty} \alpha_t = 0$.

When $\lim_{t\to\infty} \alpha_t = 0$, then there exists τ, such that $\forall t > \tau$, $\alpha_t \leq 1$.

According to the iterations in Equation (84), we can derive that

$$\begin{aligned}\frac{1-\beta_{t+1}^2}{\alpha_{t+1}} &= \frac{1+\alpha_t^2-\alpha_t^2\beta_t^2-(\alpha_t+\beta_t-\alpha_t\beta_t^2)^2}{\xi_t\alpha_t}\\ &= \frac{1+\alpha_t^2-\alpha_t^2\beta_t^2-\alpha_t^2-\beta_t^2-\alpha_t^2\beta_t^4-2\alpha_t\beta_t+2\alpha_t^2\beta_t^2+2\alpha_t\beta_t^3}{\xi_t\alpha_t}\\ &= \frac{1-\beta_t^2+\alpha_t^2(\beta_t^2-\beta_t^4)-2\alpha_t(\beta_t-\beta_t^3)}{\xi_t\alpha_t}\\ &= \frac{1-\beta_t^2}{\xi_t\alpha_t}\cdot(1+\alpha_t^2\beta_t^2-2\alpha_t\beta_t)\\ &= \frac{1-\beta_t^2}{\alpha_t}\cdot\frac{(1-\alpha_t\beta_t)^2}{\xi_t},\end{aligned}$$

then $\forall t>\tau$,

$$\begin{aligned}1-\beta_{t+1}^2 &= \alpha_{t+1}\cdot\frac{1-\beta_0^2}{\alpha_0}\prod_{i=0}^{t}\frac{(1-\alpha_i\beta_i)^2}{\xi_i}\\ &= \alpha_{t+1}\cdot\frac{\eta(0)(1-\beta_0^2)}{\alpha_0\eta(t+1)}\cdot\prod_{i=0}^{\tau}(1-\alpha_i\beta_i)^2\cdot\prod_{i=\tau+1}^{t}(1-\alpha_i\beta_i)^2 \qquad (86)\\ &\le \alpha_{t+1}\cdot\frac{\eta(0)(1-\beta_0^2)}{\varepsilon\alpha_0}\cdot\prod_{i=0}^{\tau}(1-\alpha_i\beta_i)^2,\end{aligned}$$

where the inequality is based on the fact that $1-\alpha_i\beta_i\in(0,1]$ when $0\le\beta_0\le\beta_i\le 1$, $\eta(t+1)\ge\varepsilon$, and $\alpha_i\le 1$ for $i>\tau$. Since $\frac{\eta(0)(1-\beta_0^2)}{\varepsilon\alpha_0}\cdot\prod_{i=0}^{\tau}(1-\alpha_i\beta_i)^2$ is a constant, we obtain

$$\lim_{t\to\infty}1-\beta_{t+1}^2\le\lim_{t\to\infty}\alpha_{t+1}\cdot\frac{\eta(0)(1-\beta_0^2)}{\varepsilon\alpha_0}\cdot\prod_{i=0}^{\tau}(1-\alpha_i\beta_i)^2=0,$$

as $\lim_{t\to\infty}\alpha_{t+1}=0$. This reveals $\lim_{t\to\infty}\beta_t^2=1$. Furthermore, since $\beta_t\ge 0$, we then have $\lim_{t\to\infty}\beta_t=1$. ■

Next, we are going to prove $\lim_{t\to\infty}\beta_t=1$ when $-1<\beta_0<0$. According to Lemma 11, we just need prove that $\exists\tau>0$, *s.t.*, $\beta_\tau\ge 0$.

For the sake of contradiction, suppose that $\beta_t<0$ for all $t>0$, we then have $\lim_{t\to\infty}\alpha_t=0$. As a consequence, we obtain

$$\alpha_t+\beta_t-\alpha_t\beta_t^2<0,\ \forall t\ge 0, \qquad (87)$$

and we know that $\exists t'>0$, such that

$$\alpha_t<\frac{\varepsilon}{\eta(0)}\alpha_0(1-\beta_0^2),\ \forall t\ge t'. \qquad (88)$$

According to the iterations in Equation (84), we can derive that

$$\begin{aligned}\alpha_{t+1}(1-\beta_{t+1}^2) &= \frac{\xi_t\alpha_t}{1+\alpha_t^2-\alpha_t^2\beta_t^2}\left(1-\frac{(\alpha_t+\beta_t-\alpha_t\beta_t^2)^2}{1+\alpha_t^2-\alpha_t^2\beta_t^2}\right)\\ &= \xi_t\alpha_t\left(\frac{1+\alpha_t^2-\alpha_t^2\beta_t^2-(\alpha_t+\beta_t-\alpha_t\beta_t^2)^2}{(1+\alpha_t^2-\alpha_t^2\beta_t^2)^2}\right)\\ &= \xi_t\alpha_t(1-\beta_t^2)\left(\frac{1-\alpha_t\beta_t}{1+\alpha_t^2-\alpha_t^2\beta_t^2}\right)^2\\ &= \xi_t\alpha_t(1-\beta_t^2)\left(1-\frac{\alpha_t(\alpha_t+\beta_t-\alpha_t\beta_t^2)}{1+\alpha_t^2-\alpha_t^2\beta_t^2}\right)^2,\end{aligned}$$

Since $1-\frac{\alpha_t(\alpha_t+\beta_t-\alpha_t\beta_t^2)}{1+\alpha_t^2-\alpha_t^2\beta_t^2}\geq 1$, then for $t\geq t'$,

$$\alpha_t\geq\alpha_t(1-\beta_t^2)\geq\xi_{t-1}\alpha_{t-1}(1-\beta_{t-1}^2)\geq\ldots\geq\alpha_0(1-\beta_0^2)\prod_{i=0}^{t-1}\xi_i\geq\frac{\varepsilon}{\eta(0)}\alpha_0(1-\beta_0^2),\quad(89)$$

which contradicts the fact in Equation (88). Thus, $\exists\tau>0$, s.t. $\beta_\tau\geq 0$. Consider the dynamical system with an initial time τ, we have $\lim_{t\to\infty}\beta_t=1$ according to Lemma 11.

To sum up, we have proven that $\lim_{t\to\infty}\beta_t=1$ when $\beta_0>-1$ and $\alpha_0>0$. ■

Theorem 7 (**Convergence in the Spherical Constrained Case**) *Considering the discrete dynamics in Equation* (19)*, if* $\forall i\in[N],c\in[C]$, $\hat{\boldsymbol{w}}_c^\top\hat{\boldsymbol{h}}_{i,c}(0)>-1$*, the learning rate* $\eta(t)$ *satisfies that* $\frac{\eta(t)}{\|\boldsymbol{h}_{i,c}(t)\|_2}$ *is non-increasing,* $\frac{\eta(0)(1+\gamma)}{CN\|\boldsymbol{h}_{i,c}(0)\|_2}\leq\frac{1}{\|\boldsymbol{w}_c\|_2}$, $\lim_{t\to\infty}\frac{\eta(t+1)}{\eta(t)}=1$*, and there exists a constant* $\varepsilon>0$*, s.t.,* $\eta(t)>\varepsilon$*, then we have*

$$\lim_{t\to\infty}\left\|\hat{\boldsymbol{H}}(t)-\hat{\boldsymbol{W}}(\boldsymbol{I}_C\otimes\mathbf{1}_N^\top)\right\|=0,\quad(90)$$

and further if $\lim_{t\to\infty}\|\boldsymbol{H}(t)\|<\infty$*, then there exists a constant* $\mu>0$*, such that the error above shows exponential convergence:*

$$\left\|\hat{\boldsymbol{H}}(t)-\hat{\boldsymbol{W}}(\boldsymbol{I}_C\otimes\mathbf{1}_N^\top)\right\|\leq O(e^{-\mu t}).\quad(91)$$

Moreover, if $\hat{\boldsymbol{w}}_c^\top\hat{\boldsymbol{h}}_{i,c}(0)=-1$*, then* $\boldsymbol{h}_{i,c}(t)=\boldsymbol{h}_{i,c}(0)$.

Proof Since $\boldsymbol{H}(t+1)=\boldsymbol{H}(t)+\frac{(1+\gamma)\eta(t)}{CN}\left(\frac{\partial\hat{\boldsymbol{H}}}{\partial\boldsymbol{H}}\big|_{\boldsymbol{H}=\boldsymbol{H}(t)}\right)^\top\boldsymbol{W}(\boldsymbol{I}_C\otimes\mathbf{1}_N^\top)$, then for $i\in[N],c\in[C]$,

$$\boldsymbol{h}_{i,c}(t+1)=\boldsymbol{h}_{i,c}(t)+\frac{(1+\gamma)\eta(t)}{CN\|\boldsymbol{h}_{i,c}(t)\|_2}\left(\boldsymbol{I}_p-\hat{\boldsymbol{h}}_{i,c}(t)\hat{\boldsymbol{h}}_{i,c}^\top(t)\right)\boldsymbol{w}_c.\quad(92)$$

According to Lemma 10, when $\hat{\boldsymbol{w}}_c^\top\hat{\boldsymbol{h}}_{i,c}(0)>-1$ and $\frac{\eta(0)(1+\gamma)}{CN}<\frac{\|\boldsymbol{h}_{i,c}(0)\|_2^2}{\|\boldsymbol{w}_c\|_2}$, we have $\lim_{t\to\infty}\|\hat{\boldsymbol{h}}_{i,c}-\hat{\boldsymbol{w}}_c\|=0$, then $\lim_{t\to\infty}\left\|\hat{\boldsymbol{H}}(t)-\hat{\boldsymbol{W}}(\boldsymbol{I}_C\otimes\mathbf{1}_N^\top)\right\|=0$.

If further $\lim_{t\to\infty}\|\boldsymbol{H}(t)\|_2<\infty$, let $L=\sup_{i,c,t}\|\boldsymbol{h}_{i,c}(t)\|$. According to the proof of Lemma 10, $\forall i,c$, we have $\lim_{t\to\infty}\hat{\boldsymbol{w}}_c^\top\hat{\boldsymbol{h}}_{i,c}(t)=1$, then for a given constant $\delta>0$, $\exists\tau>0$,

s.t., $\forall t > \tau$, $\hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(\tau) \geq \delta$. Consider $t > \tau$, we have

$$
\begin{aligned}
&1 - \left(\hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(t+1)\right)^2 \\
=&\tfrac{\|\boldsymbol{h}_{i,c}(t)\|_2^2}{\|\boldsymbol{h}_{i,c}(t+1)\|_2^2}\left(1 - \left(\hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(t)\right)^2\right)\left(1 - \tfrac{(1+\gamma)\eta(t)\|\boldsymbol{w}_c\|_2}{CN\|\boldsymbol{h}_{i,c}(t)\|_2^2} \cdot \hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(t)\right)^2 \\
\leq&\left(1 - \left(\hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(0)\right)^2\right)\prod_{j=0}^{t}\left(1 - \tfrac{(1+\gamma)\eta(j)\|\boldsymbol{w}_c\|_2}{CN\|\boldsymbol{h}_{i,c}(j)\|_2^2} \cdot \hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(j)\right)^2 \\
\leq&\left(1 - \left(\hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(0)\right)^2\right)\prod_{j=0}^{\tau}\left(1 - \tfrac{(1+\gamma)\eta(j)\|\boldsymbol{w}_c\|_2}{CN\|\boldsymbol{h}_{i,c}(j)\|_2^2} \cdot \hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(j)\right)^2 \prod_{j=\tau+1}^{t}\left(1 - \tfrac{(1+\gamma)\varepsilon\delta\|\boldsymbol{w}_c\|_2}{CN\|\boldsymbol{h}_{i,c}(j)\|_2^2}\right)^2 \\
\leq&\left(1 - \left(\hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(0)\right)^2\right)\prod_{j=0}^{\tau}\left(1 - \tfrac{(1+\gamma)\eta(j)\|\boldsymbol{w}_c\|_2}{CN\|\boldsymbol{h}_{i,c}(j)\|_2^2} \cdot \hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(j)\right)^2 \left(1 - \tfrac{(1+\gamma)\varepsilon\delta\|\boldsymbol{w}_c\|_2}{CNL^2}\right)^{2(t-\tau)}.
\end{aligned}
$$

where the first, the second, and the third inequalities are based on the facts that $\frac{\|\boldsymbol{h}_{i,c}(t)\|_2^2}{\|\boldsymbol{h}_{i,c}(t+1)\|_2^2} \leq 1$, $1 - \frac{(1+\gamma)\eta(j+1)\|\boldsymbol{w}_c\|_2}{CN\|\boldsymbol{h}_{i,c}(j)\|_2^2} \cdot \hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(j) \leq 1 - \frac{(1+\gamma)\varepsilon\delta\|\boldsymbol{w}_c\|_2}{CN\|\boldsymbol{h}_{i,c}(j)\|_2^2}$ for $t > \tau$, and $\|\boldsymbol{h}_{i,c}(j)\|_2 \leq L$, respectively.

