ON THE DYNAMICS UNDER THE AVERAGED SAMPLE MARGIN LOSS AND BEYOND

Anonymous authors
Paper under double-blind review

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 dynamics under gradient descent due to the intractability of loss functions and neural architectures. In this paper, we introduce a concise loss function as a surrogate, namely the Averaged Sample Margin (ASM) loss, which offers more mathematical opportunities to analyze the closed-form dynamics while requiring few simplifications or assumptions, and allows for more practical considerations. Based on the layer-peeled model that views last-layer features as free optimization variables, we build a complete analysis for the unconstrained, regularized, and spherical constrained cases. We show that these dynamics mainly *converge exponentially fast* to a solution depending on the initialization of features and classifier weights, which can help explain why the training of deep neural networks usually takes only a few hundred epochs. Our theoretical results can also aid in providing insights for improvements in practical training with the ASM loss or other losses, such as explicit feature regularization and rescaled learning rate for spherical cases. Finally, we empirically demonstrate these theoretical results and insights with extensive experiments.

1 INTRODUCTION

Deep learning with neural networks has achieved great success in a variety of tasks (Goodfellow et al., 2016), 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., 2019; Bubeck and Sellke, 2021; Mei and Montanari, 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 (DNNs), but lack convincing explanations. Due to the complexity of multilayered architectures, recent works are devoted to simplifying modeling to better understand the behavior of neural networks, and then to gain insights for new algorithms, theoretical, and experimental investigations.

To understand the implicit regularization that improves generalization of the 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). Soudry et al. (2018) proved that gradient descent iterates under exponentially-tailed loss minimization on separable data are biased toward ℓ_2-maximum-margin solutions and further showed that continuing to optimize can still lead to performance improvements even if the validation loss increases. Shamir (2021) formally showed 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 $O(1/\log t)$. However, the training of DNNs typically takes only a few hundred epochs. In this paper, we show a exponential convergence that is more realistic.

Another research line is after the empirical discovery of Neural Collapse by Papyan et al. (2020), which 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 phenomenon in simplified mathematical frameworks (Mixon et al., 2022; Fang et al., 2021; Galanti et al., 2021; Zhu et al., 2021; Hui et al., 2022; Tirer and Bruna, 2022; Lu and Steinerberger, 2022; Kothapalli et al., 2022). 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 in generalization and robustness after achieving zero training error. Although we have a clear picture of the final phase, the intermediate dynamics is hard to analyze as a result of the intractability of CE. Therefore, some work (Mixon et al., 2022; Han et al., 2022; Zhou et al., 2022a; Tirer and Bruna, 2022; Kothapalli et al., 2022) revolves around the more tractable MSE loss that performs comparably to those trained with CE (Demirkaya et al., 2020; Hui and Belkin, 2021), but they still need to make some simplifications or assumptions about the learning process. 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. Furthermore, to derive exact dynamics, 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. Therefore, MSE is still not simple enough to derive exact dynamics in certain mathematical frameworks, making it difficult to grasp the gap between the modeling and practical optimization.

In this paper, within the layer-peeled model (Fang et al., 2021) (or the unconstrained features model (Mixon et al., 2022)), we attempt to analyze the closed-form dynamics under gradient descent with as few simplifications as possible. More specifically, we introduce a new loss function, namely the Averaged Sample Margin (ASM) loss, which has a concise form that intuitively expresses the objective of classification. Compared to CE and MSE, the ASM loss offers more mathematical opportunities to let us glimpse into deep learning with the closed-form dynamics while requiring few simplifications or assumptions, and will prepare us for more practical considerations and more reasonable designs in a later section. The main contributions of our work are highlighted as follows:

- We derive exact dynamics of last-layer features and prototypes in unconstrained and regularized cases. 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 can help explain why the training of deep neural networks usually takes only a few hundred epochs.
- Moreover, we provide some insights for improvements in practical training with the ASM loss or other losses (cf. Section 4).

2 The Averaged Sample Margin Loss

In this paper, we mainly focus on the behavior of last-layer features and classifier weights in classification DNNs. As described in prior works (Fang et al., 2021; Han et al., 2022), we also consider datasets containing inputs from C different classes with N examples in each class. The last-layer features $\boldsymbol{h}_{i,c} = \boldsymbol{f}_\Theta(\boldsymbol{x}_{i,c}) \in \mathbb{R}^p$[1] extracted from the i-th example $\boldsymbol{x}_{i,c}$ by a number of parameterized layers $\boldsymbol{f}_\Theta \colon \mathcal{X} \to \mathbb{R}^p$ are usually simplified as free optimization variables (Mixon et al., 2022; Fang et al., 2021; Han et al., 2022; Ji et al., 2022). The last layer of the network, *i.e.*, the linear classifier, possesses a class prototype $\boldsymbol{w}_c \in \mathbb{R}^p$ and bias $b_c \in \mathbb{R}$ for each class $c \in [C]$, which predicts a label using the rule $\arg\max_{c'}(\langle \boldsymbol{w}_{c'}, \boldsymbol{h}_{i,c}\rangle + b_{c'})$.

[1]To obtain the closed-form solution of neural collapse, we note that Zhou et al. (2022c) assume that $p \geq C-1$, while Han et al. (2022) assume that $p > C$ since the last-layer features are usually of higher dimension than the number of classes. In this work, we will directly emphasize the relationship between p and C in some scenarios and cover all choices of the feature dimension for others.

To better understand the dynamics of features and prototypes based on gradient descent, we consider a surrogate loss that offers more mathematical opportunities than the hard-to-analyze CE loss and the MSE loss. Specifically, we introduce the Averaged Sample Margin (ASM) loss as follows:

$$L_{ASM}(\boldsymbol{W}\boldsymbol{h}+\boldsymbol{b},y) = -\boldsymbol{w}_y^\top \boldsymbol{h} - b_y + \gamma \sum_{j\neq y}(\boldsymbol{w}_j^\top \boldsymbol{h} + b_j), \tag{2.1}$$

where $\gamma > 0$ is the trade-off parameter and y denotes the class label of the feature $\boldsymbol{h}$. 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) is defined to measure the discriminativeness for a sample, which satisfies $m(\boldsymbol{h},y) \leq \frac{1}{k-1}\sum_{j\neq y}[\boldsymbol{w}_y^\top \boldsymbol{h} + b_y - (\boldsymbol{w}_j^\top \boldsymbol{h} + b_j)]$, *i.e.*, $L_{ASM}(\boldsymbol{W}\boldsymbol{h}+\boldsymbol{b},y)$ with $\gamma = \frac{1}{k-1}$ averaging the margins over all non-target classes is the lower bound of $-m(\boldsymbol{h},y)$. Moreover, $\frac{1}{k-1}\sum_{j\neq y}[\boldsymbol{w}_y^\top \boldsymbol{h} + b_y - (\boldsymbol{w}_j^\top \boldsymbol{h} + b_j)]$ can also be regarded as the unhinged version that removes the max operator and margin term in the hinge loss [2]. Here, we replace $\frac{1}{k-1}$ with an additional parameter γ that balances positive and negative logits to draw general conclusions. Furthermore, the ASM loss can also be regarded as a variant of CE and multi-binary CE loss:

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

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

Lemma 2.1 (**Neural Collapse under Averaged Sample Margin 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 $\frac{1}{CN}\sum_{i=1}^N\sum_{c=1}^C L_{ASM}(\boldsymbol{W}\boldsymbol{h}_{i,c}, y_{i,c})$ 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 is the only global optimal solution to minimize $\frac{1}{CN}\sum_{i,c} L_{ASM}(\boldsymbol{W}\boldsymbol{h}_{k,c}, y_{i,c})$ in the norm-bounded case. However, there exists an undesired direction to minimize the ASM loss in unconstrained cases, since the norm of features and prototypes tends to grow to infinity. For example, we can scale up $\boldsymbol{W}$ and $\boldsymbol{b}$ to obtain a smaller loss if $L_{ASM}(\boldsymbol{W}\boldsymbol{h}+\boldsymbol{b},y) < 0$, which will happen analogously to CE (Liu et al., 2016; Wang et al., 2017; Zhou et al., 2022c). In this paper, we will analytically characterize the direction in which features $\boldsymbol{H}$ and prototypes $\boldsymbol{W}$ diverge. Specifically, we show that the continual gradient flow exhibits an implicit bias associated with the initialization of features and prototypes.

3 Main Theoretical Results

In this section, we build a complete analysis of the closed-form dynamics of the last-layer features and prototypes under the ASM loss in unconstrained, regularized, and spherical constrained cases. We then derive the convergence analysis of these dynamics, which mainly shows exponential convergence. **All proof can be found in the Appendix A.**

3.1 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

$$\mathcal{L}_{ASM} = \frac{1}{CN}\sum_{i=1}^N\sum_{c=1}^C\Bigg[-\left(\boldsymbol{w}_{y_{i,c}}^\top \boldsymbol{h}_{i,c} + b_{y_{i,c}}\right) + \gamma\sum_{j\neq y_{i,c}}\left(\boldsymbol{w}_j^\top \boldsymbol{h}_{i,c} + b_j\right)\Bigg], \tag{3.1}$$

[2] The multi-class hinge loss is $L_{\text{hinge}}(\boldsymbol{W}\boldsymbol{h}+\boldsymbol{b},y) = \sum_{i\neq y}\max\{0, (\boldsymbol{w}_i^\top \boldsymbol{h} + b_i) - (\boldsymbol{w}_y^\top \boldsymbol{h} + b_y) + m\}$, where m is the margin term.

which can be reformulated as

$$\mathcal{L}_{ASM} = \frac{1}{CN} \operatorname{Tr}\left(\left(\gamma \mathbf{1}_C \mathbf{1}_{CN}^\top - (1+\gamma)(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top)\right)^\top \boldsymbol{W}^\top \boldsymbol{H} \right) + \frac{\gamma C - \gamma - 1}{C} \mathbf{1}_C^\top \boldsymbol{b}, \tag{3.2}$$

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, $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_C] \in \mathbb{R}^{p \times C}$ is the matrix resulting from stacking 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. Without loss of generality, we represent the label set $\{y_{i,c}\}_{1 \le i \le N, 1 \le c \le C}$ as the columns of the matrix $\boldsymbol{I}_C \otimes \mathbf{1}_N^\top$.

We follow the unconstrained features and prototypes modeling perspective (Mixon et al., 2022; Ji et al., 2022) 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[3].

Let $\boldsymbol{Y} = \frac{1}{CN}(\gamma \mathbf{1}_C \mathbf{1}_{CN}^\top - (1+\gamma)(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top))$, then $\mathcal{L}_{ASM} = \operatorname{Tr}(\boldsymbol{Y}^\top \boldsymbol{W}^\top \boldsymbol{H}) + \frac{\gamma C - \gamma - 1}{C} \mathbf{1}_C^\top \boldsymbol{b}$. Taking the partial derivative with respect to $\boldsymbol{H}$, $\boldsymbol{W}$, and $\boldsymbol{b}$, respectively, we have:

$$\nabla_{\boldsymbol{H}} \mathcal{L}_{ASM} = \boldsymbol{W}\boldsymbol{Y}, \quad \nabla_{\boldsymbol{W}} \mathcal{L}_{ASM} = \boldsymbol{H}\boldsymbol{Y}^\top, \quad \nabla_{\boldsymbol{b}} \mathcal{L}_{ASM} = \frac{\gamma C - \gamma - 1}{C} \mathbf{1}_C, \tag{3.3}$$

and the corresponding learning dynamics following Gradient Descent (GD) is

$$\boldsymbol{H}'(t) = \eta_1(t) \boldsymbol{W}(t) \boldsymbol{M}, \quad \boldsymbol{W}'(t) = \eta_2(t) \boldsymbol{H}(t) \boldsymbol{M}^\top, \quad \boldsymbol{b}'(t) = -\eta_2(t) \frac{\gamma C - \gamma - 1}{C} \mathbf{1}_C, \tag{3.4}$$

where $\boldsymbol{M} = -\boldsymbol{Y} = \frac{1}{CN}((1+\gamma)(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top) - \gamma \mathbf{1}_C \mathbf{1}_{CN}^\top)$, η_1 and η_2 are the learning rate of the features and prototypes, respectively. The reason to introduce different learning rates is that the representation $\boldsymbol{H}$ is actually the result of the interaction between a number of nonlinear layers and parameterized layers. This means that even if we use the same learning rate to optimize all parameters of the network, the feature $\boldsymbol{H}$ assumed to be a free optimization variable is almost impossible to be optimized at this learning rate[4]. Moreover, we consider dynamic learning 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 3.1 (**Dynamics of Features, Prototypes and Biases without Constraints**). *Consider the continual gradient flow (Equation 3.4) in which the dynamics follow the gradient descent direction of the Averaged Sample Margin loss in Eq. (3.2). 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*

$$\begin{aligned} \boldsymbol{Z}(t) =& \Pi_1^+ \boldsymbol{Z}_0 \left(\alpha_1^+(t)\boldsymbol{C}(t) + \beta_1^+(t)\boldsymbol{I}_{C(N+1)}\right) + \Pi_1^- \boldsymbol{Z}_0 \left(\alpha_1^-(t)\boldsymbol{C}(t) + \beta_1^-(t)\boldsymbol{I}_{C(N+1)}\right) + \Pi_3 \boldsymbol{Z}_0 \\ &+ \Pi_2^+ \boldsymbol{Z}_0 \left(\alpha_2^+(t)\boldsymbol{C}(t) + \beta_2^+(t)\boldsymbol{I}_{C(N+1)}\right) + \Pi_2^- \boldsymbol{Z}_0 \left(\alpha_2^-(t)\boldsymbol{C}(t) + \beta_2^-(t)\boldsymbol{I}_{C(N+1)}\right), \end{aligned} \tag{3.5}$$

and

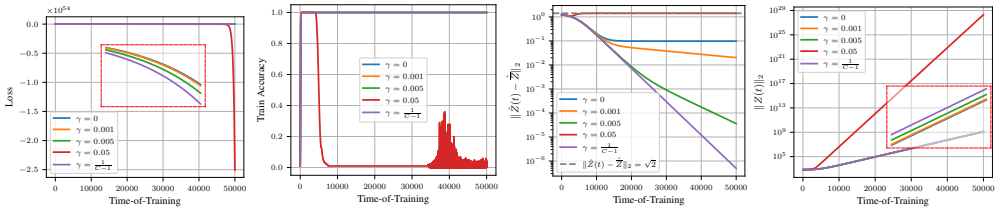
$$\boldsymbol{b}(t) = \boldsymbol{b}_0 + \frac{(1 + \gamma - \gamma C)\zeta_2(t)}{C} \mathbf{1}_C, \tag{3.6}$$