Let $c_1 = \max_{i,c}(1-(\hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(0))^2)\prod_{j=0}^{\tau}\left(1 - \frac{(1+\gamma)\eta(j)\|\boldsymbol{w}_c\|_2}{CN\|\boldsymbol{h}_{i,c}(j)\|_2^2} \cdot \hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(j)\right)^2\left(1 - \frac{(1+\gamma)\varepsilon\delta\|\boldsymbol{w}_c\|_2}{CNL^2}\right)^{-2\tau}$, and $\mu = \min_c -2\log\left(1 - \frac{(1+\gamma)\varepsilon\delta\|\boldsymbol{w}_c\|_2}{CNL^2}\right)$, then $1 - \hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(t+1) \leq \frac{c_1 e^{-\mu t}}{1+\delta}$.

Therefore, we have

$$
\left\|\hat{\boldsymbol{H}}(t) - \hat{\boldsymbol{W}}(\boldsymbol{I}_C \otimes \boldsymbol{1}_N^\top)\right\|_2^2 = 2\sum_{i,c}\left(1 - \hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(t+1)\right) \leq \frac{2c_1 CNe^{-\mu t}}{(1+\delta)}, \tag{93}
$$

i.e., $\left\|\hat{\boldsymbol{H}}(t) - \hat{\boldsymbol{W}}(\boldsymbol{I}_C \otimes \boldsymbol{1}_N^\top)\right\| = O(e^{-\mu t})$.

Moreover, if $\hat{\boldsymbol{w}}_c^\top \hat{\boldsymbol{h}}_{i,c}(0) = -1$, we have

$$
\boldsymbol{h}_{i,c}(t+1) = \boldsymbol{h}_{i,c}(t) + \frac{(1+\gamma)\eta(t)}{CN\|\boldsymbol{h}_{i,c}(t)\|_2}\left(\boldsymbol{I}_p - \hat{\boldsymbol{h}}_{i,c}(t)\hat{\boldsymbol{h}}_{i,c}^\top(t)\right)\boldsymbol{w}_c = \boldsymbol{h}_{i,c}(t), \tag{94}
$$

thus $\boldsymbol{h}_{i,c}(t) = \boldsymbol{h}_{i,c}(0)$. ■

B.7 Proof of Theorem 8

Theorem 12. *Let $\boldsymbol{z}(t)$ denote the row-first vectorization of $\begin{pmatrix}\boldsymbol{H}(t) & 0\\ 0 & \boldsymbol{W}(t)\end{pmatrix}$, $\boldsymbol{B} = \begin{pmatrix}0 & \boldsymbol{M}^\top\\ \boldsymbol{M} & 0\end{pmatrix}$, and $\boldsymbol{A} = \begin{pmatrix}0 & \nabla_{\boldsymbol{\Theta}}\boldsymbol{H}^\top\nabla_{\boldsymbol{\Theta}}\boldsymbol{H}\\ \mathbf{I}_C & 0\end{pmatrix}$. Considering the eigendecomposition $\boldsymbol{A} = \boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{U}_{\boldsymbol{A}}^{-1}$ and $\boldsymbol{B} = \boldsymbol{U}_{\boldsymbol{B}}\boldsymbol{\Lambda}_{\boldsymbol{B}}\boldsymbol{U}_{\boldsymbol{B}}^{-1}$, we have*

$$
\boldsymbol{C}\boldsymbol{z}(t) = \exp[(\boldsymbol{\Lambda}_{\boldsymbol{A}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{B}})t]\boldsymbol{C}\boldsymbol{z}(0), \tag{95}
$$

where $\boldsymbol{C} = \boldsymbol{U}_{\boldsymbol{B}}^{-1}\boldsymbol{U}_{\boldsymbol{A}}^{-1} \otimes \mathbf{I}$.

Proof Since $\boldsymbol{z}(t)$ denote the row-first vectorization of $\begin{pmatrix} \boldsymbol{H}(t) & 0 \\ 0 & \boldsymbol{W}(t) \end{pmatrix}$, $\boldsymbol{A} = \begin{pmatrix} 0 & \nabla_{\boldsymbol{\Theta}} \boldsymbol{H}^\top \nabla_{\boldsymbol{\Theta}} \boldsymbol{H} \\ \mathbf{I} & 0 \end{pmatrix}$, and $\boldsymbol{B} = \begin{pmatrix} 0 & \boldsymbol{M}^\top \\ \boldsymbol{M} & 0 \end{pmatrix}$, we have

$$\boldsymbol{z}'(t) = (\boldsymbol{A} \otimes \boldsymbol{B}^\top)\boldsymbol{z}(t) = (\boldsymbol{A} \otimes \boldsymbol{B})\boldsymbol{z}(t),$$

and then $\boldsymbol{z}(t) = e^{(\boldsymbol{A}\otimes\boldsymbol{B})t}\boldsymbol{z}(0)$. Considering the eigendecomposition $\boldsymbol{A} = \boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{U}_{\boldsymbol{A}}^{-1}$ and $\boldsymbol{B} = \boldsymbol{U}_{\boldsymbol{B}}\boldsymbol{\Lambda}_{\boldsymbol{B}}\boldsymbol{U}_{\boldsymbol{B}}^{-1}$, where $\boldsymbol{\Lambda}_{\boldsymbol{A}} = \operatorname{diag}(\lambda_1^{\boldsymbol{A}}, ..., \lambda_n^{\boldsymbol{A}})$ and $\boldsymbol{\Lambda}_{\boldsymbol{B}} = \operatorname{diag}(\lambda_1^{\boldsymbol{B}}, ..., \lambda_n^{\boldsymbol{B}})$, thus

$$\boldsymbol{A} \otimes \boldsymbol{B} = (\boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{\Lambda}_{\boldsymbol{A}}\boldsymbol{U}_{\boldsymbol{A}}^{-1}) \otimes \boldsymbol{B} = (\boldsymbol{U}_{\boldsymbol{A}} \otimes \mathbf{I})(\boldsymbol{\Lambda}_{\boldsymbol{A}} \otimes \boldsymbol{B})(\boldsymbol{U}_{\boldsymbol{A}}^{-1} \otimes \mathbf{I}),$$

where $\otimes$ denotes Kronecker product. Moreover, we have

$$\begin{aligned} \exp[(\boldsymbol{A} \otimes \boldsymbol{B})t] &= (\boldsymbol{U}_{\boldsymbol{A}} \otimes \mathbf{I}) \exp[(\boldsymbol{\Lambda}_{\boldsymbol{A}} \otimes \boldsymbol{B})t](\boldsymbol{U}_{\boldsymbol{A}}^{-1} \otimes \mathbf{I}) \\ &= (\boldsymbol{U}_{\boldsymbol{A}}\boldsymbol{U}_{\boldsymbol{B}} \otimes \mathbf{I}) \exp[(\boldsymbol{\Lambda}_{\boldsymbol{A}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{B}})t](\boldsymbol{U}_{\boldsymbol{B}}^{-1}\boldsymbol{U}_{\boldsymbol{A}}^{-1} \otimes \mathbf{I}). \end{aligned}$$

We know that $\exp[(\boldsymbol{\Lambda}_{\boldsymbol{A}}\otimes\boldsymbol{\Lambda}_{\boldsymbol{B}})t]$ is a diagonal matrix. Therefore, we have $(\boldsymbol{U}_{\boldsymbol{B}}^{-1}\boldsymbol{U}_{\boldsymbol{A}}^{-1}\otimes\mathbf{I})\boldsymbol{z}(t) = \exp[(\boldsymbol{\Lambda}_{\boldsymbol{A}} \otimes \boldsymbol{\Lambda}_{\boldsymbol{B}})t](\boldsymbol{U}_{\boldsymbol{B}}^{-1}\boldsymbol{U}_{\boldsymbol{A}}^{-1} \otimes \mathbf{I})\boldsymbol{z}(0)$, which leads to a concise closed-form dynamics on $(\boldsymbol{U}_{\boldsymbol{B}}^{-1}\boldsymbol{U}_{\boldsymbol{A}}^{-1} \otimes \mathbf{I})\boldsymbol{z}(t)$. ■

B.8 Proof of Theorem 6

Theorem 6 *Consider the continual gradient flow (Equation* (16)*) in which the prototypes* $\boldsymbol{W}$ *is fixed, we have the closed-form dynamics:*

$$\boldsymbol{H}(t) = e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\boldsymbol{H}(0) + \frac{1 - e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}}{\lambda}\boldsymbol{W}\boldsymbol{M}, \tag{96}$$

which further indicates that $\left\|\boldsymbol{H}(t) - \frac{1}{\lambda}\boldsymbol{W}\boldsymbol{M}\right\| = O\left(e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\right)$.

Proof For the first order non-homogeneous linear difference equation in Equation (16), the solution is

$$\begin{aligned} \boldsymbol{H}(t) =& e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\left(\boldsymbol{H}(0) + \int_0^t \eta(s)e^{\lambda \int_0^s \eta(\tau)\mathrm{d}\tau}\mathrm{d}s\boldsymbol{W}\boldsymbol{M}\right) \\ =& e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\boldsymbol{H}(0) + e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\int_0^t \frac{1}{\lambda}\mathrm{d}e^{\lambda \int_0^s \eta(\tau)\mathrm{d}\tau}\boldsymbol{W}\boldsymbol{M} \\ =& e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\boldsymbol{H}(0) + e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\frac{e^{\lambda \int_0^s \eta(\tau)\mathrm{d}\tau}}{\lambda}\Big|_0^t \boldsymbol{W}\boldsymbol{M} \\ =& e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\boldsymbol{H}(0) + \frac{1 - e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}}{\lambda}\boldsymbol{W}\boldsymbol{M}, \end{aligned} \tag{97}$$

and then $\|\boldsymbol{H}(t) - \frac{1}{\lambda}\boldsymbol{W}\boldsymbol{M}\| = \|e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}(\boldsymbol{H}(0) - \frac{\boldsymbol{W}\boldsymbol{M}}{\lambda})\| = O(e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau})$. ■

B.9 The Projections onto $\mathcal{E}_1^+$, $\mathcal{E}_1^-$, $\mathcal{E}_2^+$, $\mathcal{E}_2^-$ and $\mathcal{E}_3$

Lemma 13. *Let $\mathcal{S}$ denote the subspace $\mathcal{S} = \{\boldsymbol{W} : \boldsymbol{W}\mathbf{1}_n = 0, \boldsymbol{W} \in \mathbb{R}^{m\times n}\}$, then the projection of a point $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ onto $\mathcal{S}$ can be denoted as $\Pi_{\mathcal{S}}\boldsymbol{A} = \boldsymbol{A}\left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right)$.*

Proof Let $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, ..., \boldsymbol{w}_n) \in \mathcal{S}$ and $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_n) \in \mathbb{R}^{m\times n}$, we have

$$\begin{aligned}\|\boldsymbol{W} - \boldsymbol{A}\|_F^2 &= \sum_{i=1}^n \|\boldsymbol{w}_i - \boldsymbol{a}_i\|_2^2 \geq \frac{1}{n}\left\|\sum_{i=1}^n \boldsymbol{w}_i - \sum_{i=1}^n \boldsymbol{a}_i\right\|_2^2 \\ &= \frac{1}{n}\|\boldsymbol{W}\mathbf{1}_n - \boldsymbol{A}\mathbf{1}_n\|_2^2 = \frac{1}{n}\|\boldsymbol{A}\mathbf{1}_n\|_2^2\end{aligned}$$

where we used the Cauchy-Schwarz inequality, and the equality holds if and only if $\boldsymbol{w}_i - \boldsymbol{a}_i = \boldsymbol{w}_n - \boldsymbol{a}_n$, $\forall i \in [n]$, and $\sum_{i=1}^n \boldsymbol{w}_i = 0$, *i.e.*, $\boldsymbol{W} = \boldsymbol{A}\left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right)$. Therefore, the projection of $\boldsymbol{A}$ onto $\mathcal{S}$ is $\Pi_{\mathcal{S}}\boldsymbol{A} = \arg\min_{\boldsymbol{W}\in\mathcal{S}} \|\boldsymbol{W} - \boldsymbol{A}\|_F^2 = \boldsymbol{A}\left(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\right)$. ■

Lemma 14. *Let $\mathcal{S}$ denote the subspace $\{\boldsymbol{W} : \boldsymbol{W}(\boldsymbol{I}_c \otimes \mathbf{1}_n) = 0, \boldsymbol{W} \in \mathbb{R}^{m\times cn}\}$, then the projection of a point $\boldsymbol{A} \in \mathbb{R}^{m\times cn}$ onto $\mathcal{S}$ can be denoted as $\Pi_{\mathcal{S}}\boldsymbol{A} = \boldsymbol{A}\left(\boldsymbol{I}_{cn} - \frac{1}{n}\boldsymbol{I}_c \otimes \mathbf{1}_n\mathbf{1}_n^\top\right)$.*

Proof The proof is similar to Lemma 13. We can simply let $\boldsymbol{W} = (\boldsymbol{W}_1, ..., \boldsymbol{W}_n) \in \mathcal{S}$ and $\boldsymbol{A} = (\boldsymbol{A}_1, ..., \boldsymbol{A}_n) \in \mathbb{R}^{m\times cn}$, where $\boldsymbol{W}_i, \boldsymbol{A}_i \in \mathbb{R}^{m\times c}$, then we have

$$\|\boldsymbol{W} - \boldsymbol{A}\|_F^2 = \sum_{i=1}^n \|\boldsymbol{W}_i - \boldsymbol{A}_i\|_F^2 \geq \frac{1}{n}\left\|\sum_{i=1}^n \boldsymbol{W}_i - \sum_{i=1}^n \boldsymbol{A}_i\right\|_2^2 = \frac{1}{n}\|\boldsymbol{A}(\boldsymbol{I}_c \otimes \mathbf{1}_n)\|_2^2$$

where the equality holds if and only if $\boldsymbol{W}_i - \boldsymbol{A}_i = \boldsymbol{W}_n - \boldsymbol{A}_n$, $\forall i \in [n]$, and $\sum_{i=1}^n \boldsymbol{W}_i = 0$, *i.e.*, $\boldsymbol{W} = \boldsymbol{A}(\boldsymbol{I}_{cn} - \frac{1}{n}\boldsymbol{I}_c \otimes \mathbf{1}_n\mathbf{1}_n^\top)$. Therefore, the projection of $\boldsymbol{A}$ onto $\mathcal{S}$ is $\Pi_{\mathcal{S}}\boldsymbol{A} = \boldsymbol{A}\left(\boldsymbol{I}_{cn} - \frac{1}{n}\boldsymbol{I}_c \otimes \mathbf{1}_n\mathbf{1}_n^\top\right)$. ■

Lemma 15. *For $\boldsymbol{H} \in \mathbb{R}^{p\times CN}$, $\boldsymbol{W} \in \mathbb{R}^{p\times C}$, the projection of $(\boldsymbol{H}, \boldsymbol{W})$ onto $\mathcal{E}_1^\epsilon$ is*

$$\Pi_1^\epsilon(\boldsymbol{H}, \boldsymbol{W}) = (\tfrac{\epsilon}{\sqrt{N}}(\boldsymbol{P} \otimes \mathbf{1}_N^\top), \boldsymbol{P}), \tag{98}$$

where $\boldsymbol{P} = \frac{1}{2}(\frac{\epsilon}{\sqrt{N}}\boldsymbol{H}(\boldsymbol{I}_C \otimes \mathbf{1}_N) + \boldsymbol{W})(\boldsymbol{I}_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top)$.

Proof Let $\mathcal{S} = \{\boldsymbol{Z} : \boldsymbol{Z}\mathbf{1}_C = 0, \boldsymbol{Z} \in \mathbb{R}^{p\times C}\}$ and $\boldsymbol{H} = \{\boldsymbol{H}_1, ..., \boldsymbol{H}_N\}$ (where $\boldsymbol{H}_i \in \mathbb{R}^{p\times C}$), the minimizer of $\boldsymbol{Z} \in \mathcal{S}$ is

$$\begin{aligned}
&\arg\min_{\boldsymbol{Z}\in\mathcal{S}} \|\tfrac{\epsilon}{\sqrt{N}}(\boldsymbol{Z}\otimes\mathbf{1}_N^\top) - \boldsymbol{H}\|_F^2 + \|\boldsymbol{Z}-\boldsymbol{W}\|_F^2 \\
=&\arg\min_{\boldsymbol{Z}\in\mathcal{S}} \sum_{i=1}^N \|\tfrac{1}{\sqrt{N}}\boldsymbol{Z} - \epsilon\boldsymbol{H}_i\|_F^2 + \|\boldsymbol{Z}-\boldsymbol{W}\|_F^2 \\
=&\arg\min_{\boldsymbol{Z}\in\mathcal{S}} \|\boldsymbol{Z}\|_F^2 - \tfrac{2\epsilon}{\sqrt{N}}\sum_{i=1}^N\langle\boldsymbol{Z},\boldsymbol{H}_i\rangle + \|\boldsymbol{H}\|_F^2 + \|\boldsymbol{Z}-\boldsymbol{W}\|_F^2 \\
=&\arg\min_{\boldsymbol{Z}\in\mathcal{S}} \|\boldsymbol{Z} - \tfrac{\epsilon}{\sqrt{N}}\sum_{i=1}^N\boldsymbol{H}_i\|_F^2 - \|\frac{1}{\sqrt{N}}\sum_{i=1}^N\boldsymbol{H}_i\|_F^2 + \|\boldsymbol{H}\|_F^2 + \|\boldsymbol{Z}-\boldsymbol{W}\|_F^2 \\
=&\arg\min_{\boldsymbol{Z}\in\mathcal{S}} \|\boldsymbol{Z} - \tfrac{\epsilon}{\sqrt{N}}\boldsymbol{H}(\boldsymbol{I}_C\otimes\mathbf{1}_N)\|_F^2 + \|\boldsymbol{Z}-\boldsymbol{W}\|_F^2 + \|\boldsymbol{H}\|_F^2 - \|\tfrac{\epsilon}{\sqrt{N}}\boldsymbol{H}(\boldsymbol{I}_C\otimes\mathbf{1}_N)\|_F^2 \\
=&\arg\min_{\boldsymbol{Z}\in\mathcal{S}} \|\boldsymbol{Z} - \tfrac{1}{2}(\tfrac{\epsilon}{\sqrt{N}}\boldsymbol{H}(\boldsymbol{I}_C\otimes\mathbf{1}_N) + \boldsymbol{W})\|_F^2 \\
=&\tfrac{1}{2}(\tfrac{\epsilon}{\sqrt{N}}\boldsymbol{H}(\boldsymbol{I}_C\otimes\mathbf{1}_N) + \boldsymbol{W})(\boldsymbol{I}_C - \tfrac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top).
\end{aligned} \tag{99}$$

Thus, $\Pi_1^\epsilon(\boldsymbol{H},\boldsymbol{W}) = (\frac{\epsilon}{\sqrt{N}}(\boldsymbol{P}\otimes\mathbf{1}_N^\top), \boldsymbol{P})$ with $\boldsymbol{P} = \frac{1}{2}(\frac{\epsilon}{\sqrt{N}}\boldsymbol{H}(\boldsymbol{I}_C\otimes\mathbf{1}_N) + \boldsymbol{W})(\boldsymbol{I}_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top)$. ■

Lemma 16. *For $\boldsymbol{H} \in \mathbb{R}^{p\times CN}$, $\boldsymbol{W} \in \mathbb{R}^{p\times C}$, the projection of $(\boldsymbol{H},\boldsymbol{W})$ onto $\mathcal{E}_2^\epsilon$ is*

$$\Pi_2^\epsilon(\boldsymbol{H},\boldsymbol{W}) = (\tfrac{\epsilon}{\sqrt{N}}\boldsymbol{h}\mathbf{1}_{CN}^\top, \boldsymbol{h}\mathbf{1}_C^\top), \tag{100}$$

where $\boldsymbol{h} = \frac{1}{2C}(\frac{\epsilon}{\sqrt{N}}\boldsymbol{H}\mathbf{1}_{CN} + \boldsymbol{W}\mathbf{1}_C)$.

Proof We have

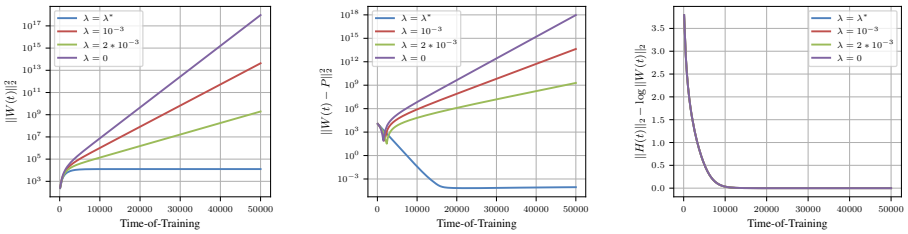
$$\begin{aligned}
&\arg\min_{\boldsymbol{h}\in\mathbb{R}^p} \|\tfrac{\epsilon}{\sqrt{N}}\boldsymbol{h}\mathbf{1}_{CN}^\top - \boldsymbol{H}\|_F^2 + \|\boldsymbol{h}\mathbf{1}_C^\top - \boldsymbol{W}\|_F^2 \\
=&\arg\min_{\boldsymbol{h}\in\mathbb{R}^p} CN\|\tfrac{1}{\sqrt{N}}\boldsymbol{h} - \tfrac{\epsilon}{CN}\boldsymbol{H}\mathbf{1}_{CN}\|_2^2 + C\|\boldsymbol{h} - \tfrac{1}{C}\boldsymbol{W}\mathbf{1}_C\|_2^2 \\
=&\arg\min_{\boldsymbol{h}\in\mathbb{R}^p} \|\boldsymbol{h} - \tfrac{\epsilon}{C\sqrt{N}}\boldsymbol{H}\mathbf{1}_{CN}\|_2^2 + \|\boldsymbol{h} - \tfrac{1}{C}\boldsymbol{W}\mathbf{1}_C\|_2^2 \\
=&\arg\min_{\boldsymbol{h}\in\mathbb{R}^p} \|\boldsymbol{h} - \tfrac{1}{2C}(\tfrac{\epsilon}{\sqrt{N}}\boldsymbol{H}\mathbf{1}_{CN} + \boldsymbol{W}\mathbf{1}_C)\|_2^2 \\
=&\tfrac{1}{2C}(\tfrac{\epsilon}{\sqrt{N}}\boldsymbol{H}\mathbf{1}_{CN} + \boldsymbol{W}\mathbf{1}_C)
\end{aligned} \tag{101}$$