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 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:

$$\begin{aligned} \mathcal{E}_1^\epsilon &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = \epsilon \cdot \tfrac{1}{\sqrt{N}}(\boldsymbol{W} \otimes \mathbf{1}_N^\top), \boldsymbol{W}\mathbf{1}_C = 0\}, \\ \mathcal{E}_2^\epsilon &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = \epsilon \cdot \tfrac{1}{\sqrt{N}} \boldsymbol{h}\mathbf{1}_{CN}^\top, \boldsymbol{W} = \boldsymbol{h}\mathbf{1}_C^\top, \boldsymbol{h} \in \mathbb{R}^p\}, \\ \mathcal{E}_3 &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H}(\boldsymbol{I}_C \otimes \mathbf{1}_N) = 0, \boldsymbol{W} = 0\}. \end{aligned} \tag{3.7}$$

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

[4] 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.*, $\eta_1(t) = s \cdot \eta_2(t)$



(a) Train Loss (b) Train Accuracy (c) $\|\widehat{\boldsymbol{Z}}(t) - \widehat{\overline{\boldsymbol{Z}}}\|_2$ (d) $\|\boldsymbol{Z}(t)\|_2$

Figure 1: Verification of the behavior of gradient descent iterates in Equation (3.4) 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 3.2), and then randomly initialize $\boldsymbol{H}_0$ and $\boldsymbol{W}_0$. The curve in the red box represents the zoomed-in curve 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) denote the distance between $\widehat{\boldsymbol{Z}}(t) = \boldsymbol{Z}(t)/\|\boldsymbol{Z}(t)\|_2$ and $\widehat{\overline{\boldsymbol{Z}}} = \overline{\boldsymbol{Z}}/\|\overline{\boldsymbol{Z}}\|_2$. As expected in Eq. (3.9), the convergence rate is exponential when $0 < \gamma < \frac{2}{C-2}$, and will be fastest if $\gamma = \frac{1}{C-1}$. (d) denotes the norm of $\boldsymbol{Z}(t)$ that increases exponentially. As can be noticed, $\widehat{\boldsymbol{Z}}(t)$ does not converge to $\widehat{\overline{\boldsymbol{Z}}}$ but tend to be orthogonal to $\widehat{\overline{\boldsymbol{Z}}}$ when $\gamma = 0.05 > \frac{2}{C-2}$, that is, $\lim_{t\to\infty} \|\widehat{\boldsymbol{Z}}(t) - \widehat{\overline{\boldsymbol{Z}}}\|_2 = \sqrt{2}$.

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: *variability collapse* and *convergence to self-duality* (Papyan et al., 2020). In the following, we show that $\frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|}$ will converge to a solution in $\mathcal{E}$:

Corollary 3.2 (Convergence in the Unconstrained Case). *Under the conditions and notation of Theorem 3.1, 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 Eq. (3.4)) will behave as:*

$$e^{-\frac{(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}} \boldsymbol{Z}(t) = \overline{\boldsymbol{Z}} + \boldsymbol{\Delta}(t), \tag{3.8}$$

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 least 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*

$$\left\|\frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|} - \frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|}\right\| = O\left(e^{\frac{\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\cdot\max\{-\gamma C,(C-2)\gamma-2\}}\right), \tag{3.9}$$

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

This corollary shows that even without any explicit regularization, when minimizing the ASM loss using gradient descent, we prove that $\frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|}$ converges to a special point $\frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|} \in \mathcal{E}$ determined by the initialization of $\boldsymbol{Z}_0$ and the ratio s, but does not necessarily perform as a neural collapse solution that forms a simplex equiangular tight frame as CE or MSE do (Papyan et al., 2020; Han et al., 2022). In addition, the rate of convergence is exponential as a function of the integral of the learning rate, *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 explains 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 Eq. (3.9), that is, $\left\| \frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|} - \frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|} \right\| = O\left(e^{-\frac{\sqrt{\zeta_1(t)\zeta_2(t)}}{(C-1)\sqrt{N}}}\right)$. For example, when $\eta_1(t) = \eta$ is a constant learning rate, then $\zeta_1(t) = \eta t \to \infty$ as $t \to \infty$, and the gradient flow in Eq. (3.4) shows an exponential convergence rate of $\frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|}$ to $\frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|}$, but it also leads to an exponential increase with the rate $e^{\frac{(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}}$ in the norm of features and prototypes, which is almost unbearable for the practical training of the models. Therefore, we need to limit the excessive growth of these norms.

3.2 Regularized Case

We then consider the following regularized optimization problem that explicitly introduces ℓ_2 regularization (often called "weight decay") on features, prototypes, and biases to avoid large values:

$$\min_{\boldsymbol{W},\boldsymbol{H},\boldsymbol{b}} \operatorname{Tr}(\boldsymbol{Y}^\top \boldsymbol{W}^\top \boldsymbol{H}) + \tfrac{\gamma C-\gamma-1}{C} \mathbf{1}_C^\top \boldsymbol{b} + \tfrac{\lambda}{2}\left(\|\boldsymbol{H}\|_F^2 + \|\boldsymbol{W}\|_F^2 + \|\boldsymbol{b}\|_2^2\right), \tag{3.10}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and $\lambda > 0$ is the regularization parameter.

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

$$\nabla_{\boldsymbol{H}} \mathcal{L}_{ASM} = \boldsymbol{W}\boldsymbol{Y} + \lambda \boldsymbol{H}(t),\ \nabla_{\boldsymbol{W}} \mathcal{L}_{ASM} = \boldsymbol{H}\boldsymbol{Y}^\top + \lambda \boldsymbol{W}(t),\ \nabla_{\boldsymbol{b}} \mathcal{L}_{ASM} = \tfrac{\gamma C-\gamma-1}{C}\mathbf{1}_C + \lambda \boldsymbol{b}(t),$$

and the corresponding learning dynamics following gradient descent with different learning rates and ℓ_2-norm regularization can be formulated as

$$\begin{aligned} \boldsymbol{H}'(t) &= \eta_1(t)\boldsymbol{W}(t)\boldsymbol{M} - \lambda\eta_1(t)\boldsymbol{H}(t), \\ \boldsymbol{W}'(t) &= \eta_2(t)\boldsymbol{H}(t)\boldsymbol{M}^\top - \lambda\eta_2(t)\boldsymbol{W}(t), \\ \boldsymbol{b}'(t) &= -\eta_2(t)\tfrac{\gamma C-\gamma-1}{C}\mathbf{1}_C - \lambda\eta_2(t)\boldsymbol{b}(t). \end{aligned} \tag{3.11}$$

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

Theorem 3.3 (Dynamics of Features, Prototypes, and Biases under Weight Decay). *Consider the continual gradient flow (Equation 3.11) in which the dynamics follows the gradient descent direction of the Averaged Sample Margin loss in Eq. (3.10). 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.*

$$\begin{aligned} \boldsymbol{Z}(t) =& \Pi_1^+ \boldsymbol{Z}_0 \begin{pmatrix} a_1^+(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_1^+(t)\boldsymbol{I}_C \end{pmatrix} + \Pi_1^- \boldsymbol{Z}_0 \begin{pmatrix} a_1^-(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_1^-(t)\boldsymbol{I}_C \end{pmatrix} + \\ & \Pi_2^+ \boldsymbol{Z}_0 \begin{pmatrix} a_2^+(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_2^+(t)\boldsymbol{I}_C \end{pmatrix} + \Pi_2^- \boldsymbol{Z}_0 \begin{pmatrix} a_2^-(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_2^-(t)\boldsymbol{I}_C \end{pmatrix} + \\ & \Pi_3 \boldsymbol{Z}_0 \begin{pmatrix} a_3(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_3(t)\boldsymbol{I}_C \end{pmatrix}, \end{aligned} \tag{3.12}$$

and

$$\boldsymbol{b}(t) = \phi(t)\left(\boldsymbol{b}_0 + \tfrac{1+\gamma-\gamma C}{C}\psi(t)\mathbf{1}_C\right), \tag{3.13}$$

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 3.1, 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, η_1, and η_2 (where the detailed forms can be seen in Appendix A), $\phi(t) = \exp(-\lambda\int_0^t \eta_2(\tau)\mathrm{d}\tau)$, and $\psi(t) = \int_0^t \zeta_2(\tau)\exp(\lambda\int_0^\tau \eta_2(s)\mathrm{d}s)\mathrm{d}\tau$.

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

Corollary 3.4 (Convergence Under ℓ_2 Regularization). *Under the conditions and notation of Theorem 3.3, 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 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 Eq. (3.11)) behaves as:*

$$\left\| \frac{\boldsymbol{H}(t)}{\|\boldsymbol{H}(t)\|} - \frac{\pi_h^+\boldsymbol{H}_1^+ + \pi_h^-\boldsymbol{H}_1^-}{\|\pi_h^+\boldsymbol{H}_1^+ + \pi_h^-\boldsymbol{H}_1^-\|} \right\| + \left\| \frac{\boldsymbol{W}(t)}{\|\boldsymbol{W}(t)\|} - \frac{\pi_w^+\boldsymbol{W}_1^+ + \pi_w^-\boldsymbol{W}_1^-}{\|\pi_w^+\boldsymbol{H}_1^+ + \pi_w^-\boldsymbol{H}_1^-\|} \right\| = O(e^{-\omega\zeta_2(t)}), \tag{3.14}$$

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 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* $\lim_{t\to\infty} \|\boldsymbol{Z}(t)\| = \infty$.

Remark. The results in Corollary 3.4 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 is fragile, as depicted in Fig. 19. As a consequence, we emphasize *adding explicit regularization to features, not just implicit penalization attached by other components* in Sec. 4.2.

3.3 Spherical Constrained Case

We finally 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 to prevent arithmetic overflow or underflow happening in training DNNs, and we fix the prototypes[5] to satisfy $\boldsymbol{W}\mathbf{1}_C = 0$[6], then the optimization problem in Eq. (3.1) can be reformulated as

$$\min_{\boldsymbol{H}} -\frac{1+\gamma}{CN} \operatorname{Tr}((\boldsymbol{I}_C \otimes \mathbf{1}_N)\boldsymbol{W}^\top \widehat{\boldsymbol{H}}), \tag{3.15}$$

where $\widehat{\boldsymbol{H}} = (\widehat{\boldsymbol{h}}_{1,1}, ..., \widehat{\boldsymbol{h}}_{c,N})$, and $\widehat{\boldsymbol{h}}_{i,c} = \frac{\boldsymbol{h}_{i,c}}{\|\boldsymbol{h}_{i,c}\|_2}$. We then take the partial derivative with respect to $\boldsymbol{H}$, and the discrete dynamical system based on gradient descent is formulated as

$$\boldsymbol{H}(t+1) = \boldsymbol{H}(t) + \frac{(1+\gamma)\eta(t)}{CN}\left(\frac{\partial \widehat{\boldsymbol{H}}}{\partial \boldsymbol{H}}\Big|_{\boldsymbol{H}=\boldsymbol{H}(t)}\right)^\top \boldsymbol{W}(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top), \tag{3.16}$$

where $(\frac{\partial \widehat{\boldsymbol{H}}}{\partial \boldsymbol{H}})^\top$ is an element-wise operator, that is $(\frac{\partial \widehat{\boldsymbol{H}}}{\partial \boldsymbol{H}})^\top \boldsymbol{H}' = \left(\frac{1}{\|\boldsymbol{h}_{i,c}\|_2}(\boldsymbol{I}_p - \widehat{\boldsymbol{h}}_{i,c}\widehat{\boldsymbol{h}}_{i,c}^\top)\boldsymbol{h}'_{i,c}\right)$ for $\boldsymbol{H}' \in \mathbb{R}^{p\times CN}$.