■

Lemma 17. *For $\boldsymbol{H} \in \mathbb{R}^{p\times CN}$, $\boldsymbol{W} \in \mathbb{R}^{p\times C}$, the projection of $(\boldsymbol{H},\boldsymbol{W})$ onto $\mathcal{E}_3$ is*

$$\Pi_3(\boldsymbol{H},\boldsymbol{W}) = (\boldsymbol{H}(\boldsymbol{I}_{CN} - \tfrac{1}{N}\boldsymbol{I}_C\otimes\mathbf{1}_N\mathbf{1}_N^\top), 0). \tag{102}$$

Proof This can be easily derived by Lemma 14. ■



(a) $\log \|\boldsymbol{W}(t)\|_2$ (b) $\log \|\boldsymbol{W}(t) - \boldsymbol{P}\|_2$ (c) $\log \|\boldsymbol{H}(t)\|_2 - \log \|\boldsymbol{W}(t)\|_2$

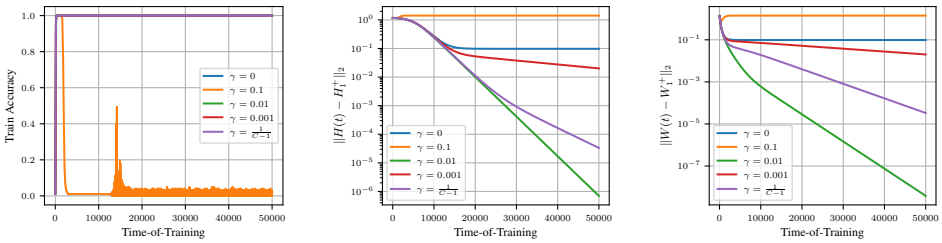
Figure 4: Behavior of gradient descent iterates under the unhinged loss with different weight decay coefficients and $\eta_1(t) = \eta_2(t) = 0.1$ (*i.e.*, $s = 1$), where $\boldsymbol{P}$ denotes the component $\boldsymbol{W}_1^+$ in the projection $\Pi_1^+ \boldsymbol{Z}_0$ calculated according to Lemma 15. (a) The logarithm of the norm of $\boldsymbol{W}(t)$. As expected, the norm increases exponentially when $\lambda < \lambda^*$; (b) The difference between $\boldsymbol{W}(t)$ and $\boldsymbol{P}$. As expected (Corollary 5), $\boldsymbol{W}(t)$ converges to $\boldsymbol{P}$ when $\lambda = \lambda^*$, while other differences are dominated by $\|\boldsymbol{W}(t)\|_2$; (c) The difference in ℓ_2 norm between $\boldsymbol{H}(t)$ and $\boldsymbol{W}(t)$. The convergence is the same even if the weight decay is different.

Table 4: Test accuracies on imbalanced CIFAR-10 under different explicit feature regularization.

Dataset	Imbalanced CIFAR-10							
Imbalance Type	long-tailed				step			
Imbalance Ratio	100	50	20	10	100	50	20	10
baseline	67.81	72.93	83.97	88.37	61.24	68.10	78.73	85.49
$\lambda = 5e-6$	**67.84**	72.85	83.17	**89.06**	60.79	**68.41**	**80.20**	**86.69**
$\lambda = 1e-5$	67.74	**76.14**	**84.17**	**89.19**	**61.50**	67.71	**80.97**	**87.18**
$\lambda = 5e-5$	**69.74**	**77.29**	**84.92**	**88.64**	60.69	**70.27**	**81.27**	**87.17**

Table 5: Test accuracies on imbalanced CIFAR-100 under different explicit feature regularization.

Dataset	Imbalanced CIFAR-100							
Imbalance Type	long-tailed				step			
Imbalance Ratio	100	50	20	10	100	50	20	10
baseline	33.37	39.40	42.96	56.38	40.89	42.69	51.92	57.52
$\lambda = 5e-6$	**36.00**	**41.92**	**50.75**	**60.13**	**41.90**	**43.85**	47.80	56.74
$\lambda = 1e-5$	**36.61**	**42.36**	**49.21**	**58.91**	**41.48**	**43.77**	49.64	56.49
$\lambda = 5e-5$	**34.88**	**42.74**	**54.72**	**60.84**	**40.97**	**43.20**	48.96	**57.97**



(a) Train Accuracy (b) $\|\boldsymbol{H}(t) - \overline{\boldsymbol{H}}\|_2$ (c) $\|\boldsymbol{W}(t) - \overline{\boldsymbol{W}}\|_2$

(d) $\|\boldsymbol{H}(t)\|_2^2$ (e) $\|\boldsymbol{W}(t)\|_2^2$

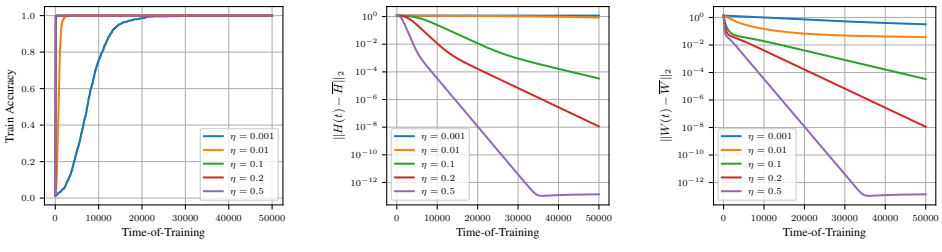
Figure 5: Verification of the behavior of regularized gradient descent iterates in Equation (12) with $\gamma \in \{0, 0.1, 0.01, 0.001, \frac{1}{C-1}\}$. We set $p = 512$, $C = 100$, $N = 10$, $\lambda = \frac{(1+\gamma)}{C\sqrt{N}}$, $\eta_1(t) = \eta_2(t) = 0.1$ (*i.e.*, $s = 1$), thus we have $\lim_{t\to\infty} \boldsymbol{Z}(t) = \Pi_1^+ \boldsymbol{Z}_0$, according to Corollary 5, and then randomly initialize $\boldsymbol{H}_0$ and $\boldsymbol{W}_0$. (a) The training accuracy with the prediction rule $\arg\max_c \boldsymbol{w}_c^\top \boldsymbol{h}$. As expected, the features align to their corresponding prototypes when $\gamma < \frac{2}{C-2}$. (b) The ℓ_2 distance between $\boldsymbol{H}(t)$ and $\boldsymbol{H}_1^+$. As expected Theorem 5, the distance will decrease as exponential rate when $0 < \gamma < \frac{2}{C-2}$. (c) The ℓ_2 distance between $\boldsymbol{W}(t)$ and $\boldsymbol{W}_1^+$. (d) and (e) denote the norm of features and prototypes, respectively. As can be seen, $\|\boldsymbol{H}\|_2$ and $\|\boldsymbol{W}\|_2$ do not grow exponentially as in the unconstrained case, which confirms that weight decay can avoid excessive growth of feature norm and prototype norm.

Appendix C. Experiments

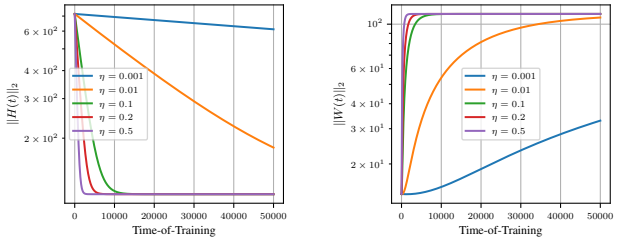
In this section, we provide experimental details, including datasets, network architectures, optimization methods, hyperparameter settings, and more results.

C.1 Numerical Experiments

For numerical experiments in Figures 1, 4, 5, 6, 7, 8, and 9, we set $p = 512$, $C = 100$, $N = 10$, and then randomly initialize $\boldsymbol{H}_0$ and $\boldsymbol{W}_0$. We use the SGD optimizer to optimize these free variables.



(a) Train Accuracy (b) $\|\boldsymbol{H}(t) - \overline{\boldsymbol{H}}\|_2$ (c) $\|\boldsymbol{W}(t) - \overline{\boldsymbol{W}}\|_2$



(d) $\|\boldsymbol{H}(t)\|_2$ (e) $\|\boldsymbol{W}(t)\|_2$

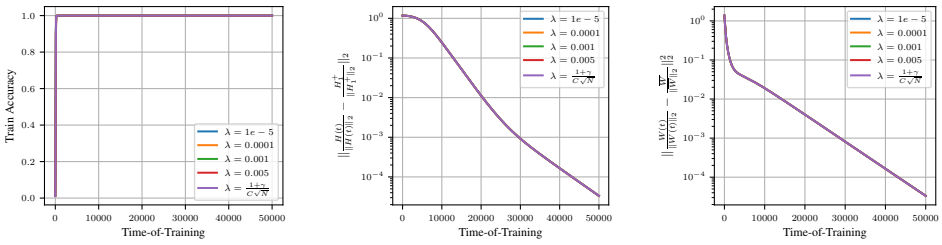
Figure 6: Verification of the behavior of regularized gradient descent iterates in Equation (12) with different learning rates ($\eta \in \{0.001, 0.01, 0.1, 0.2, 0.5\}$). We set $p = 512$, $C = 100$, $N = 10$, $\eta_1 = \eta_2 = \eta$ ($s = 1$), $\gamma = \frac{1}{C-1}$, and $\lambda = \frac{1+\gamma}{C\sqrt{N}}$. As can be seen, features and prototypes converge to $(\overline{\boldsymbol{H}}, \overline{\boldsymbol{W}})$ exponentially, and larger learning rates can accelerate convergence.

C.2 Visual Classification

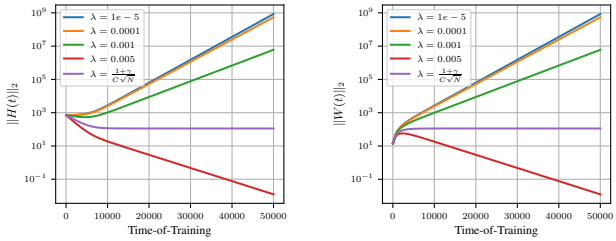
For classification experiments in Figure 2, Figure 12, Figure 13, Figure 14, we experiment with ResNet-18, ResNet-34, and ResNet-50 (He et al., 2016) trained on CIFAR-10, CIFAR-100 (Krizhevsky and Hinton, 2009), and ImageNet-100 that takes the first 100 classes of ImageNet (Deng et al., 2009) , respectively. The networks are trained for 200 epochs and 100 epochs for CIFAR-10/-100 and ImageNet-100, respectively. For all training, we use SGD optimizer with momentum 0.9 and cosine learning rate annealing Loshchilov and Hutter (2017) with $T_{\max}$ being the corresponding epochs. The initial learning rate is set to 0.1, weight decay is set to 5×10^{-4}, and batch size is set to 256. Typical data augmentations including random width/height shift and horizontal flip are applied. Moreover, to use the PAL and FNPAL (Zhou et al., 2022c) that anchors prototypes with a neural collapse solution, we remove the ReLU layer before the linear classifier in the last layer.

C.3 Imbalanced Classification

For the experiments of imbalanced learning in Table 2, Table 4, Table 5, Figure 10, and Figure 11, we utilize the same network architectures, and optimization settings as visual classification. We only use the imbalanced versions of CIFAR-10 and CIFAR-100 by following the setting in (Zhou et al., 2022d). The number of training examples is reduced for per class, and the test set keeps unchanged, where we use the imbalance ratio $\rho = \frac{\max_i n_i}{\min_i n_i}$ to



(a) Train accuracy (b) $\|\hat{\boldsymbol{H}}(t) - \hat{\boldsymbol{H}}_1^+\|_2$ (c) $\|\hat{\boldsymbol{W}}(t) - \hat{\boldsymbol{W}}_1^+\|_2$



(d) $\|\boldsymbol{H}(t)\|_2$ (e) $\|\boldsymbol{W}(t)\|_2$

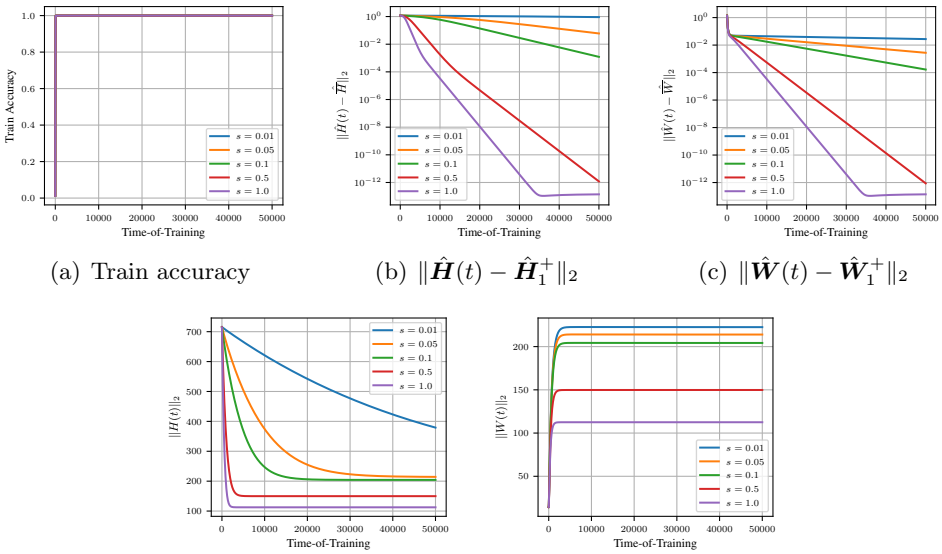
Figure 7: Verification of the behavior of regularized gradient descent iterates in Equation (15) with different weight decay coefficients ($\lambda = \{1e-5, 1e-4, 1e-3, 5e-3, \frac{1+\gamma}{C\sqrt{N}}\}$). We set $p = 512$, $C = 100$, $N = 10$, $\eta_1(t) = \eta_2(t) = 0.1$ (*i.e.*, $s = 1$), $\gamma = \frac{1}{C-1}$, where $\overline{\boldsymbol{H}} = \pi_h^+ \boldsymbol{H}_1^+ + \pi_h^- \boldsymbol{H}_1^-$ and $\overline{\boldsymbol{W}} = \pi_w^+ \boldsymbol{W}_1^+ + \pi_w^- \boldsymbol{W}_1^-$ in Corollary 5. (a) The logarithm of the norm of $\boldsymbol{W}(t)$. As expected, the norm increases exponentially when $\lambda < \lambda^* = \frac{1+\gamma}{C\sqrt{N}}$; (b) The difference between $\boldsymbol{W}(t)$ and $\boldsymbol{P}$. As expected in Corollary 5, $\boldsymbol{W}(t)$ converges to $\boldsymbol{P}$ when $\lambda = \lambda^*$, while other differences are dominated by $\|\boldsymbol{W}(t)\|_2$; (c) The difference in ℓ_2 norm between $\boldsymbol{H}(t)$ and $\boldsymbol{W}(t)$. The convergence is the same even if the weight decay is different.

denote the ratio between sample sizes of the most frequent and least frequent class. Moreover, long-tailed imbalance (Cui et al., 2019) that utilizes an exponential decay in samples sizes and step imbalance (Buda et al., 2018)(that sets all minority classes to have the same number of samples, as do all majority classes) are considered.

For imbalanced learning, we utilize expected calibration error (ECE) to measure calibration of the models (Zhong et al., 2021), where all predictions are grouped into several interval bins of equal size and then calculate the error between the accuracy and confidence for each interval bin, *i.e*,

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|\mathcal{S}_b|}{N} \left|\mathrm{acc}(\mathcal{S}_b) - \mathrm{conf}(\mathcal{S}_b)\right| \times 100\%, \tag{103}$$

where N denotes the number of predictions, B is the number of interval bins, $\mathcal{S}_b$ is the set of samples whose prediction scores fall into Bin-b, $\mathrm{acc}(\cdot)$ and $\mathrm{conf}(\cdot)$ denote the accuracy and predicted confidence of $\mathcal{S}_b$, respectively.



(a) Train accuracy (b) $\|\hat{\boldsymbol{H}}(t) - \hat{\boldsymbol{H}}_1^+\|_2$ (c) $\|\hat{\boldsymbol{W}}(t) - \hat{\boldsymbol{W}}_1^+\|_2$

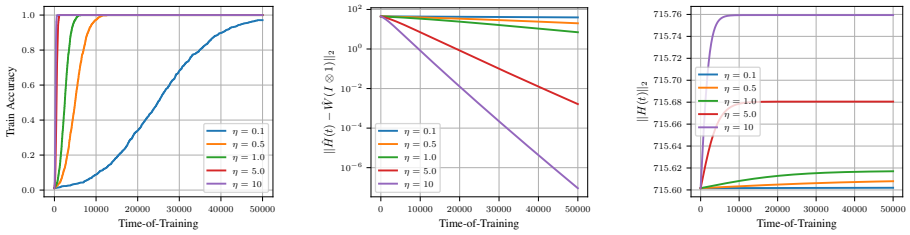
(d) $\|\boldsymbol{H}(t)\|_2$ (e) $\|\boldsymbol{W}(t)\|_2$

Figure 8: Verification of the behavior of regularized gradient descent iterates in Equation (15) with different scale parameters $s = \frac{\eta_1(0)}{\eta_2(0)} \in \{0.01, 0.05, 0.1, 0.5, 1.0\}$. We set $p = 512$, $C = 100$, $N = 10$, $\eta_2 = 0.5$, $\gamma = \frac{1}{C-1}$, where $\overline{\boldsymbol{H}} = \pi_h^+ \boldsymbol{H}_1^+ + \pi_h^- \boldsymbol{H}_1^-$ and $\overline{\boldsymbol{W}} = \pi_w^+ \boldsymbol{W}_1^+ + \pi_w^- \boldsymbol{W}_1^-$ in Corollary 5. As can be seen, larger scale parameter s can achieve faster convergence speed.

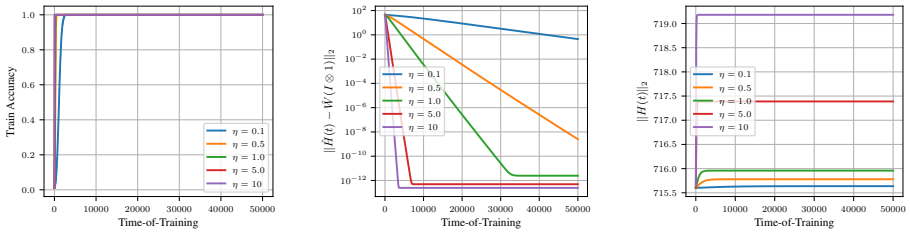
As shown in Table 4 and Table 5, explicit feature regularization can improve imbalanced learning on CIFAR-10/-100 in most cases.

C.4 Out-of-Distribution Detection

For the experiments of OOD detection in Figure 3, Table 6, Figure 16, Figure 17, and Figure 18, we use a ResNet-18 on CIFAR-10 and a ResNet-34 on CIFAR-100 to train the classification models, and use their test dataset as the in-distribution data $\mathcal{D}_{\text{in}}^{\text{test}}$. For the OOD test dataset $\mathcal{D}_{\text{out}}^{\text{test}}$, we simply use a common benchmark: SVHN (Netzer et al., 2011). We measure the performance with the following metrics: (1) the false positive rate (FPR95) of OOD examples when true positive rate of in-distribution examples is at 95%; (2) the area under the receiver operating characteristic curve (AUROC); and (3) the area under the precision-recall curve (AUPR). We then consider the softmax-based score (Hendrycks and Gimpel, 2016), energy-based score (Liu et al., 2020), and our proposed feature norm-based score to assessing the improvement of explicit feature regularization over the normal training.



(a) Train Accuracy (b) $\|\hat{\boldsymbol{H}}(t) - \hat{\boldsymbol{W}}(\boldsymbol{I}_C \otimes \mathbf{1}_N)\|_2$ (c) $\|\boldsymbol{H}(t)\|_2$



(d) Train Accuracy (e) $\|\hat{\boldsymbol{H}}(t) - \hat{\boldsymbol{W}}(\boldsymbol{I}_C \otimes \mathbf{1}_N)\|_2$ (f) $\|\boldsymbol{H}(t)\|_2$

Figure 9: Verification of the behavior of discrete gradient descent iterates in Equation (18) under anchored prototypes with different learning rates $\eta \in \{0.1, 0.5, 1.0, 5.0, 10\}$ and without (a-c) or with (d-f) rescaled learning rates. We set $p = 512$, $C = 100$, and $N = 10$. As expected in Theorem 7, the feature norm $\|\boldsymbol{H}(t)\|_2$ is non-decreasing, and the error $\|\hat{\boldsymbol{H}}(t) - \hat{\boldsymbol{W}}(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top)\|_2$ shows exponential decrease.

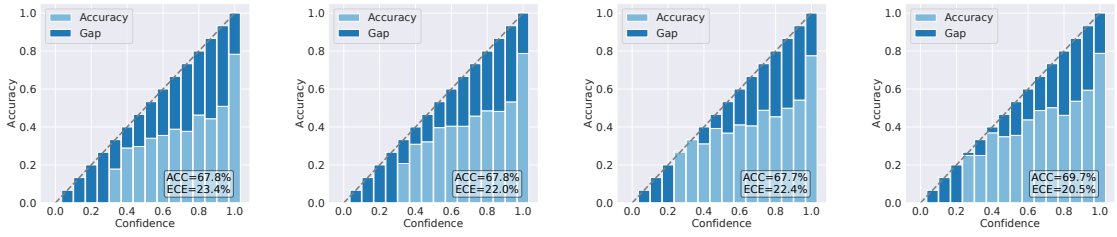
Appendix D. Other Potential Insights

D.1 A Good Initialization of Prototypes

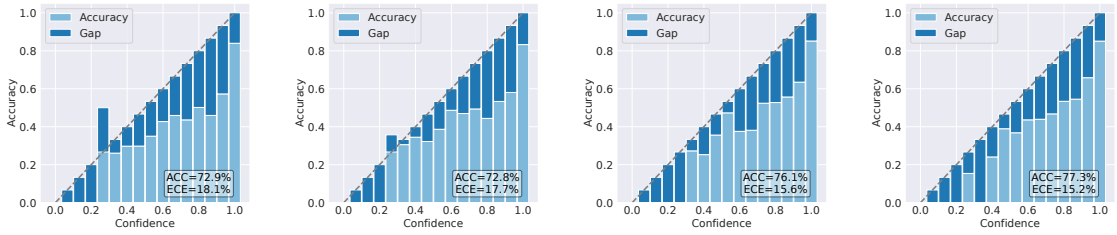
As depicted in Section 3 and B.9, the dynamics under the unhinged loss is dependent on the initialization of both features and prototypes, such as $\Pi_1^+ \boldsymbol{Z}_0 = (\frac{1}{\sqrt{N}}(\boldsymbol{P} \otimes \mathbf{1}_N^\top), \boldsymbol{P})$, where $\boldsymbol{P} = \frac{1}{2}\left(\frac{1}{\sqrt{N}}\boldsymbol{H}_0(\boldsymbol{I}_C \otimes \mathbf{1}_N) + \boldsymbol{W}_0\right)(\boldsymbol{I}_C - \frac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top)$. However, these features $\boldsymbol{H}_0$ extracted from a dataset by some nonlinear layers and parameterized layers are practically intractable, but we can elaborately initialize $\boldsymbol{W}_0$ and highlight its role in the whole. To do this, we consider two ways: (1) Initializing the structure of $\boldsymbol{W}_0$. Inspired by the neural collapse solution that maximizes class separation, we can initialize $\boldsymbol{W}_0$ as this structure, *i.e.*, $\hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_j = \frac{-1}{C-1}, \forall i \neq j$; (2) Increasing the importance of $\boldsymbol{W}_0$. A simple strategy is scaling up $\boldsymbol{W}_0$, thereby implicitly weakening the importance of $\boldsymbol{H}_0$. However, it is difficult to handle the initialization of features because they are obtained by a complex processing a large dataset, thus we seek to initialize the prototypes in the last layer of the network.