Despite the fact that the optimization objective in Eq. (3.15) does not show convexity, Lipschitzness, and β-smoothness on $\boldsymbol{H}$ due to the ℓ_2 normalization operator, the normalized features that obey the gradient descent iterates in Eq. (3.16) can still converge to their corresponding normalized prototypes, *i.e.*, achieve the global minimum of Eq. (3.15):

Theorem 3.5. *Considering the discrete dynamics in Eq. (3.16), if* $\forall i \in [N], c \in [C]$, $\widehat{\boldsymbol{w}}_c^\top \widehat{\boldsymbol{h}}_{i,c}(0) > -1$, *the learning rate* $\eta(t)$ *satisfies that* $\frac{\eta(t)}{\|\boldsymbol{h}_{i,c}(t)\|_2}$ *is non-increasing,* $\frac{\eta(0)(1+\gamma)}{CN\|\boldsymbol{h}_{i,c}(0)\|_2} \le \frac{1}{\|\boldsymbol{w}_c\|_2}$, $\lim_{t\to\infty} \frac{\eta(t+1)}{\eta(t)} = 1$, *and there exists a constant* $\varepsilon > 0$, *s.t.,* $\eta(t) > \varepsilon$, *then we have*

$$\lim_{t\to\infty} \left\|\widehat{\boldsymbol{H}}(t) - \widehat{\boldsymbol{W}}(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top)\right\| = 0, \tag{3.17}$$

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:*

$$\left\|\widehat{\boldsymbol{H}}(t) - \widehat{\boldsymbol{W}}(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top)\right\| = O(e^{-\mu t}). \tag{3.18}$$

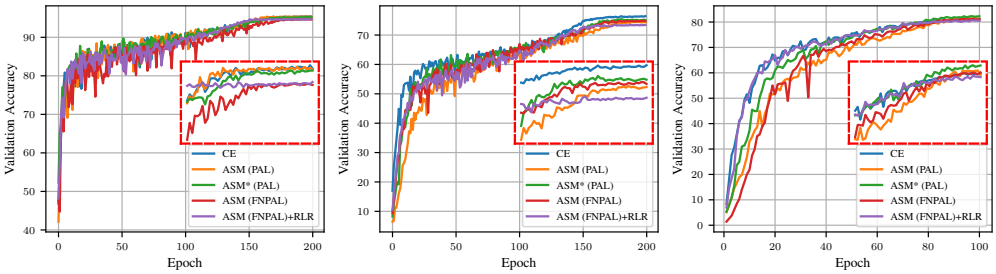
Moreover, if $\widehat{\boldsymbol{w}}_c^\top \widehat{\boldsymbol{h}}_{i,c}(0) = -1$, *then* $\boldsymbol{h}_{i,c}(t) = \boldsymbol{h}_{i,c}(0)$.

4 Insights and Experiments

In this section, we provide some insights to better train DNNs according to the conclusions in Sec. 3. We then corroborate our theoretical results and insights with extensive experiments. **More details and results can be found in Appendix B.**

[5]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.

[6]This aims to simplify Eq. (2.1) as the objective of feature alignment, that is, $L_{ASM}(\boldsymbol{W}\widehat{\boldsymbol{h}}, y) = -\boldsymbol{w}_y^\top \widehat{\boldsymbol{h}} + \gamma \sum_{j\neq y} \boldsymbol{w}_j^\top \widehat{\boldsymbol{h}} = \frac{(1+\gamma)\|\boldsymbol{w}\|_2}{2}(\|\widehat{\boldsymbol{h}} - \widehat{\boldsymbol{w}}_y\|_2^2 - 2)$, and the global minimum will be obtained at $\widehat{\boldsymbol{h}} = \widehat{\boldsymbol{w}}_y$.



(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., 2022b). The curve in the red box represents the zoomed-in curve of the last 50 epochs. As can be seen, DNNs trained with the ASM loss can achieve comparative or even better performance compared to those of the CE loss.

4.1 The ASM Loss with Prototype-anchored learning

Since directly using the ASM loss will lead to volatile effects, which is mainly reflected in the rapid increases of feature norms and the imbalance between class prototypes when training DNNs with the stochastic gradient method, as shown in Fig. 19. Inspired by recent works (Zhou et al., 2022b; Kasarla et al., 2022) that use the neural collapse solution as an inductive bias, we can anchor prototypes $\boldsymbol{W}$ during training, and then the dynamics of $\boldsymbol{H}$ with ℓ_2-norm-based regularization in Eq. (3.10) can be formulated as the first-order non-homogeneous linear difference equation:

$$\boldsymbol{H}'(t) = \eta(t)\boldsymbol{W}\boldsymbol{M} - \lambda\eta(t)\boldsymbol{H}(t), \tag{4.1}$$

and the solution to the non-homogeneous linear DE can be easily derived:

Theorem 4.1. *Consider the continual gradient flow (Equation (4.1)) in which the prototypes $\boldsymbol{W}$ are fixed, we have the closed-form dynamics:*

$$\boldsymbol{H}(t) = e^{-\lambda\int_0^t \eta(\tau)\mathrm{d}\tau}\boldsymbol{H}(0) + \frac{1 - e^{-\lambda\int_0^t \eta(\tau)\mathrm{d}\tau}}{\lambda}\boldsymbol{W}\boldsymbol{M}, \tag{4.2}$$

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

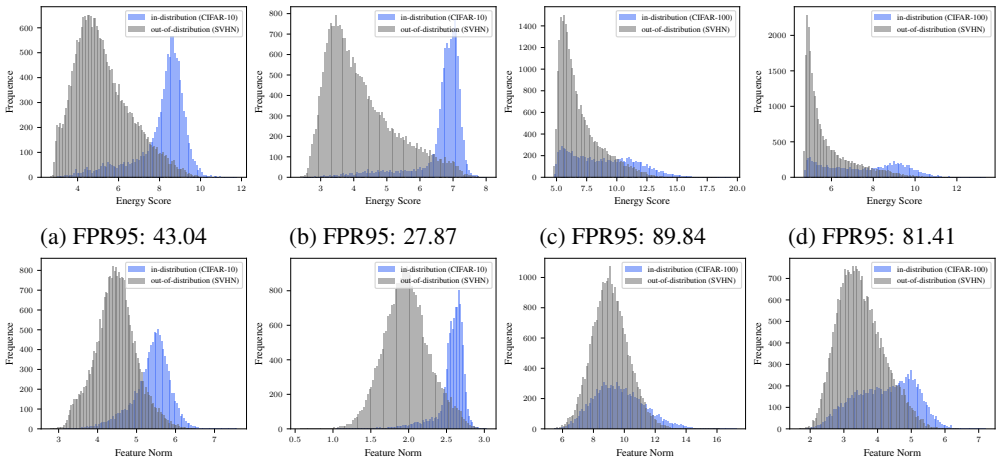
Prototype-anchored learning (PAL) is very effective in alleviating the instability by transforming the classification problem as a feature alighment problem (Zhou et al., 2022b). As shown in Fig. 2, the ASM loss with PAL and FNPAL can achieves comparable or even better results compared to CE.

Table 1: Validation accuracies on long-tailed CIFAR-10/-100 with CE and different explicit feature regularization ($\lambda \in \{0, 5e-6, 1e-5, 5e-5\}$). 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 baseline. The best results are underlined. As can be seen, explicit feature regularization effectively improve the performance on long-tailed classification in most cases, even for normal classification.

Dataset	Long-tailed CIFAR-10					Long-tailed CIFAR-100				
Imbalance Ratio	100	50	20	10	1	100	50	20	10	1
$\lambda = 0$	67.81	72.93	83.97	88.37	95.28	33.37	39.40	42.96	56.38	75.42
$\lambda = 5e-6$	**67.84**	72.85	83.17	**89.06**	95.27	**36.00**	**41.92**	**50.75**	**60.13**	**76.48**
$\lambda = 1e-5$	67.74	**76.14**	**84.17**	**89.19**	95.23	**36.61**	**42.36**	**49.21**	**58.91**	**77.34**
$\lambda = 5e-5$	**69.74**	**77.29**	**84.92**	**88.64**	**95.39**	**34.88**	**42.74**	**54.72**	**60.84**	**76.19**

4.2 Explicit Feature Regularization

In this paper, we directly consider explicit feature regularization to avoid excessive growth of feature norms, rather than fully adopt the implicit penalization described in prior works (Fang et al., 2021;



(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 improves the distinction between ID samples and OOD samples in the feature norm.

Zhou et al., 2022a). Explicit regularization on features can significantly remedy over-confidence and even improve generalization, which can be demonstrated in Tab. 1 and Fig. 3, where experiments on long-tailed classification (Zhong et al., 2021) and out-of-distribution (OOD) detection (Liu et al., 2020) are conducted, respectively. Moreover, adding explicit feature regularization can speed up the convergence of $\widehat{\boldsymbol{H}}(t)$ to $\widehat{\boldsymbol{W}\boldsymbol{M}}$ according to Theorem 4.1, as verified in Fig. 15.

4.3 Rescaled Learning Rate for The Spherical Constrained Case

As shown in Sec. 3.3, the gradient of the objective in Eq. (3.15) with respect to $\boldsymbol{h}_{i,c}$ is $-\frac{1+\gamma}{CN\|\boldsymbol{h}_{i,c}\|_2}(\boldsymbol{I}_p - \widehat{\boldsymbol{h}}_{i,c}\widehat{\boldsymbol{h}}_{i,c}^\top)\boldsymbol{w}_c$, which shows that the gradient is significantly influenced by $\frac{1}{\|\boldsymbol{h}_{i,c}\|_2}$. For example, the gradient will disappear when $\|\boldsymbol{h}_{i,c}\|_2$ is too large, and the gradient will be too large when $\|\boldsymbol{h}_{i,c}\|$ is too small. A natural solution would be scaling up the learning rate $\eta(t)$ with the feature norm $\|\boldsymbol{h}_{i,c}\|_2$ for each sample[7], and we can still guarantee the convergence in Theorem 3.5 if $\eta(t)$ is non-increasing and satisfies $\frac{\eta(0)(1+\gamma)}{CN} \leq \frac{1}{\|\boldsymbol{w}_c\|_2}$. As shown in Fig. 2c, the ASM loss with the rescaled learning rate (RLR) obviously converges faster than without it.

5 Conclusion

In this paper, we introduce the Averaged Sample Margin loss (ASM) as a surrogate to analyze the behavior of last-layer features and prototypes. Thanks to the conciseness of the ASM loss, we derive the exact dynamics under gradient descent in unconstrained, regularized, and spherical constrained cases. We then prove that these dynamics will converge exponentially to a particular solution relying on the initialization. Inspired by these results, we further provide some insights for improvements, such as the ASM loss with prototype-anchored learning, explicit feature regularization, and rescaled learning rate for spherical cases. We finally verify these theoretical results and insights with extensive experiments, including numerical analysis, visual classification, imbalanced learning, and out-of-distribution detection.

We expect that the ASM loss can be an excellent tool to help the community further understand the behavior of deep neural networks and beyond the scope of the paper.

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

REFERENCES

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.

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

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

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

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.

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.

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.

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.

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.

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.

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.

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.

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 2022b.

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

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.

Appendix for "Dynamics of The ASM Loss and Beyond"

A Proofs for Lemmas, Theorems, Propositions and Corollaries

A.1 Proof of Lemma 2.1

Lemma 2.1 (The Neural Collapse of The Averaged Sample Margin 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 $\frac{1}{CN}\sum_{i=1}^{N}\sum_{c=1}^{C} L_{ASM}(\boldsymbol{W}\boldsymbol{h}_{i,c}, y_{i,c})$ 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 $\frac{1}{CN}\sum_{i=1}^{N}\sum_{c=1}^{C} L_{ASM}(\boldsymbol{W}\boldsymbol{h}_{i,c}, y_{i,c})$ by a sequence of inequalities that holds if and only if the solution forms Neural Collapse (Papyan et al., 2020). Let $\widehat{\boldsymbol{w}} = \frac{1}{C}\sum_{c=1}^{C}\boldsymbol{w}_c$, according to the definition of L_{SM}, we have

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

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 $(C\widehat{\boldsymbol{w}} - \boldsymbol{w}_{y_{i,c}})^\top \boldsymbol{h}_{i,c} \geq -E_2\|C\widehat{\boldsymbol{w}} - \boldsymbol{w}_{y_{i,c}}\|_2$, 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 $\|\widehat{\boldsymbol{w}}\|_2 \geq 0$.