D.2 Refined Decision-makings

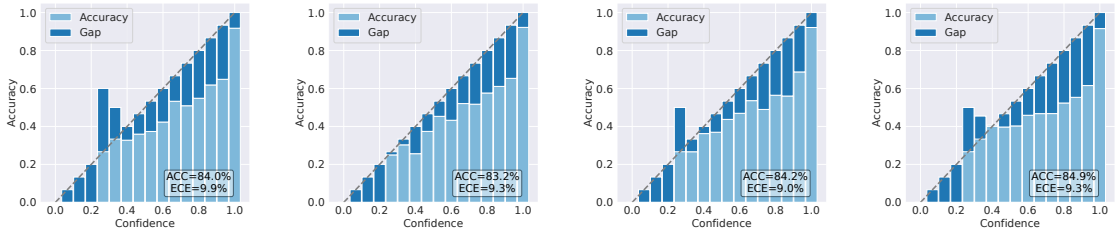
Recalling the rule—$\arg\max_{c'} \langle \boldsymbol{w}_{c'}, \boldsymbol{h} \rangle + b_{c'}$ that makes decision by selecting the class with the largest logit (where the inner product $\langle \boldsymbol{w}_{c'}, \boldsymbol{h} \rangle$ is dominant), which may not be good to directly use the learned features and prototypes, since learning with the unhinged loss



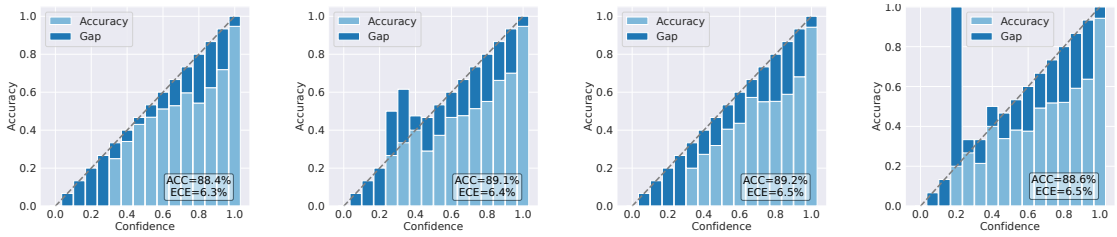
(a) CE ($\lambda = 0$) (b) CE ($\lambda = 5e-6$) (c) CE ($\lambda = 1e-5$) (d) CE ($\lambda = 5e-5$)



(e) CE ($\lambda = 0$) (f) CE ($\lambda = 5e-6$) (g) CE ($\lambda = 1e-5$) (h) CE ($\lambda = 5e-5$)

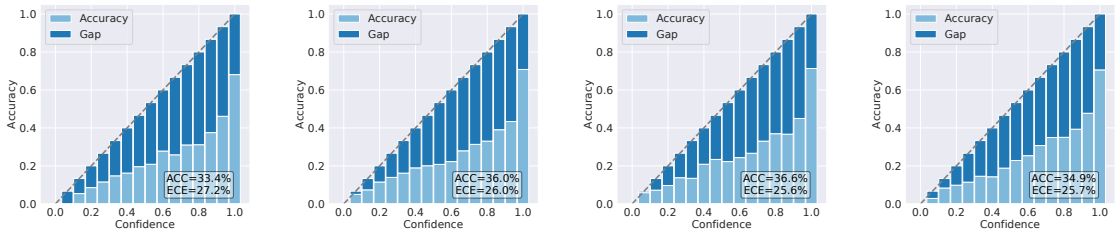


(i) CE ($\lambda = 0$) (j) CE ($\lambda = 5e-6$) (k) CE ($\lambda = 1e-5$) (l) CE ($\lambda = 5e-5$)

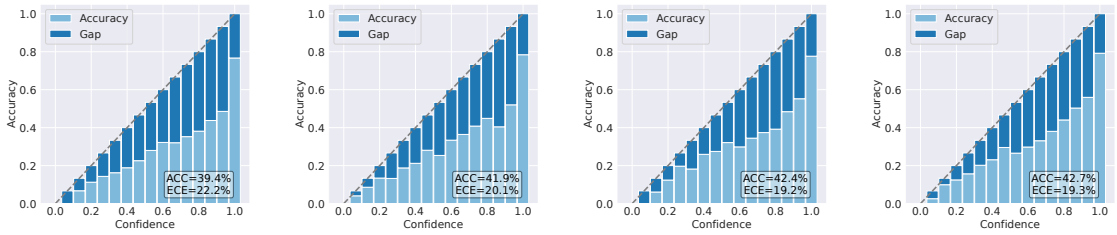


(m) CE ($\lambda = 0$) (n) CE ($\lambda = 5e-6$) (o) CE ($\lambda = 1e-5$) (p) CE ($\lambda = 5e-5$)

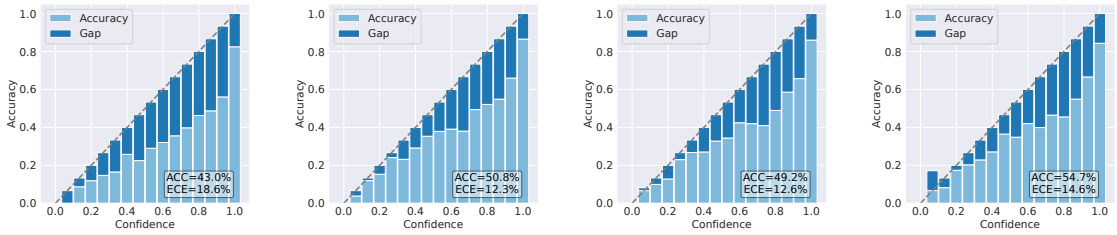
Figure 10: Reliability diagrams of ResNet-34 (He et al., 2016) trained by CE on CIFAR-10-LT with imbalance ratio $\rho \in \{100, 50, 20, 10\}$ under different explicit feature regularization ($\lambda \in \{0, 5e-6, 1e-5, 5e-5\}$). As can be seen, an appropriate larger weight decay can improve both accuracy and confidence



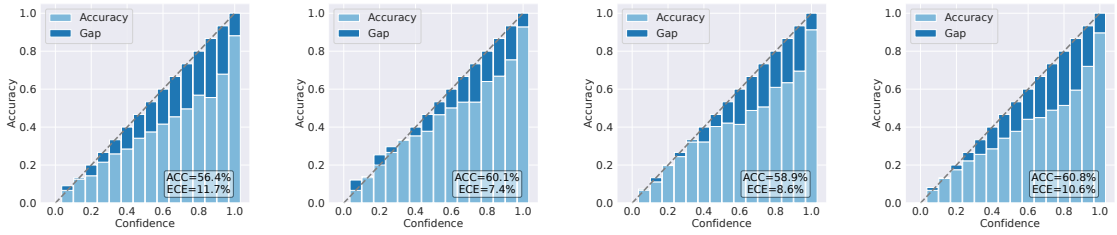
(a) CE ($\lambda = 0$) (b) CE ($\lambda = 5e-6$) (c) CE ($\lambda = 1e-5$) (d) CE ($\lambda = 5e-5$)



(e) CE ($\lambda = 0$) (f) CE ($\lambda = 5e-6$) (g) CE ($\lambda = 1e-5$) (h) CE ($\lambda = 5e-5$)

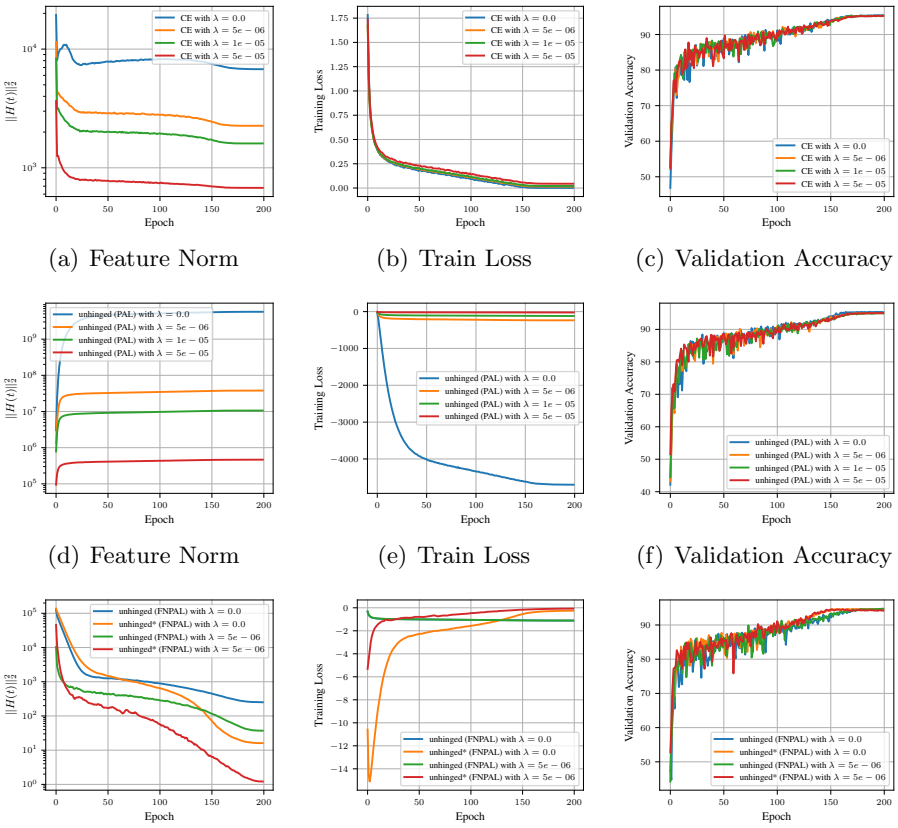


(i) CE ($\lambda = 0$) (j) CE ($\lambda = 5e-6$) (k) CE ($\lambda = 1e-5$) (l) CE ($\lambda = 5e-5$)



(m) CE ($\lambda = 0$) (n) CE ($\lambda = 5e-6$) (o) CE ($\lambda = 1e-5$) (p) CE ($\lambda = 5e-5$)

Figure 11: Reliability diagrams of ResNet-34 (He et al., 2016) trained by CE on CIFAR-100-LT with imbalance ratio $\rho \in \{100, 50, 20, 10\}$ under different explicit feature regularization ($\lambda \in \{0.0, 5e-6, 1e-5, 5e-5\}$), where ECE denotes the expected calibration error (Zhong et al., 2021). As can be seen, an appropriate larger weight decay can improve both accuracy and confidence calibration.