According the above derivation, the equality holds if and only if $\forall i \in [N], \forall c \in [C]$, $\boldsymbol{w}_{y_{i,c}}^\top \boldsymbol{h}_{i,c} = E_1E_2$, $(C\widehat{\boldsymbol{w}} - \boldsymbol{w}_{y_{i,c}})^\top \boldsymbol{h}_{i,c} = -E_2\|C\widehat{\boldsymbol{w}} - \boldsymbol{w}_{y_{i,c}}\|_2$, $\|C\widehat{\boldsymbol{w}} - \boldsymbol{w}_c\|_2 = \|C\widehat{\boldsymbol{w}} - \boldsymbol{w}_C\|_2$, $\|\boldsymbol{w}_c\|_2 = E_1$, and $\|\widehat{\boldsymbol{w}}\|_2 = 0$. 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. □

A.2 Proof of Theorem 3.1

In this section, we will provide the proof of Theorem 3.1. 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 3.1. Their concrete projection operator can be found in Appendix A.8. In the following, we show that these five subspaces are orthogonal:

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

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

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

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

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

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

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

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

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

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

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

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

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

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 3.1 (Dynamics of Features, Prototypes and Biases without Constraints). *Consider the continual gradient flow (Equation 3.4) in which the dynamics follow the gradient descent direction of Sample Margin loss in Eq. (3.2). 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*

$$\begin{aligned}
\boldsymbol{Z}(t) =& \Pi_1^+\boldsymbol{Z}_0\left(\alpha_1^+(t)\boldsymbol{C}(t) + \beta_1^+(t)\boldsymbol{I}_{C(N+1)}\right) + \Pi_1^-\boldsymbol{Z}_0\left(\alpha_1^-(t)\boldsymbol{C}(t) + \beta_1^-(t)\boldsymbol{I}_{C(N+1)}\right) + \Pi_3\boldsymbol{Z}_0 \\
&+ \Pi_2^+\boldsymbol{Z}_0\left(\alpha_2^+(t)\boldsymbol{C}(t) + \beta_2^+(t)\boldsymbol{I}_{C(N+1)}\right) + \Pi_2^-\boldsymbol{Z}_0\left(\alpha_2^-(t)\boldsymbol{C}(t) + \beta_2^-(t)\boldsymbol{I}_{C(N+1)}\right),
\end{aligned} \tag{A.2}$$

and

$$\boldsymbol{b}(t) = \boldsymbol{b}_0 + \frac{1+\gamma-\gamma C}{C}\int_0^t \eta_2(\tau)d\tau, \tag{A.3}$$

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 can be seen in the appendix),* $\boldsymbol{Z}_0 = (\boldsymbol{H}_0, \boldsymbol{W}_0)$*,* $\boldsymbol{C}(t) =$

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

$$\begin{aligned} \mathcal{E}_1^\epsilon &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = \epsilon \cdot \tfrac{1}{\sqrt{N}}(\boldsymbol{W} \otimes \mathbf{1}_N^\top), \boldsymbol{W}\mathbf{1}_C = 0\}, \\ \mathcal{E}_2^\epsilon &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = \epsilon \cdot \tfrac{1}{\sqrt{N}}\boldsymbol{h}\mathbf{1}_{CN}^\top, \boldsymbol{W} = \boldsymbol{h}\mathbf{1}_C^\top, \boldsymbol{h} \in \mathbb{R}^p\}, \\ \mathcal{E}_3 &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H}(\boldsymbol{I}_C \otimes \mathbf{1}_N) = 0, \boldsymbol{W} = 0\}. \end{aligned} \tag{A.4}$$

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

$$\boldsymbol{Z}'(t) = \boldsymbol{Z}(t) \begin{pmatrix} 0 & \boldsymbol{M}^\top \\ \boldsymbol{M} & 0 \end{pmatrix} \begin{pmatrix} \eta_1(t)\boldsymbol{I}_{CN} & 0 \\ 0 & \eta_2(t)\boldsymbol{I}_C \end{pmatrix}, \tag{A.5}$$

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:

$$\boldsymbol{Z}'(t) = \boldsymbol{Z}(t)\boldsymbol{A}(t), \quad \boldsymbol{Z}(0) = \boldsymbol{Z}_0. \tag{A.6}$$

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

$$\boldsymbol{Z}(t) = \boldsymbol{Z}_0 \exp\left(\int_0^t \boldsymbol{A}(\tau)\mathrm{d}\tau\right) = \boldsymbol{Z}_0 \exp\begin{pmatrix} 0 & \int_0^t \eta_2(\tau)\mathrm{d}\tau \boldsymbol{M}^\top \\ \int_0^t \eta_1(\tau)\mathrm{d}\tau \boldsymbol{M} & 0 \end{pmatrix}. \tag{A.7}$$

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

$$\boldsymbol{Z}(t) = \boldsymbol{Z}_0 \exp(\boldsymbol{L}(t)) = \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\boldsymbol{L}(t))^k}{k!}. \tag{A.8}$$

Moreover, we have

$$\begin{aligned} (\boldsymbol{L}(t))^2 &= \begin{pmatrix} 0 & \zeta_2(t)\boldsymbol{M}^\top \\ \zeta_1(t)\boldsymbol{M} & 0 \end{pmatrix} \begin{pmatrix} 0 & \zeta_2(t)\boldsymbol{M}^\top \\ \zeta_1(t)\boldsymbol{M} & 0 \end{pmatrix} \\ &= \begin{pmatrix} \zeta_1(t)\zeta_2(t)\boldsymbol{M}^\top\boldsymbol{M} & 0 \\ 0 & \zeta_1(t)\zeta_2(t)\boldsymbol{M}\boldsymbol{M}^\top \end{pmatrix} \\ &= \zeta_1(t)\zeta_2(t)\boldsymbol{B}^2, \end{aligned} \tag{A.9}$$

thus we obtain

$$\begin{aligned}\boldsymbol{Z}(t) =& \boldsymbol{Z}_0 \left(\sum_{k=0}^{\infty} \frac{(\boldsymbol{L}(t))^{2k+1}}{(2k+1)!} + \sum_{k=0}^{\infty} \frac{(\boldsymbol{L}(t))^{2k}}{(2k)!} \right) \\ =& \boldsymbol{Z}_0 \left(\sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k+1} \boldsymbol{C}(t)}{(2k+1)!} + \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k}}{(2k)!} \right) \\ =& \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k+1} \boldsymbol{C}(t)}{(2k+1)!} + \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k}}{(2k)!}\end{aligned} \tag{A.10}$$

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

Considering the following five subspaces:

$$\begin{aligned}\mathcal{E}_1^{\epsilon} &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = \epsilon \cdot \tfrac{1}{\sqrt{N}} (\boldsymbol{W} \otimes \mathbf{1}_N^\top), \boldsymbol{W}\mathbf{1}_C = 0\}, \\ \mathcal{E}_2^{\epsilon} &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H} = \epsilon \cdot \tfrac{1}{\sqrt{N}} \boldsymbol{h}\mathbf{1}_{CN}^\top, \boldsymbol{W} = \boldsymbol{h}\mathbf{1}_C^\top, \boldsymbol{h} \in \mathbb{R}^p\}, \\ \mathcal{E}_3 &:= \{(\boldsymbol{H}, \boldsymbol{W}) : \boldsymbol{H}(\boldsymbol{I}_C \otimes \mathbf{1}_N) = 0, \boldsymbol{W} = 0\}.\end{aligned} \tag{A.11}$$

where $\epsilon \in \{\pm\}$. According to Lemma A.1, 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

$$\boldsymbol{B} = \tfrac{1}{C\sqrt{N}}\left[(1+\gamma)(\Pi_1^+ - \Pi_1^-) + (1+\gamma-\gamma C)(\Pi_2^+ - \Pi_2^-)\right]. \tag{A.12}$$

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

$$\begin{aligned}
&\Pi_1^\epsilon \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k+1} \boldsymbol{C}(t)}{(2k+1)!} \\
=&\Pi_1^\epsilon \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \left(\frac{\epsilon(1+\gamma)}{C\sqrt{N}}\right)^{2k+1} \boldsymbol{C}(t)}{(2k+1)!} \\
=&\frac{\Pi_1^\epsilon \boldsymbol{Z}_0 \boldsymbol{C}(t)}{\sqrt{\zeta_1(t)\zeta_2(t)}} \sum_{k=0}^{\infty} \frac{\left(\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right)^{2k+1}}{(2k+1)!} \\
=&\frac{\Pi_1^\epsilon \boldsymbol{Z}_0 \boldsymbol{C}(t)}{2\sqrt{\zeta_1(t)\zeta_2(t)}} \left(e^{\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}} - e^{-\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}}\right),
\end{aligned} \tag{A.13}$$

and

$$\begin{aligned}
&\Pi_1^\epsilon \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k}}{(2k)!} \\
=&\Pi_1^\epsilon \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \left(\frac{\epsilon(1+\gamma)}{C\sqrt{N}}\right)^{2k}}{(2k)!} \\
=&\frac{\Pi_1^\epsilon \boldsymbol{Z}_0}{2} \left(e^{\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}} + e^{-\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}}\right),
\end{aligned} \tag{A.14}$$

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

$$\Pi_1^\epsilon \boldsymbol{Z}_0 \exp(\boldsymbol{L}(t)) = \Pi_1^\epsilon \boldsymbol{Z}_0 (\alpha_1^\epsilon \boldsymbol{C}(t) + \beta_1^\epsilon \boldsymbol{I}_{C(N+1)}), \tag{A.15}$$

with

$$\begin{aligned}
\alpha_1^\epsilon(t) &= \frac{\exp\left(\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right) - \exp\left(-\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right)}{2\sqrt{\zeta_1(t)\zeta_2(t)}}, \\
\beta_1^\epsilon(t) &= \frac{\exp\left(\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right) + \exp\left(-\frac{\epsilon(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right)}{2}.
\end{aligned} \tag{A.16}$$

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

$$\Pi_2^\epsilon \boldsymbol{Z}_0 \exp(\boldsymbol{L}(t)) = \Pi_2^\epsilon \boldsymbol{Z}_0 (\alpha_2^\epsilon \boldsymbol{C}(t) + \beta_2^\epsilon \boldsymbol{I}_{C(N+1)}), \tag{A.17}$$

with

$$\begin{aligned} \alpha_2^\epsilon(t) &= \frac{\exp\left(\frac{\epsilon(1+\gamma-\gamma C)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right) - \exp\left(-\frac{\epsilon(1+\gamma-\gamma C)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right)}{2\sqrt{\zeta_1(t)\zeta_2(t)}}, \\ \beta_2^\epsilon(t) &= \frac{\exp\left(\frac{\epsilon(1+\gamma-\gamma C)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right) + \exp\left(-\frac{\epsilon(1+\gamma-\gamma C)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\right)}{2}. \end{aligned} \tag{A.18}$$

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

$$\Pi_3 \boldsymbol{Z}_0 \left(\sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k+1} \boldsymbol{C}(t)}{(2k+1)!} + \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k}}{(2k)!} \right) = \Pi_3 \boldsymbol{Z}_0 \tag{A.19}$$

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

$$\begin{aligned} \boldsymbol{Z}(t) =& \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k+1} \boldsymbol{C}(t)}{(2k+1)!} + \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\zeta_1(t)\zeta_2(t))^k \boldsymbol{B}^{2k}}{(2k)!} \\ =& (\Pi_1^+ \boldsymbol{Z}_0 + \Pi_1^- \boldsymbol{Z}_0 + \Pi_2^+ \boldsymbol{Z}_0 + \Pi_2^- \boldsymbol{Z}_0 + \Pi_3 \boldsymbol{Z}_0) \exp(\boldsymbol{L}(t)) \\ =& \Pi_1^+ \boldsymbol{Z}_0 \left(\alpha_1^+(t)\boldsymbol{C}(t) + \beta_1^+(t)\boldsymbol{I}_{C(N+1)}\right) + \Pi_1^- \boldsymbol{Z}_0 \left(\alpha_1^-(t)\boldsymbol{C}(t) + \beta_1^-(t)\boldsymbol{I}_{C(N+1)}\right) + \Pi_3 \boldsymbol{Z}_0 \\ &+ \Pi_2^+ \boldsymbol{Z}_0 \left(\alpha_2^+(t)\boldsymbol{C}(t) + \beta_2^+(t)\boldsymbol{I}_{C(N+1)}\right) + \Pi_2^- \boldsymbol{Z}_0 \left(\alpha_2^-(t)\boldsymbol{C}(t) + \beta_2^-(t)\boldsymbol{I}_{C(N+1)}\right) \end{aligned} \tag{A.20}$$

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

$$\boldsymbol{b}(t) = \boldsymbol{b}(0) + \int_0^t -\eta_2(\tau)\frac{\gamma C-\gamma-1}{C}\mathrm{d}\tau \mathbf{1}_C = \boldsymbol{b}_0 + \frac{(1+\gamma-\gamma C)\zeta_2(t)}{C}\mathbf{1}_C, \tag{A.21}$$

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

A.3 Proof of Corollary 3.2

Corollary 3.2 *Under the conditions and notation of Theorem 3.1, 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 Eq. (3.4)) will behave as:*

$$e^{-\frac{(1+\gamma)\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}} \boldsymbol{Z}(t) = \left(\tfrac{1+\sqrt{s}}{2}\boldsymbol{H}_1^+ + \tfrac{1-\sqrt{s}}{2}\boldsymbol{H}_1^-, \tfrac{1+\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^+ - \tfrac{1-\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^-\right) + \boldsymbol{\Delta}(t), \tag{A.22}$$

where $(\boldsymbol{H}_1^+, \boldsymbol{W}_1^+) = \Pi_1^+ \boldsymbol{Z}_0$, $(\boldsymbol{H}_1^-, \boldsymbol{W}_1^-) = \Pi_1^- \boldsymbol{Z}_0$, and the residual term $\boldsymbol{\Delta}(t)$ decreases at least 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 let $\overline{\boldsymbol{Z}} = \left(\frac{1+\sqrt{s}}{2}\boldsymbol{H}_1^+ + \frac{1-\sqrt{s}}{2}\boldsymbol{H}_1^-, \frac{1+\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^+ - \frac{1-\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^-\right)$, the normalized $\boldsymbol{Z}(t)$ converges to the normalized $\overline{\boldsymbol{Z}}$ in

$$\left\| \frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|} - \frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|} \right\| = O\left(e^{\frac{\sqrt{\zeta_1(t)\zeta_2(t)}}{C\sqrt{N}}\cdot\max\{-\gamma C,(C-2)\gamma-2\}}\right), \tag{A.23}$$

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 3.1, we have

$$\begin{aligned} \boldsymbol{H}(t) &= \sum_{\substack{i\in\{1,2\}\\ \epsilon\in\{\pm\}}} (\alpha_i^\epsilon(t)\zeta_1(t) + \beta_i^\epsilon(t))\boldsymbol{H}_i^\epsilon + \boldsymbol{H}_3, \\ \boldsymbol{W}(t) &= \sum_{\substack{i\in\{1,2\}\\ \epsilon\in\{\pm\}}} (\alpha_i^\epsilon(t)\zeta_2(t) + \beta_i^\epsilon(t))\boldsymbol{W}_i^\epsilon + \boldsymbol{W}_3. \end{aligned} \tag{A.24}$$

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 (A.24) to obtain

$$\begin{aligned} e^{-p(t)}\boldsymbol{H}(t) &= \frac{1+\sqrt{s}}{2}\boldsymbol{H}_1^+ + \frac{1-\sqrt{s}}{2}\boldsymbol{H}_1^- + \boldsymbol{\Delta}_1(t), \\ e^{-p(t)}\boldsymbol{W}(t) &= \frac{1+\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^+ - \frac{1-\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^- + \boldsymbol{\Delta}_2(t), \end{aligned} \tag{A.25}$$

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

$$e^{-p(t)}\boldsymbol{Z}(t) = \left(\tfrac{1+\sqrt{s}}{2}\boldsymbol{H}_1^+ + \tfrac{1-\sqrt{s}}{2}\boldsymbol{H}_1^-, \tfrac{1+\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^+ - \tfrac{1-\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^-\right) + \boldsymbol{\Delta}(t), \tag{A.26}$$

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

$$\left\|\frac{\boldsymbol{Z}(t)}{\|\boldsymbol{Z}(t)\|} - \frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|}\right\| = \left\|\frac{\overline{\boldsymbol{Z}}+\boldsymbol{\Delta}(t)}{\|\overline{\boldsymbol{Z}}+\boldsymbol{\Delta}(t)\|} - \frac{\overline{\boldsymbol{Z}}}{\|\overline{\boldsymbol{Z}}\|}\right\| \le \frac{2\|\overline{\boldsymbol{Z}}\|\|\boldsymbol{\Delta}(t)\|}{\|\overline{\boldsymbol{Z}}+\boldsymbol{\Delta}(t)\|\|\overline{\boldsymbol{Z}}\|} = \frac{2\|\boldsymbol{\Delta}(t)\|}{\|\overline{\boldsymbol{Z}}+\boldsymbol{\Delta}(t)\|}, \tag{A.27}$$

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

$$\overline{\boldsymbol{Z}} = \left(\tfrac{\sqrt{s}}{\sqrt{N}}\left(\tfrac{1+\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^+ - \tfrac{1-\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^-\right)\otimes \boldsymbol{1}_N^\top, \tfrac{1+\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^+ - \tfrac{1-\sqrt{s}}{2\sqrt{s}}\boldsymbol{W}_1^-\right), \tag{A.28}$$

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

A.4 PROOF OF THEOREM 3.3

Theorem 3.3 (Dynamics of Features and Prototypes with ℓ_2 Regularization) *Under the conditions and notations of Theorem 3.1, consider the continual gradient flow (Equation 3.11) in which the dynamics follow the gradient descent direction of Sample Margin loss in Eq. (3.10). 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*

$$\begin{aligned}\boldsymbol{Z}(t) =&\Pi_1^+ \boldsymbol{Z}_0 \begin{pmatrix} a_1^+(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_1^+(t)\boldsymbol{I}_C \end{pmatrix} + \Pi_1^- \boldsymbol{Z}_0 \begin{pmatrix} a_1^-(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_1^-(t)\boldsymbol{I}_C \end{pmatrix} + \\ &\Pi_2^+ \boldsymbol{Z}_0 \begin{pmatrix} a_2^+(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_2^+(t)\boldsymbol{I}_C \end{pmatrix} + \Pi_2^- \boldsymbol{Z}_0 \begin{pmatrix} a_2^-(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_2^-(t)\boldsymbol{I}_C \end{pmatrix} + \\ &\Pi_3 \boldsymbol{Z}_0 \begin{pmatrix} a_3(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_3(t)\boldsymbol{I}_C \end{pmatrix}\end{aligned} \tag{A.29}$$

$$\boldsymbol{Z}(t) = \boldsymbol{Z}(t) \begin{pmatrix} \phi_1(t)\boldsymbol{I}_{CN} & 0 \\ 0 & \phi_2(t)\boldsymbol{I}_C \end{pmatrix}, \text{ and } \boldsymbol{b}(t) = \phi_3(t) \left(\boldsymbol{b}(0) - \tfrac{\gamma C - 1 - \gamma}{C} \psi(t) \mathbf{1}_C \right), \tag{A.30}$$

where $\boldsymbol{Z}(t)$ is defined in Eq. (3.5), $\phi_i(t) = e^{-\int_0^t \lambda_i(r)\mathrm{d}r}$ for $i \in \{1, 2, 3\}$, and $\psi(t) = \int_0^t \eta(s) e^{\int_0^s \lambda_3(r)\mathrm{d}r} \mathrm{d}s$.

Proof. According to the gradient flow in Eq. (3.11), and the notations in the proof of Theorem 3.1, we have:

$$\boldsymbol{Z}'(t) = \boldsymbol{Z}(t)\boldsymbol{A}(t) - \lambda \boldsymbol{Z}(t) \begin{pmatrix} \eta_1(t)\boldsymbol{I}_{CN} & 0 \\ 0 & \eta_2(t)\boldsymbol{I}_C \end{pmatrix}, \tag{A.31}$$

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

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

$$\begin{aligned}&[\boldsymbol{A}(t_1), \boldsymbol{A}(t_2)] \\ =&\boldsymbol{A}(t_1)\boldsymbol{A}(t_2) - \boldsymbol{A}(t_2)\boldsymbol{A}(t_1) \\ =&[\boldsymbol{A}(t_1) - \lambda\boldsymbol{\Lambda}(t_1)][\boldsymbol{A}(t_2) - \lambda\boldsymbol{\Lambda}(t_2)] - [\boldsymbol{A}(t_2) - \lambda\boldsymbol{\Lambda}(t_2)][\boldsymbol{A}(t_1) - \lambda\boldsymbol{\Lambda}(t_1)] \\ =&-\lambda[\boldsymbol{\Lambda}(t_1)\boldsymbol{A}(t_2) + \boldsymbol{A}(t_1)\boldsymbol{\Lambda}(t_2) - \boldsymbol{\Lambda}(t_2)\boldsymbol{A}(t_1) - \boldsymbol{A}(t_2)\boldsymbol{\Lambda}(t_1)] \\ =&-\lambda \begin{pmatrix} 0 & [\eta_1(t_1)\eta_2(t_2) - \eta_1(t_2)\eta_2(t_1)]\boldsymbol{M}^\top \\ [\eta_2(t_1)\eta_1(t_2) - \eta_2(t_2)\eta_1(t_2)]\boldsymbol{M} & 0 \end{pmatrix} \\ =&0\end{aligned} \tag{A.32}$$

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

$$\boldsymbol{Z}(t) = \boldsymbol{Z}_0 \exp\left(\int_0^t \boldsymbol{A}(\tau)\mathrm{d}\tau \right) = \boldsymbol{Z}_0 \exp \begin{pmatrix} -\lambda\zeta_1(t)\boldsymbol{I}_{CN} & \zeta_2(t)\boldsymbol{M}^\top \\ \zeta_1(t)\boldsymbol{M} & -\lambda\zeta_2(t)\boldsymbol{I}_C \end{pmatrix}, \tag{A.33}$$

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 $\boldsymbol{B} = \boldsymbol{B} - \lambda \boldsymbol{I}_{C(N+1)}$, we have

$$\boldsymbol{Z}(t) = \boldsymbol{Z}_0 \exp(\boldsymbol{B}\boldsymbol{C}(t)) = \boldsymbol{Z}_0 \sum_{k=0}^{\infty} \frac{(\boldsymbol{B}\boldsymbol{C}(t))^k}{k!}. \tag{A.34}$$

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 3.1, we have

$$\begin{aligned} &\Pi_1^\epsilon \boldsymbol{Z}_0 \boldsymbol{B} = \tfrac{\epsilon(1+\gamma)}{C\sqrt{N}} \Pi_1^\epsilon \boldsymbol{Z}_0, \\ &\Pi_2^\epsilon \boldsymbol{Z}_0 \boldsymbol{B} = \tfrac{\epsilon(1+\gamma-\gamma C)}{C\sqrt{N}} \Pi_2^\epsilon \boldsymbol{Z}_0, \\ &\Pi_3 \boldsymbol{Z}_0 \boldsymbol{B} = 0, \end{aligned} \quad \Rightarrow \quad \begin{aligned} &\Pi_1^\epsilon \boldsymbol{Z}_0 \boldsymbol{B} = \tfrac{\epsilon(1+\gamma)-\lambda C\sqrt{N}}{C\sqrt{N}} \Pi_1^\epsilon \boldsymbol{Z}_0, \\ &\Pi_2^\epsilon \boldsymbol{Z}_0 \boldsymbol{B} = \tfrac{\epsilon(1+\gamma-\gamma C)-\lambda C\sqrt{N}}{C\sqrt{N}} \Pi_2^\epsilon \boldsymbol{Z}_0, \\ &\Pi_3 \boldsymbol{Z}_0 \boldsymbol{B} = -\lambda \Pi_3 \boldsymbol{Z}_0. \end{aligned} \tag{A.35}$$

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

$$\boldsymbol{D}\boldsymbol{B} = \sigma \boldsymbol{D},\ \boldsymbol{H}\boldsymbol{M}^\top = \sigma \boldsymbol{W},\ \text{and}\ \boldsymbol{W}\boldsymbol{M} = \sigma \boldsymbol{H}. \tag{A.36}$$

In the following, we will prove that there exist two scalars $a(t)$ and $b(t)$, such that $\boldsymbol{D}\exp(\boldsymbol{B}\boldsymbol{C}(t)) = \boldsymbol{D}\begin{pmatrix} a(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b(t)\boldsymbol{I}_C \end{pmatrix}$.

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

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

$$\begin{aligned}
&\boldsymbol{D}(\boldsymbol{B}\boldsymbol{C}(t))^{n+1} \\
=&\boldsymbol{D}(\boldsymbol{B}\boldsymbol{C}(t))^n(\boldsymbol{B}\boldsymbol{C}(t)) \\
=&(a_n(t)\boldsymbol{H}, b_n(t)\boldsymbol{W})(\boldsymbol{B}\boldsymbol{C}(t)) \\
=&(a_n(t)\boldsymbol{H}, b_n(t)\boldsymbol{W})\begin{pmatrix} -\lambda\boldsymbol{I}_{CN} & \boldsymbol{M}^\top \\ \boldsymbol{M} & -\lambda\boldsymbol{I}_C \end{pmatrix}\begin{pmatrix} \zeta_1(t)\boldsymbol{I}_{CN} & 0 \\ 0 & \zeta_2(t)\boldsymbol{I}_C \end{pmatrix} \\
=&(b_n(t)\boldsymbol{W}\boldsymbol{M} - \lambda a_n(t)\boldsymbol{H}, a_n(t)\boldsymbol{H}\boldsymbol{M}^\top - \lambda b_n(t)\boldsymbol{W})\begin{pmatrix} \zeta_1(t)\boldsymbol{I}_{CN} & 0 \\ 0 & \zeta_2(t)\boldsymbol{I}_C \end{pmatrix} \\
=&(\zeta_1(t)(\sigma b_n(t) - \lambda a_n(t))\boldsymbol{H}, \zeta_2(t)(\sigma a_n(t) - \lambda b_n(t))\boldsymbol{H})
\end{aligned}, \tag{A.37}$$

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

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

$$\begin{pmatrix} a_k(t) \\ b_k(t) \end{pmatrix} = \begin{pmatrix} -\lambda\zeta_1(t) & \sigma\zeta_1(t) \\ \sigma\zeta_2(t) & -\lambda\zeta_2(t) \end{pmatrix}\begin{pmatrix} a_{k-1}(t) \\ b_{k-1}(t) \end{pmatrix} \quad \text{with} \quad \begin{pmatrix} a_0 \\ b_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{A.38}$$

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

Therefore, we have

$$\boldsymbol{D}\exp(\boldsymbol{B}\boldsymbol{C}(t)) = \boldsymbol{D}\sum_{k=0}^\infty \frac{(\boldsymbol{B}\boldsymbol{C}(t))^k}{k!} = \boldsymbol{D}\begin{pmatrix} a(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b(t)\boldsymbol{I}_C \end{pmatrix}, \tag{A.39}$$