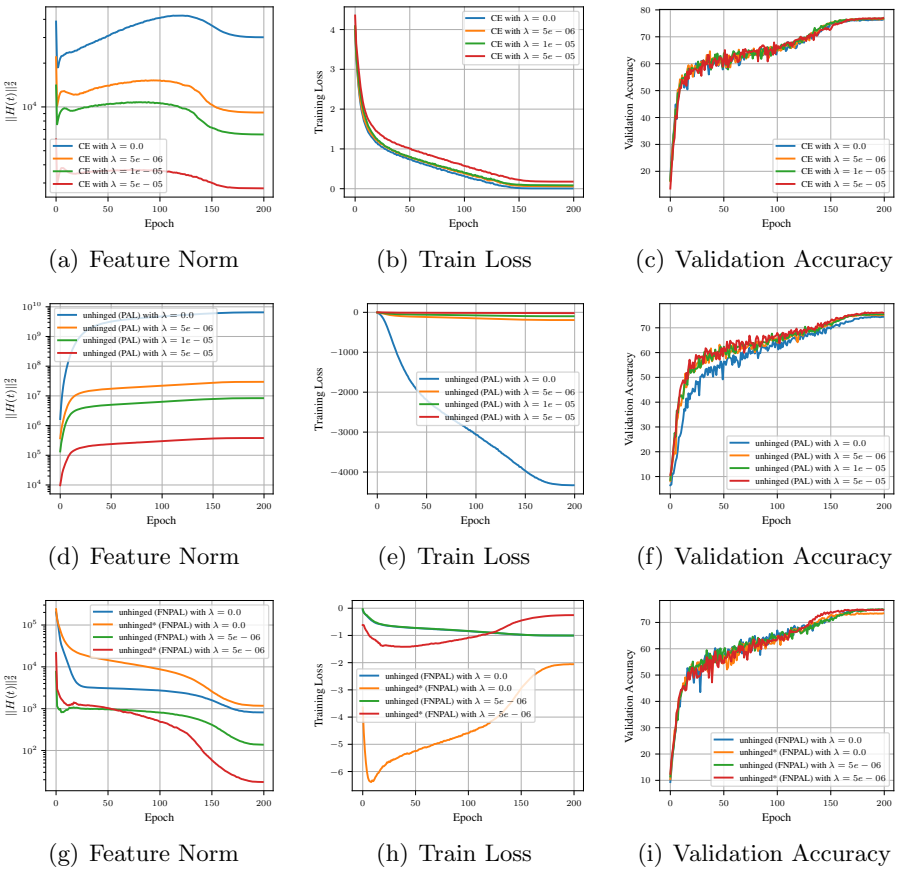


(a) Feature Norm (b) Train Loss (c) Validation Accuracy

(d) Feature Norm (e) Train Loss (f) Validation Accuracy

(g) Feature Norm (h) Train Loss (i) Validation Accuracy

Figure 12: Behavior of visual classification on CIFAR-10 with CE, the unhinged loss (PAL) and the unhinged loss (FNPAL) under different weight decay coefficients.

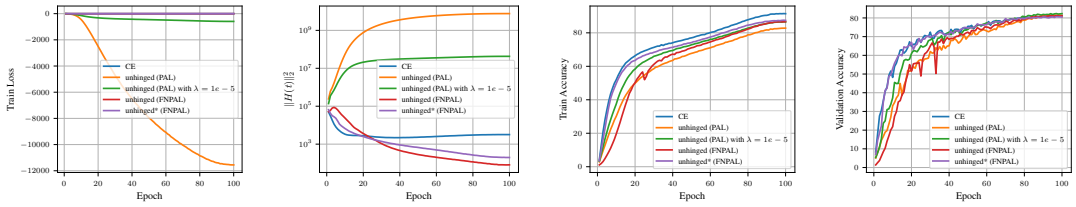


(a) Feature Norm (b) Train Loss (c) Validation Accuracy

(d) Feature Norm (e) Train Loss (f) Validation Accuracy

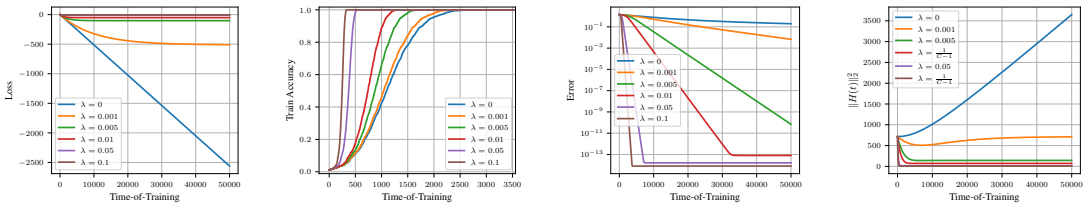
(g) Feature Norm (h) Train Loss (i) Validation Accuracy

Figure 13: Behavior of visual classification on CIFAR-100 with CE, the unhinged loss (PAL) and the unhinged loss (FNPAL) under different weight decay coefficients.



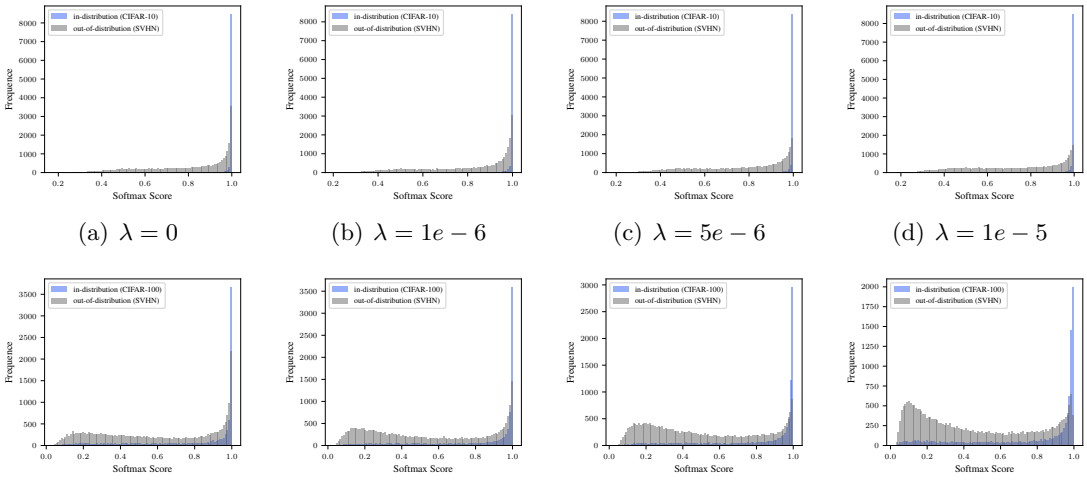
(a) Train Loss (b) Feature Norm (c) Train Accuracy (d) Validation Accuracy

Figure 14: Behavior of visual classification on ImageNet-100 with CE, the unhinged loss (PAL) and the unhinged loss (FNPAL) under different weight decay coefficients.



(a) Train Loss (b) Train Accuracy (c) $\|\hat{\boldsymbol{H}}(t) - \boldsymbol{W}\hat{\boldsymbol{M}}\|_2$ (d) Feature Norm

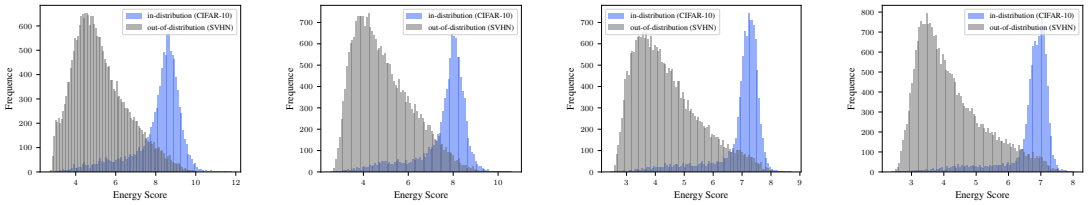
Figure 15: Behavior of gradient descent iterates of the unhinged (PAL) loss in Theorem 6 with different explicit feature regularization ($\lambda \in \{0, 0.001, 0.005, 0.01, 0.05, 0.1\}$). We set $p = 512$, $C = 100$, $N = 10$, and $\eta = 0.1$. We randomly initialize $\boldsymbol{H}_0$ and $\boldsymbol{W}$, and then anchor prototypes $\boldsymbol{W}$ during training. As expected in Theorem 6, the error $\|\hat{\boldsymbol{H}}(t) - \boldsymbol{W}\hat{\boldsymbol{M}}\|_2$ decreases as an exponential rate $O(e^{-\lambda\eta t})$, and a larger λ can accelerate the convergence.



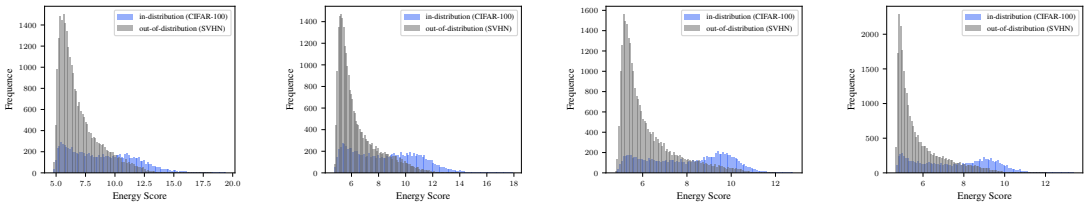
(a) $\lambda = 0$ (b) $\lambda = 1e-6$ (c) $\lambda = 5e-6$ (d) $\lambda = 1e-5$

(e) $\lambda = 0.0$ (f) $\lambda = 1e-6$ (g) $\lambda = 5e-6$ (h) $\lambda = 1e-5$

Figure 16: Distribution of softmax scores (Hendrycks and Gimpel, 2016) from models trained with different explicit feature regularization, where CE is the loss function.

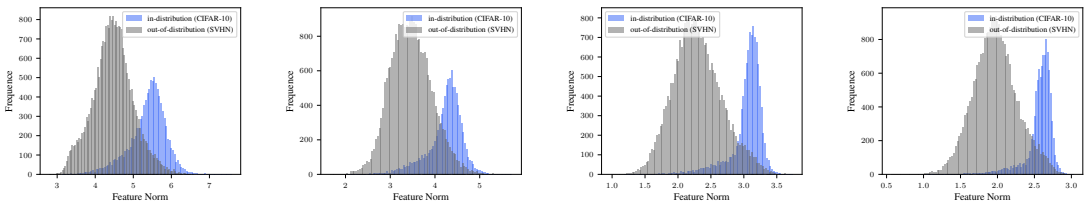


(a) $\lambda = 0$ (b) $\lambda = 1e-6$ (c) $\lambda = 5e-6$ (d) $\lambda = 1e-5$

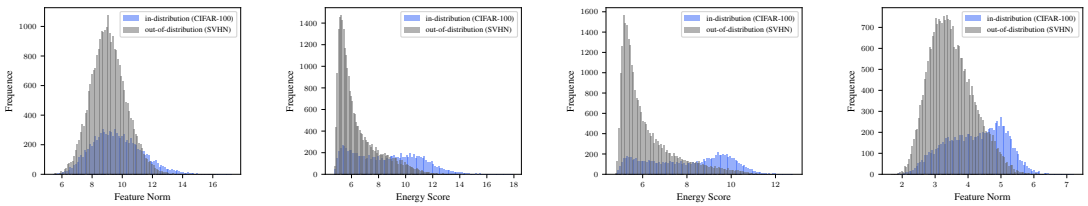


(e) $\lambda = 0$ (f) $\lambda = 1e-6$ (g) $\lambda = 5e-6$ (h) $\lambda = 1e-5$

Figure 17: Distribution of energy scores (Liu et al., 2020) from models trained with different explicit feature regularization, where CE is the loss function.