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.*,

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

Next, we are going to derive a concrete expression of $\exp(\boldsymbol{S}(\sigma, \lambda, \zeta_1(t), \zeta_2(t)))$. Let the determinant $|\boldsymbol{S}(\sigma, \lambda, \zeta_1(t), \zeta_2(t) - \theta\boldsymbol{I}| = 0$, we can derive that the eigenvalues of $\boldsymbol{S}(\sigma, \lambda, \zeta_1(t), \zeta_2(t)$ are $\theta_1 = \frac{-\lambda(\zeta_1(t)+\zeta_2(t)) - \sqrt{\lambda^2(\zeta_1(t)-\zeta_2(t))^2 + 4\sigma^2\zeta_1(t)\zeta_2(t)}}{2} < 0$ and $\theta_2 = \frac{-\lambda(\zeta_1(t)+\zeta_2(t)) + \sqrt{\lambda^2(\zeta_1(t)-\zeta_2(t))^2 + 4\sigma^2\zeta_1(t)\zeta_2(t)}}{2} = \frac{(\sqrt{\lambda^2(s-1)^2 + 4s\sigma^2} - \lambda(s+1))\zeta_2(t)}{2}$, and the corresponding eigenvectors are

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

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

$$\boldsymbol{S}(\sigma, \lambda, \zeta_1(t), \zeta_2(t) = \boldsymbol{P}\begin{pmatrix} \theta_1 & 0 \\ 0 & \theta_2 \end{pmatrix}\boldsymbol{P}^{-1},\ \boldsymbol{P}^{-1} = \frac{\sigma\zeta_1(t)}{\theta_2 - \theta_1}\begin{pmatrix} \frac{\lambda\eta_1(t)+\theta_2}{\sigma\zeta_1(t)} & -1 \\ -\frac{\lambda\zeta_1(t)+\theta_1}{\sigma\zeta_1(t)} & 1 \end{pmatrix}, \tag{A.40}$$

and

$$\begin{aligned}
&\exp(\boldsymbol{S}(\sigma,\lambda,\zeta_1(t),\zeta_2(t))) = \boldsymbol{P}\begin{pmatrix} e^{\theta_1} & 0 \\ 0 & e^{\theta_2}\end{pmatrix}\boldsymbol{P}^{-1} \\
=&\frac{\sigma\zeta_1(t)}{\theta_2-\theta_1}\begin{pmatrix} 1 & 1 \\ \frac{\lambda\zeta_1(t)+\theta_1}{\sigma\zeta_1(t)} & \frac{\lambda\zeta_1(t)+\theta_2}{\sigma\zeta_1(t)}\end{pmatrix}\begin{pmatrix} e^{\theta_1} & 0 \\ 0 & e^{\theta_2}\end{pmatrix}\begin{pmatrix} \frac{\lambda\eta_1(t)+\theta_2}{\sigma\zeta_1(t)} & -1 \\ -\frac{\lambda\zeta_1(t)+\theta_1}{\sigma\zeta_1(t)} & 1\end{pmatrix} \\
=&\frac{1}{\theta_2-\theta_1}\begin{pmatrix} (\lambda\zeta_1(t)+\theta_2)e^{\theta_1}-(\lambda\zeta_1(t)+\theta_1)e^{\theta_2} & \sigma\zeta_1(t)e^{\theta_2}-\sigma\zeta_1(t)e^{\theta_1} \\ (\lambda\zeta_1(t)+\theta_2)e^{\theta_2}-(\lambda\zeta_1(t)+\theta_1)e^{\theta_1} & \sigma\zeta_2(t)e^{\theta_2}-\sigma\theta_2(t)e^{\theta_1}\end{pmatrix}
\end{aligned}, \tag{A.41}$$

thus

$$\begin{aligned}
a(t) &= \frac{e^{\theta_2}}{\theta_2-\theta_1}\left[(\sigma\zeta_1(t)-\lambda\zeta_1(t)-\theta_1)-(\sigma\zeta_1(t)-\lambda\zeta_1(t)-\theta_2)e^{\theta_1-\theta_2}\right], \\
b(t) &= \frac{e^{\theta_2}}{\theta_2-\theta_1}\left[(\lambda\zeta_1(t)+\theta_2+\sigma\zeta_2(t))+(\lambda\zeta_1(t)+\theta_1-\sigma\zeta_2(t))e^{\theta_1-\theta_2}\right].
\end{aligned} \tag{A.42}$$

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

$$\begin{aligned}
\boldsymbol{Z}(t) =&(\Pi_1^+\boldsymbol{Z}_0+\Pi_1^-\boldsymbol{Z}_0+\Pi_2^+\boldsymbol{Z}_0+\Pi_2^-\boldsymbol{Z}_0+\Pi_3\boldsymbol{Z}_0)\exp(\boldsymbol{L}(t)) \\
=&\Pi_1^+\boldsymbol{Z}_0\begin{pmatrix} a_1^+(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_1^+(t)\boldsymbol{I}_C\end{pmatrix}+\Pi_1^-\boldsymbol{Z}_0\begin{pmatrix} a_1^-(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_1^-(t)\boldsymbol{I}_C\end{pmatrix}+ \\
&\Pi_2^+\boldsymbol{Z}_0\begin{pmatrix} a_2^+(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_2^+(t)\boldsymbol{I}_C\end{pmatrix}+\Pi_2^-\boldsymbol{Z}_0\begin{pmatrix} a_2^-(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_2^-(t)\boldsymbol{I}_C\end{pmatrix}+ \\
&\Pi_3\boldsymbol{Z}_0\begin{pmatrix} a_3(t)\boldsymbol{I}_{CN} & 0 \\ 0 & b_3(t)\boldsymbol{I}_C\end{pmatrix},
\end{aligned} \tag{A.43}$$

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

$$\begin{aligned}
\begin{pmatrix} a_1^\epsilon(t) \\ b_1^\epsilon(t)\end{pmatrix} &= \exp\left(\boldsymbol{S}\left(\tfrac{\epsilon(1+\gamma)}{C\sqrt{N}},\lambda,\zeta_1(t),\zeta_2(t)\right)\right)\begin{pmatrix}1\\1\end{pmatrix}, \\
\begin{pmatrix} a_2^\epsilon(t) \\ b_2^\epsilon(t)\end{pmatrix} &= \exp\left(\boldsymbol{S}\left(\tfrac{\epsilon(1+\gamma-\gamma C)}{C\sqrt{N}},\lambda,\zeta_1(t),\zeta_2(t)\right)\right)\begin{pmatrix}1\\1\end{pmatrix}, \\
\begin{pmatrix} a_3(t) \\ b_3(t)\end{pmatrix} &= \exp\left(\boldsymbol{S}\left(0,\lambda,\zeta_1(t),\zeta_2(t)\right)\right)\begin{pmatrix}1\\1\end{pmatrix}.
\end{aligned} \tag{A.44}$$

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

$$\boldsymbol{b}(t)=\phi(t)\left(\boldsymbol{b}_0+\tfrac{1+\gamma-\gamma C}{C}\psi(t)\mathbf{1}_C\right), \tag{A.45}$$

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

A.5 Proof of Corollary 3.4

Corollary 3.4. *Under the conditions and notation of Theorem 3.3, 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 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 behaves as:*

$$\left\|\frac{\boldsymbol{H}(t)}{\|\boldsymbol{H}(t)\|}-\frac{\pi_h^+\boldsymbol{H}_1^++\pi_h^-\boldsymbol{H}_1^-}{\|\pi_h^+\boldsymbol{H}_1^++\pi_h^-\boldsymbol{H}_1^-\|}\right\|+\left\|\frac{\boldsymbol{W}(t)}{\|\boldsymbol{W}(t)\|}-\frac{\pi_w^+\boldsymbol{W}_1^++\pi_w^-\boldsymbol{W}_1^-}{\|\pi_w^+\boldsymbol{H}_1^++\pi_w^-\boldsymbol{H}_1^-\|}\right\|=O(e^{-\omega\zeta_2(t)}), \tag{A.46}$$

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 $\boldsymbol{S}(\sigma, \lambda, \zeta_1(t), \zeta_2(t))$ are $\theta_1 = -\frac{\lambda(s+1)+\sqrt{\lambda^2(s-1)^2+4s\sigma^2}}{2}\zeta_2(t) \to -\infty$ as $t \to \infty$ and $\theta_2 = \frac{(\sqrt{\lambda^2(s-1)^2+4s\sigma^2}-\lambda(s+1))}{2}\zeta_2(t)$.

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

$$\begin{aligned} a(t) &= \frac{e^{\omega_2\zeta_2(t)}}{\omega_2-\omega_1}\left[(s\sigma - s\lambda - \omega_1) - (s\sigma - s\lambda - \omega_2)e^{(\omega_1-\omega_2)\zeta_2(t)}\right] \\ &= \tfrac{s\sigma-s\lambda-\omega_1}{\omega_2-\omega_1}e^{\omega_2\zeta_2(t)} + O(e^{(\omega_1)\zeta_2(t)}), \\ b(t) &= \frac{e^{\omega_2\zeta_2(t)}}{\omega_2-\omega_1}\left[(s\lambda + \omega_2 + \sigma) + (s\lambda + \omega_1 - \sigma)e^{(\omega_1-\omega_2)\zeta_2(t)}\right] \\ &= \tfrac{s\lambda+\omega_2+\sigma}{\omega_2-\omega_1}e^{\omega_2\zeta_2(t)} + O(e^{\omega_1\zeta_2(t)}). \end{aligned} \tag{A.47}$$

Moreover, according to Theorem 3.3, we have

$$\boldsymbol{H}(t) = \sum_{\substack{\epsilon\in\{\pm\}\\ i\in\{1,2\}}} a_i^\epsilon(t)\boldsymbol{H}_i^\epsilon + a_3(t)\boldsymbol{H}_3, \quad \boldsymbol{W}(t) = \sum_{\substack{\epsilon\in\{\pm\}\\ i\in\{1,2\}}} b_i^\epsilon(t)\boldsymbol{W}_i^\epsilon + b_3(t)\boldsymbol{W}_3, \tag{A.48}$$

with

$$\begin{aligned} \begin{pmatrix} a_1^\epsilon(t) \\ b_1^\epsilon(t) \end{pmatrix} &= \exp\left(\boldsymbol{S}\left(\tfrac{\epsilon(1+\gamma)}{C\sqrt{N}}, \lambda, \zeta_1(t), \zeta_2(t)\right)\right)\begin{pmatrix}1\\1\end{pmatrix}, \\ \begin{pmatrix} a_2^\epsilon(t) \\ b_2^\epsilon(t) \end{pmatrix} &= \exp\left(\boldsymbol{S}\left(\tfrac{\epsilon(1+\gamma-\gamma C)}{C\sqrt{N}}, \lambda, \zeta_1(t), \zeta_2(t)\right)\right)\begin{pmatrix}1\\1\end{pmatrix}, \\ \begin{pmatrix} a_3(t) \\ b_3(t) \end{pmatrix} &= \exp\left(\boldsymbol{S}\left(0, \lambda, \zeta_1(t), \zeta_2(t)\right)\right)\begin{pmatrix}1\\1\end{pmatrix}. \end{aligned} \tag{A.49}$$

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 $\epsilon \in \{\pm 1\}$, thus we have

$$\begin{aligned} e^{-\omega_2\left(\frac{1+\gamma}{C\sqrt{N}}, \lambda, s\right)\zeta_2(t)}\boldsymbol{H}(t) &= \pi_h^+(\lambda, \gamma, s, C, N)\boldsymbol{H}_1^+ + \pi_h^-(\lambda, \gamma, s, C, N)\boldsymbol{H}_1^- + \boldsymbol{\Delta}_1, \\ e^{-\omega_2\left(\frac{1+\gamma}{C\sqrt{N}}, \lambda, s\right)\zeta_2(t)}\boldsymbol{W}(t) &= \pi_w^+(\lambda, \gamma, s, C, N)\boldsymbol{W}_1^+ + \pi_w^-(\lambda, \gamma, s, C, N)\boldsymbol{W}_1^- + \boldsymbol{\Delta}_2, \end{aligned} \tag{A.50}$$

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

$$\begin{aligned} \pi_h^+(\lambda, \gamma, s, C, N) &= \frac{s\frac{1+\gamma}{C\sqrt{N}} - s\lambda - \omega_1\left(\frac{1+\gamma}{C\sqrt{N}}, \lambda, s\right)}{\omega_2\left(\frac{1+\gamma}{C\sqrt{N}}, \lambda, s\right) - \omega_1\left(\frac{1+\gamma}{C\sqrt{N}}, \lambda, s\right)}, \\ \pi_h^-(\lambda, \gamma, s, C, N) &= \frac{-s\frac{1+\gamma}{C\sqrt{N}} - s\lambda - \omega_1\left(\frac{1+\gamma}{C\sqrt{N}}, \lambda, s\right)}{\omega_2\left(\frac{1+\gamma}{C\sqrt{N}}, \lambda, s\right) - \omega_1\left(\frac{1+\gamma}{C\sqrt{N}}, \lambda, s\right)}, \\ \pi_w^+(\lambda, \gamma, s, C, N) &= \frac{s\lambda + \omega_2\left(\frac{1+\gamma}{C\sqrt{N}}, \lambda, s\right) + \frac{1+\gamma}{C\sqrt{N}}}{\omega_2\left(\frac{1+\gamma}{C\sqrt{N}}, \lambda, s\right) - \omega_1\left(\frac{1+\gamma}{C\sqrt{N}}, \lambda, s\right)}, \\ \pi_w^-(\lambda, \gamma, s, C, N) &= \frac{s\lambda + \omega_2\left(\frac{1+\gamma}{C\sqrt{N}}, \lambda, s\right) - \frac{1+\gamma}{C\sqrt{N}}}{\omega_2\left(\frac{1+\gamma}{C\sqrt{N}}, \lambda, s\right) - \omega_1\left(\frac{1+\gamma}{C\sqrt{N}}, \lambda, s\right)}. \end{aligned} \tag{A.51}$$

Therefore, we have

$$\begin{aligned}\lim_{t\to\infty}\frac{\boldsymbol{H}(t)}{\|\boldsymbol{H}(t)\|}&=\frac{\pi_h^+(\lambda,\gamma,s,C,N)\boldsymbol{H}_1^+ + \pi_h^-(\lambda,\gamma,s,C,N)\boldsymbol{H}_1^-}{\|\pi_h^+(\lambda,\gamma,s,C,N)\boldsymbol{H}_1^+ + \pi_h^-(\lambda,\gamma,s,C,N)\boldsymbol{H}_1^-\|},\\ \lim_{t\to\infty}\frac{\boldsymbol{W}(t)}{\|\boldsymbol{W}(t)\|}&=\frac{\pi_w^+(\lambda,\gamma,s,C,N)\boldsymbol{W}_1^+ + \pi_w^-(\lambda,\gamma,s,C,N)\boldsymbol{W}_1^-}{\|\pi_w^+(\lambda,\gamma,s,C,N)\boldsymbol{W}_1^+ + \pi_w^-(\lambda,\gamma,s,C,N)\boldsymbol{W}_1^-\|},\end{aligned} \tag{A.52}$$

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

$$\lim_{t\to\infty}\boldsymbol{H}(t)=\boldsymbol{H}_1^+ + \frac{1-s}{1+s}\boldsymbol{H}_1^-,\ \lim_{t\to\infty}\boldsymbol{W}(t)=\boldsymbol{W}_1^+ - \frac{1-s}{1+s}\boldsymbol{W}_1^-. \tag{A.53}$$

- 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. □

A.6 Proof of Theorem 3.5

Lemma A.2. *For $\boldsymbol{h}(t),\boldsymbol{w}\in\mathbb{R}^p$, $\eta(t)>0$, let $\widehat{\boldsymbol{v}}=\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}$ denote the ℓ_2-normalized vector of $\boldsymbol{v}$, considering the discrete dynamical system $\boldsymbol{h}(t+1)=\boldsymbol{h}(t)+\frac{\eta(t)}{\|\boldsymbol{h}(t)\|_2}\left(\boldsymbol{I}_p-\widehat{\boldsymbol{h}}(t)\widehat{\boldsymbol{h}}^\top(t)\right)\boldsymbol{w}$, if $\widehat{\boldsymbol{w}}^\top\widehat{\boldsymbol{h}}(0)>-1$, 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*

$$\lim_{t\to\infty}\left\|\frac{\boldsymbol{h}(t)}{\|\boldsymbol{h}(t)\|_2}-\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}\right\|=0. \tag{A.54}$$

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 $\beta_t=\widehat{\boldsymbol{w}}^\top\widehat{\boldsymbol{h}}(t)$ denote the cosine similarity between $\widehat{\boldsymbol{w}}$ and $\widehat{\boldsymbol{h}}(t)$, 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 $\left(\boldsymbol{I}_p-\widehat{\boldsymbol{h}}(t)\widehat{\boldsymbol{h}}^\top(t)\right)\boldsymbol{w}$, thus we have

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

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 $\boldsymbol{w}$:

$$\boldsymbol{h}(t+1)=\left(1-\frac{\eta(t)\boldsymbol{w}^\top\widehat{\boldsymbol{h}}(t)}{\|\boldsymbol{h}(t)\|_2}\right)\boldsymbol{h}(t)+\frac{\eta(t)}{\|\boldsymbol{h}(t)\|_2}\boldsymbol{w}, \tag{A.55}$$

so $\widehat{\boldsymbol{w}}^\top\widehat{\boldsymbol{h}}(t+1)\geq\widehat{\boldsymbol{w}}^\top\widehat{\boldsymbol{h}}(t)$, *i.e.*, β_t *is monotonically increasing*, which is based on the facts that $1-\frac{\eta(t)\boldsymbol{w}^\top\widehat{\boldsymbol{h}}(t)}{\|\boldsymbol{h}(t)\|_2}\geq 1-\frac{\eta(t)\|\boldsymbol{w}\|_2}{\|\boldsymbol{h}(t)\|_2}\geq 1-\frac{\eta(0)\|\boldsymbol{w}\|_2}{\|\boldsymbol{h}(0)\|_2}>0$, $\frac{\eta(t)}{\|\boldsymbol{h}(t)\|_2}>0$, 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 Eq. (A.55) as follows:

$$\begin{aligned}
\beta_{t+1} &= \frac{\boldsymbol{w}^\top \boldsymbol{h}(t+1)}{\|\boldsymbol{w}\|_2 \|\boldsymbol{h}(t+1)\|_2} \\
&= \frac{\boldsymbol{w}^\top \left(\boldsymbol{h}(t) + \frac{\eta(t)}{\|\boldsymbol{h}(t)\|_2}\left(\boldsymbol{I}_p - \widehat{\boldsymbol{h}}(t)\widehat{\boldsymbol{h}}^\top(t)\right)\boldsymbol{w}\right)}{\|\boldsymbol{w}\|_2 \sqrt{\|\boldsymbol{h}(t)\|_2^2 + \frac{\eta^2(t)}{\|\boldsymbol{h}(t)\|_2^2}\left\|\left(\boldsymbol{I}_p - \widehat{\boldsymbol{h}}(t)\widehat{\boldsymbol{h}}^\top(t)\right)\boldsymbol{w}\right\|_2^2}} \\
&= \frac{\beta_t + \frac{\eta(t)\|\boldsymbol{w}\|_2}{\|\boldsymbol{h}(t)\|_2^2}(1-\beta_t^2)}{\sqrt{1 + \frac{\eta^2(t)\|\boldsymbol{w}\|_2^2}{\|\boldsymbol{h}(t)\|_2^4}(1-\beta_t^2)}} \\
&= \frac{\beta_t + \alpha_t(1-\beta_t^2)}{\sqrt{1+\alpha_t^2(1-\beta_t^2)}}, \\
\alpha_{t+1} &= \frac{\eta(t+1)\|\boldsymbol{w}\|_2}{\|\boldsymbol{h}(t+1)\|_2^2} \\
&= \frac{\xi_t \eta(t)\|\boldsymbol{w}\|_2}{\|\boldsymbol{h}(t)\|_2^2 + \frac{\eta^2(t)}{\|\boldsymbol{h}(t)\|_2^2}\left\|\left(\boldsymbol{I}_p - \widehat{\boldsymbol{h}}(t)\widehat{\boldsymbol{h}}^\top(t)\right)\boldsymbol{w}\right\|_2^2} \\
&= \frac{\xi_t \alpha_t}{1+\alpha_t^2(1-\beta_t^2)},
\end{aligned} \tag{A.56}$$

with $\beta_0 = \widehat{\boldsymbol{w}}^\top \widehat{\boldsymbol{h}}(0) > -1$, $\alpha_0 = \frac{\eta(0)\|\boldsymbol{w}\|_2}{\|\boldsymbol{h}(0)\|_2^2} > 0$, $\xi_t \leq 1$ 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

$$\lim_{t\to\infty} \alpha_{t+1} = \lim_{t\to\infty} \frac{\xi_t \alpha_t}{1+\alpha_t^2(1-\beta_t^2)} \quad \Rightarrow \quad a = \frac{a}{1+\lim_{t\to\infty}\alpha_t^2(1-\beta_t^2)}, \tag{A.57}$$

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 A.3. *For the discrete dynamical system in Eq. (A.56), 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 Eq. (A.56), 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 $\forall t > \tau$,

$$\begin{aligned}
1-\beta_{t+1}^2 &= \alpha_{t+1}\cdot\frac{1-\beta_0^2}{\alpha_0}\prod_{i=0}^{t}\frac{(1-\alpha_i\beta_i)^2}{\xi_i} \\
&= \alpha_{t+1}\cdot\frac{\eta(0)(1-\beta_0^2)}{\alpha_0\eta(t+1)}\cdot\prod_{i=0}^{\tau}(1-\alpha_i\beta_i)^2\cdot\prod_{i=\tau+1}^{t}(1-\alpha_i\beta_i)^2 \qquad \text{(A.58)} \\
&\leq \alpha_{t+1}\cdot\frac{\eta(0)(1-\beta_0^2)}{\varepsilon\alpha_0}\cdot\prod_{i=0}^{\tau}(1-\alpha_i\beta_i)^2,
\end{aligned}$$

where the inequality is based on the fact that $1-\alpha_i\beta_i \in (0,1]$ when $0 \leq \beta_0 \leq \beta_i \leq 1$, $\eta(t+1) \geq \varepsilon$, and $\alpha_i \leq 1$ 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

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

as $\lim_{t\to\infty}\alpha_{t+1} = 0$. This reveals $\lim_{t\to\infty}\beta_t^2 = 1$. Furthermore, since $\beta_t \geq 0$, 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 A.3, we just need prove that $\exists\tau > 0$, *s.t.*, $\beta_\tau \geq 0$.

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

$$\alpha_t + \beta_t - \alpha_t\beta_t^2 < 0,\ \forall t \geq 0, \qquad \text{(A.59)}$$

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

$$\alpha_t < \frac{\varepsilon}{\eta(0)}\alpha_0(1-\beta_0^2),\ \forall t \geq t'. \qquad \text{(A.60)}$$

According to the iterations in Eq. (A.56), 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 $t \geq t'$,

$$\alpha_t \geq \alpha_t(1-\beta_t^2) \geq \xi_{t-1}\alpha_{t-1}(1-\beta_{t-1}^2) \geq \ldots \geq \alpha_0(1-\beta_0^2)\prod_{i=0}^{t-1}\xi_i \geq \frac{\varepsilon}{\eta(0)}\alpha_0(1-\beta_0^2), \qquad \text{(A.61)}$$

which contradicts the fact in Eq. (A.60). 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 A.3.

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

Theorem 3.5 *Considering the discrete dynamics in Eq. (3.16), if* $\forall i \in [N], c \in [C]$, $\widehat{\boldsymbol{w}}_c^\top\widehat{\boldsymbol{h}}_{i,c}(0) > -1$, *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*

$$\lim_{t\to\infty}\left\|\widehat{\boldsymbol{H}}(t) - \widehat{\boldsymbol{W}}(\boldsymbol{I}_C\otimes\boldsymbol{1}_N^\top)\right\| = 0, \qquad \text{(A.62)}$$

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:*

$$\left\|\widehat{\boldsymbol{H}}(t)-\widehat{\boldsymbol{W}}(\boldsymbol{I}_C\otimes\mathbf{1}_N^\top)\right\|\leq O(e^{-\mu t}). \tag{A.63}$$

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

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

$$\boldsymbol{h}_{i,c}(t+1)=\boldsymbol{h}_{i,c}(t)+\frac{(1+\gamma)\eta(t)}{CN\|\boldsymbol{h}_{i,c}(t)\|_2}\left(\boldsymbol{I}_p-\widehat{\boldsymbol{h}}_{i,c}(t)\widehat{\boldsymbol{h}}_{i,c}^\top(t)\right)\boldsymbol{w}_c. \tag{A.64}$$

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

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 A.2, $\forall i,c$, we have $\lim_{t\to\infty}\widehat{\boldsymbol{w}}_c^\top\widehat{\boldsymbol{h}}_{i,c}(t)=1$, then for a given constant $\delta>0$, $\exists\tau>0$, *s.t.*, $\forall t>\tau$, $\widehat{\boldsymbol{w}}_c^\top\widehat{\boldsymbol{h}}_{i,c}(\tau)\geq\delta$. Consider $t>\tau$, we have

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

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

Let $c_1=\max_{i,c}(1-(\widehat{\boldsymbol{w}}_c^\top\widehat{\boldsymbol{h}}_{i,c}(0))^2)\prod_{j=0}^{\tau}\left(1-\frac{(1+\gamma)\eta(j)\|\boldsymbol{w}_c\|_2}{CN\|\boldsymbol{h}_{i,c}(j)\|_2^2}\cdot\widehat{\boldsymbol{w}}_c^\top\widehat{\boldsymbol{h}}_{i,c}(j)\right)^2\left(1-\frac{(1+\gamma)\varepsilon\delta\|\boldsymbol{w}_c\|_2}{CNL^2}\right)^{-2\tau}$, and $\mu=\min_c -2\log\left(1-\frac{(1+\gamma)\varepsilon\delta\|\boldsymbol{w}_c\|_2}{CNL^2}\right)$, then $1-\widehat{\boldsymbol{w}}_c^\top\widehat{\boldsymbol{h}}_{i,c}(t+1)\leq\frac{c_1e^{-\mu t}}{1+\delta}$.