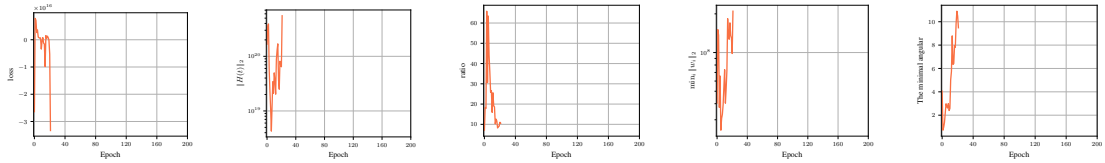


(a) $\lambda = 0$ (b) $\lambda = 1e-6$ (c) $\lambda = 5e-6$ (d) $\lambda = 1e-5$



(e) $\lambda = 0$ (f) $\lambda = 1e-6$ (g) $\lambda = 5e-6$ (h) $\lambda = 1e-5$

Figure 18: Distribution of feature norms from models trained with different explicit feature regularization, where CE is the loss function.



(a) (b) (c) (d) (e)

Figure 19: The behavior of features and prototypes when directly training ResNet-18 with the unhinged loss in Equation (1) on CIFAR-10. We set the weight decay coefficient as $5e-4$. (a) The train accuracy. (b) The feature norm. (c) the ratio $\frac{\max_i \|\boldsymbol{w}_i\|_2}{\min_i \|\boldsymbol{w}_i\|_2}$. (d) $\min_i \|\boldsymbol{w}_i\|_2$. (e) The minimal angular between prototypes: $\arccos \max_{i \neq j} \hat{\boldsymbol{w}}_i^\top \hat{\boldsymbol{w}}_j$. In these figures, we only show the curves for the first 21 epochs, since "NaN" appears at the 22-th epoch. We can find that implicit penalization attached by other components (e.g., network architectures and weight decays) does not limit the rapid growth of the feature norm and prototype norm, indicating implicit penalization is fragile. Moreover, the ratio $\frac{\max_i \|\boldsymbol{w}_i\|_2}{\min_i \|\boldsymbol{w}_i\|_2}$ starts our very large and the minimal angular is very small, which indicates that there are two prototypes that are particular imbalanced.

Table 6: OOD detection performance using softmax-based (Hendrycks and Gimpel, 2016), energy-based (Liu et al., 2020), and feature norm-based approaches while model training with feature regularization ($\lambda = \{0, 1e-6, 5e-6, 1e-5\}$). We use ResNet-18 and ResNet-34 to train on the in-distribution datasets CIFAR-10 and CIFAR-100, respectively. We then use SVHN (Netzer et al., 2011) as the OOD dataset to evaluate the performance of OOD detection. All values are percentages. ↑ indicates large values are better, and ↓ indicates smaller values are better. The best results are underlined.

Dataset $\mathcal{D}_{\text{in}}^{\text{test}}$	λ	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑
CIFAR-10		Softmax-based / Energy-based / Feature Norm-based		
	0	52.09 / 43.04 / 52.10	91.67 / 91.94 / 89.54	84.11 / 82.80 / 77.06
	1e-6	54.00 / 43.72 / **51.45**	91.44 / **92.12** / 89.08	82.31 / 81.77 / 74.16
	5e-6	**45.37** / **33.92** / **26.93**	**93.08** / **93.78** / **94.03**	**84.31** / **83.73** / **82.79**
	1e-5	**37.39** / **27.87** / **24.94**	**93.90** / **94.60** / **94.17**	**85.48** / **85.34** / **83.15**
CIFAR-100		Softmax-based / Energy-based / Feature Norm-based		
	0	87.75 / 89.84 / 95.54	71.01 / 71.94 / 59.54	55.42 / 56.69 / 43.21
	1e-6	**82.08** / **82.57** / **88.77**	**75.36** / **76.28** / **68.83**	**61.40** / **61.90** / **51.58**
	5e-6	**79.01** / **78.68** / **85.94**	**78.70** / **79.15** / **70.32**	**62.58** / **62.39** / **48.39**
	1e-5	**81.48** / **81.41** / **87.83**	**77.02** / **78.03** / **73.91**	**62.92** / **63.66** / **58.81**

within limited iterations (that means $\zeta_1(t) < \infty$) will introduce some residual $\boldsymbol{\Delta}(t)$ caused by gradient descent regardless of the unconstrained case or regularized case.

Example 1. *If we add a perturbation Δ for all features while adding $s\Delta$ for all prototypes, then the perturbed decision-making will be* $\arg\max_{c'} \langle \boldsymbol{w}_{c'} + s\Delta, \boldsymbol{h} + \Delta \rangle + b_{c'}$*, which may not be equivalent to* $\arg\max_{c'} \langle \boldsymbol{w}_{c'}, \boldsymbol{h} \rangle + b_{c'}$*.*

D.3 Adjusted Sample Margin Loss

As aforementioned in Section 3.4 and the proof of Theorem 7, we will encounter zero gradients when the cosine similarity $\hat{\boldsymbol{w}}_y^\top \hat{\boldsymbol{h}}$ is -1 or 1, so we can adjust the loss to avoid the issue by to $\boldsymbol{w}_y$ and accelerate convergence:

$$L'_\gamma(\boldsymbol{W}\hat{\boldsymbol{h}}, y) = \begin{cases} L_\gamma(\boldsymbol{W}\hat{\boldsymbol{h}}, y) & \text{if } \hat{\boldsymbol{w}}_y^\top \hat{\boldsymbol{h}} \geq -1 + \varepsilon, \\ -(1+\gamma)(\boldsymbol{w}_y + \boldsymbol{\delta})^\top \hat{\boldsymbol{h}} & \text{if } \hat{\boldsymbol{w}}_y^\top \hat{\boldsymbol{h}} < -1 + \varepsilon, \end{cases} \tag{104}$$

where $\varepsilon \in (0, 1)$ is a hyperparameter and $\boldsymbol{\delta} = -\left(1 + \frac{\hat{\boldsymbol{w}}_y^\top \hat{\boldsymbol{h}} \sqrt{1-(1-\varepsilon)^2}}{(1-\varepsilon)\sqrt{1-(\hat{\boldsymbol{w}}_y^\top \hat{\boldsymbol{h}})^2}}\right)(\boldsymbol{w}_y + \hat{\boldsymbol{h}}\hat{\boldsymbol{h}}^\top \boldsymbol{w}_y)$ (performed with a stop-gradient) satisfying $\frac{(\boldsymbol{w}_y + \boldsymbol{\delta})^\top \hat{\boldsymbol{h}}}{\|\boldsymbol{w}_y + \boldsymbol{\delta}\|_2} = -1 + \varepsilon$.

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. URL `https://arxiv.org/abs/1701.07875`.

Peter Auer, Nicolo Cesa-Bianchi, and Claudio Gentile. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

Sébastien Bubeck and Mark Sellke. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34:28811–28822, 2021.

Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

Ahmet Demirkaya, Jiasi Chen, and Samet Oymak. Exploring the role of loss functions in multiclass classification. In *2020 54th annual conference on information sciences and systems (ciss)*, pages 1–5. IEEE, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Cong Fang, Hanze Dong, and Tong Zhang. Mathematical models of overparameterized neural networks. *Proceedings of the IEEE*, 109(5):683–703, 2021a.

Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43):e2103091118, 2021b.

Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. In *International Conference on Learning Representations*, 2021.

Peifeng Gao, Qianqian Xu, Peisong Wen, Huiyang Shao, Zhiyong Yang, and Qingming Huang. A study of neural collapse phenomenon: Grassmannian frame, symmetry, generalization. *arXiv preprint arXiv:2304.08914*, 2023.

Suriya Gunasekar, Jason D Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems*, 31, 2018.

XY 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.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

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*, pages 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *International Conference on Learning Representations*, 2021.

Like Hui, Mikhail Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J Su. An unconstrained layer-peeled perspective on neural collapse. In *International Conference on Learning Representations*, 2022.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory*, pages 1772–1798. PMLR, 2019.

Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pages 2109–2136. PMLR, 2020.

Tejaswi Kasarla, Gertjan J Burghouts, Max van Spengler, Elise van der Pol, Rita Cucchiara, and Pascal Mettes. Maximum class separation as inductive bias in one matrix. *arXiv preprint arXiv:2206.08704*, 2022.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? *Advances in Neural Information Processing Systems*, 34:28648–28662, 2021.

Vignesh Kothapalli, Ebrahim Rasromani, and Vasudev Awatramani. Neural collapse: A review on modelling principles and generalization. *arXiv preprint arXiv:2206.04041*, 2022.

A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Computer Science Department, University of Toronto, Tech. Rep*, 1, 01 2009.

Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. What do we want from explainable artificial intelligence (xai)? – a stakeholder perspective on xai and a conceptual model guiding interdisciplinary xai research. *Artificial Intelligence*, 296:103473, 2021. ISSN 0004-3702.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.

Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *International Conference on Machine Learning*, pages 507–516. PMLR, 2016.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

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

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75, 2021.

Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 20(2):1–13, 2022.

Mor Shpigel Nacson, J. Lee, Suriya Gunasekar, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *AISTATS*, 2019.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2019.

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

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

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

Tomaso Poggio and Qianli Liao. Explicit regularization and implicit bias in deep network classifiers trained with the square loss. *arXiv preprint arXiv:2101.00072*, 2020.

Zhongang Qi, Saeed Khorram, and Li Fuxin. Embedding deep networks into visual explanations. *Artificial Intelligence*, 292:103435, 2021. ISSN 0004-3702.

Ayush Sekhari, Karthik Sridharan, and Satyen Kale. Sgd: The role of implicit regularization, batch-size and multiple-epochs. *Advances In Neural Information Processing Systems*, 34:27422–27433, 2021.

Ohad Shamir. Gradient methods never overfit on separable data. *J. Mach. Learn. Res.*, 22: 85:1–85:20, 2021.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

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

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pages 10268–10278. PMLR, 2021.

Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 21478–21505. PMLR, 2022.

Tom Tirer, Haoxiang Huang, and Jonathan Niles-Weed. Perturbation analysis of neural collapse. In *International Conference on Machine Learning*, pages 34301–34329. PMLR, 2023.

Brendan Van Rooyen, Aditya Menon, and Robert C Williamson. Learning with symmetric label noise: The importance of being unhinged. *Advances in neural information processing systems*, 28, 2015.

Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. Normface: L2 hypersphere embedding for face verification. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1041–1049, 2017.

Ke Wang, Vidya Muthukumar, and Christos Thrampoulidis. Benign overfitting in multiclass classification: All roads lead to interpolation. *Advances in Neural Information Processing Systems*, 34:24164–24179, 2021.

Mengjia Xu, Akshay Rangamani, Qianli Liao, Tomer Galanti, and Tomaso Poggio. Dynamics in deep classifiers trained with the square loss: Normalization, low rank, neural collapse, and generalization bounds. *Research*, 6:0024, 2023.

Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR, 2021.

Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representation*, 2017.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021.

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 *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27179–27202. PMLR, 17–23 Jul 2022a.

Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. In *Advances in Neural Information Processing Systems*, 2022b.

Xiong Zhou, Xianming Liu, Deming Zhai, Junjun Jiang, Xin Gao, and Xiangyang Ji. Prototype-anchored learning for learning with imperfect annotations. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27245–27267. PMLR, 17–23 Jul 2022c.

Xiong Zhou, Xianming Liu, Deming Zhai, Junjun Jiang, Xin Gao, and Xiangyang Ji. Learning towards the largest margins. In *International Conference on Learning Representations*, 2022d.

Xiong Zhou, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Asymmetric loss functions for noise-tolerant learning: Theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

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