Therefore, we have

$$\left\|\widehat{\boldsymbol{H}}(t)-\widehat{\boldsymbol{W}}(\boldsymbol{I}_C\otimes\mathbf{1}_N^\top)\right\|_2^2=2\sum_{i,c}\left(1-\widehat{\boldsymbol{w}}_c^\top\widehat{\boldsymbol{h}}_{i,c}(t+1)\right)\leq\frac{2c_1CNe^{-\mu t}}{(1+\delta)}, \tag{A.65}$$

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

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

$$\boldsymbol{h}_{i,c}(t+1)=\boldsymbol{h}_{i,c}(t)+\frac{(1+\gamma)\eta(t)}{CN\|\boldsymbol{h}_{i,c}(t)\|_2}\left(\boldsymbol{I}_p-\widehat{\boldsymbol{h}}_{i,c}(t)\widehat{\boldsymbol{h}}_{i,c}^\top(t)\right)\boldsymbol{w}_c=\boldsymbol{h}_{i,c}(t), \tag{A.66}$$

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

A.7 PROOF OF THEOREM 4.1

Theorem 4.1 *Consider the continual gradient flow (Equation (4.1)) in which the prototypes $\boldsymbol{W}$ is fixed, we have the closed-form dynamics:*

$$\boldsymbol{H}(t) = e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\boldsymbol{H}(0) + \frac{1 - e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}}{\lambda}\boldsymbol{W}\boldsymbol{M}, \tag{A.67}$$

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 Eq. (4.1), the solution is

$$\begin{aligned}
\boldsymbol{H}(t) =& e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\left(\boldsymbol{H}(0) + \int_0^t \eta(s)e^{\lambda \int_0^s \eta(\tau)\mathrm{d}\tau}\mathrm{d}s\boldsymbol{W}\boldsymbol{M}\right) \\
=& e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\boldsymbol{H}(0) + e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\int_0^t \frac{1}{\lambda}\mathrm{d}e^{\lambda \int_0^s \eta(\tau)\mathrm{d}\tau}\boldsymbol{W}\boldsymbol{M} \\
=& e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\boldsymbol{H}(0) + e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\frac{e^{\lambda \int_0^s \eta(\tau)\mathrm{d}\tau}}{\lambda}\Big|_0^t \boldsymbol{W}\boldsymbol{M} \\
=& e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}\boldsymbol{H}(0) + \frac{1 - e^{-\lambda \int_0^t \eta(\tau)\mathrm{d}\tau}}{\lambda}\boldsymbol{W}\boldsymbol{M},
\end{aligned} \tag{A.68}$$

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

A.8 THE PROJECTIONS ONTO $\mathcal{E}_1^+$, $\mathcal{E}_1^-$, $\mathcal{E}_2^+$, $\mathcal{E}_2^-$ AND $\mathcal{E}_3$

Lemma A.4. *Let $\mathcal{S}$ denote the subspace $\mathcal{S} = \{\boldsymbol{W} : \boldsymbol{W}\boldsymbol{1}_n = 0, \boldsymbol{W} \in \mathbb{R}^{m\times n}\}$, then the projection of a point $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ onto $\mathcal{S}$ can be denoted as $\Pi_{\mathcal{S}}\boldsymbol{A} = \boldsymbol{A}\left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n^\top\right)$.*

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

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

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.*, $\boldsymbol{W} = \boldsymbol{A}\left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n^\top\right)$. Therefore, the projection of $\boldsymbol{A}$ onto $\mathcal{S}$ is $\Pi_{\mathcal{S}}\boldsymbol{A} = \arg\min_{\boldsymbol{W}\in\mathcal{S}} \|\boldsymbol{W} - \boldsymbol{A}\|_F^2 = \boldsymbol{A}\left(\boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n^\top\right)$. □

Lemma A.5. *Let $\mathcal{S}$ denote the subspace $\{\boldsymbol{W} : \boldsymbol{W}(\boldsymbol{I}_c \otimes \boldsymbol{1}_n) = 0, \boldsymbol{W} \in \mathbb{R}^{m\times cn}\}$, then the projection of a point $\boldsymbol{A} \in \mathbb{R}^{m\times cn}$ onto $\mathcal{S}$ can be denoted as $\Pi_{\mathcal{S}}\boldsymbol{A} = \boldsymbol{A}\left(\boldsymbol{I}_{cn} - \frac{1}{n}\boldsymbol{I}_c \otimes \boldsymbol{1}_n\boldsymbol{1}_n^\top\right)$.*

Proof. The proof is similar to Lemma A.4. 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

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

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.*, $\boldsymbol{W} = \boldsymbol{A}(\boldsymbol{I}_{cn} - \frac{1}{n}\boldsymbol{I}_c \otimes \boldsymbol{1}_n\boldsymbol{1}_n^\top)$. Therefore, the projection of $\boldsymbol{A}$ onto $\mathcal{S}$ is $\Pi_{\mathcal{S}}\boldsymbol{A} = \boldsymbol{A}\left(\boldsymbol{I}_{cn} - \frac{1}{n}\boldsymbol{I}_c \otimes \boldsymbol{1}_n\boldsymbol{1}_n^\top\right)$. □

Lemma A.6. *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*

$$\Pi_1^\epsilon(\boldsymbol{H}, \boldsymbol{W}) = (\tfrac{\epsilon}{\sqrt{N}}(\boldsymbol{P} \otimes \boldsymbol{1}_N^\top), \boldsymbol{P}), \tag{A.69}$$

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

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

$$\begin{aligned}
&\arg\min_{\boldsymbol{Z}\in\mathcal{S}} \|\tfrac{\epsilon}{\sqrt{N}}(\boldsymbol{Z}\otimes\mathbf{1}_N^\top) - \boldsymbol{H}\|_F^2 + \|\boldsymbol{Z}-\boldsymbol{W}\|_F^2 \\
=&\arg\min_{\boldsymbol{Z}\in\mathcal{S}} \sum_{i=1}^N \|\tfrac{1}{\sqrt{N}}\boldsymbol{Z} - \epsilon\boldsymbol{H}_i\|_F^2 + \|\boldsymbol{Z}-\boldsymbol{W}\|_F^2 \\
=&\arg\min_{\boldsymbol{Z}\in\mathcal{S}} \|\boldsymbol{Z}\|_F^2 - \tfrac{2\epsilon}{\sqrt{N}}\sum_{i=1}^N\langle\boldsymbol{Z},\boldsymbol{H}_i\rangle + \|\boldsymbol{H}\|_F^2 + \|\boldsymbol{Z}-\boldsymbol{W}\|_F^2 \\
=&\arg\min_{\boldsymbol{Z}\in\mathcal{S}} \|\boldsymbol{Z} - \tfrac{\epsilon}{\sqrt{N}}\sum_{i=1}^N\boldsymbol{H}_i\|_F^2 - \|\frac{1}{\sqrt{N}}\sum_{i=1}^N\boldsymbol{H}_i\|_F^2 + \|\boldsymbol{H}\|_F^2 + \|\boldsymbol{Z}-\boldsymbol{W}\|_F^2 \\
=&\arg\min_{\boldsymbol{Z}\in\mathcal{S}} \|\boldsymbol{Z} - \tfrac{\epsilon}{\sqrt{N}}\boldsymbol{H}(\boldsymbol{I}_C\otimes\mathbf{1}_N)\|_F^2 + \|\boldsymbol{Z}-\boldsymbol{W}\|_F^2 + \|\boldsymbol{H}\|_F^2 - \|\tfrac{\epsilon}{\sqrt{N}}\boldsymbol{H}(\boldsymbol{I}_C\otimes\mathbf{1}_N)\|_F^2 \\
=&\arg\min_{\boldsymbol{Z}\in\mathcal{S}} \|\boldsymbol{Z} - \tfrac{1}{2}(\tfrac{\epsilon}{\sqrt{N}}\boldsymbol{H}(\boldsymbol{I}_C\otimes\mathbf{1}_N) + \boldsymbol{W})\|_F^2 \\
=&\tfrac{1}{2}(\tfrac{\epsilon}{\sqrt{N}}\boldsymbol{H}(\boldsymbol{I}_C\otimes\mathbf{1}_N) + \boldsymbol{W})(\boldsymbol{I}_C - \tfrac{1}{C}\mathbf{1}_C\mathbf{1}_C^\top).
\end{aligned} \tag{A.70}$$

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 A.7. *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*

$$\Pi_2^\epsilon(\boldsymbol{H},\boldsymbol{W}) = (\tfrac{\epsilon}{\sqrt{N}}\boldsymbol{h}\mathbf{1}_{CN}^\top, \boldsymbol{h}\mathbf{1}_C^\top), \tag{A.71}$$

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

Proof. We have

$$\begin{aligned}
&\arg\min_{\boldsymbol{h}\in\mathbb{R}^p} \|\tfrac{\epsilon}{\sqrt{N}}\boldsymbol{h}\mathbf{1}_{CN}^\top - \boldsymbol{H}\|_F^2 + \|\boldsymbol{h}\mathbf{1}_C^\top - \boldsymbol{W}\|_F^2 \\
=&\arg\min_{\boldsymbol{h}\in\mathbb{R}^p} CN\|\tfrac{1}{\sqrt{N}}\boldsymbol{h} - \tfrac{\epsilon}{CN}\boldsymbol{H}\mathbf{1}_{CN}\|_2^2 + C\|\boldsymbol{h} - \tfrac{1}{C}\boldsymbol{W}\mathbf{1}_C\|_2^2 \\
=&\arg\min_{\boldsymbol{h}\in\mathbb{R}^p} \|\boldsymbol{h} - \tfrac{\epsilon}{C\sqrt{N}}\boldsymbol{H}\mathbf{1}_{CN}\|_2^2 + \|\boldsymbol{h} - \tfrac{1}{C}\boldsymbol{W}\mathbf{1}_C\|_2^2 \\
=&\arg\min_{\boldsymbol{h}\in\mathbb{R}^p} \|\boldsymbol{h} - \tfrac{1}{2C}(\tfrac{\epsilon}{\sqrt{N}}\boldsymbol{H}\mathbf{1}_{CN} + \boldsymbol{W}\mathbf{1}_C)\|_2^2 \\
=&\tfrac{1}{2C}(\tfrac{\epsilon}{\sqrt{N}}\boldsymbol{H}\mathbf{1}_{CN} + \boldsymbol{W}\mathbf{1}_C)
\end{aligned} \tag{A.72}$$

□

Lemma A.8. *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*

$$\Pi_3(\boldsymbol{H},\boldsymbol{W}) = (\boldsymbol{H}(\boldsymbol{I}_{CN} - \tfrac{1}{N}\boldsymbol{I}_C\otimes\mathbf{1}_N\mathbf{1}_N^\top), 0). \tag{A.73}$$

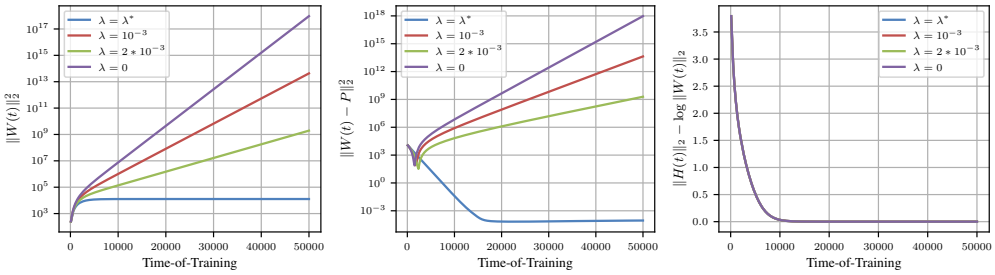
Proof. This can be easily derived by Lemma A.5. □

B EXPERIMENTS

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

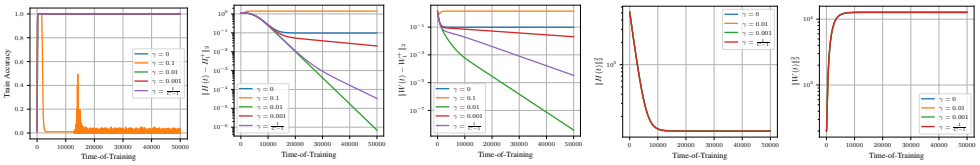
B.1 NUMERICAL EXPERIMENTS

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) $\log \|\boldsymbol{W}(t)\|_2$ (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 Averaged Sample Margin 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 A.6. (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 3.4), $\boldsymbol{W}(t)$ 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.



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

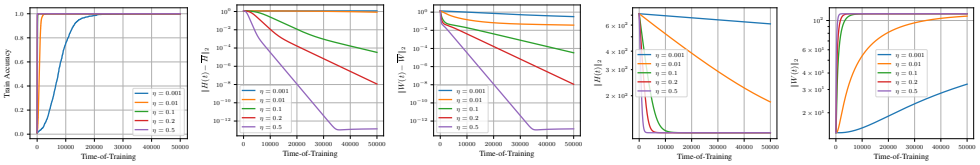
Figure 5: Verification of the behavior of regularized gradient descent iterates in Equation (3.11) with $\gamma \in \{0, 0.1, 0.01, 0.001, \frac{1}{C-1}\}$. We set $p = 512$, $C = 100$, $N = 10$, $\lambda = \frac{(1+\gamma)}{C\sqrt{N}}$, $\eta_1(t) = \eta_2(t) = 0.1$ (*i.e.*, $s = 1$), thus we have $\lim_{t\to\infty} \boldsymbol{Z}(t) = \Pi_1^+ \boldsymbol{Z}_0$, according to Corollary 3.4, 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 3.4, 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.

B.2 Visual Classification

For classification experiments in Fig. 2, Fig. 12, Fig. 13, Fig. 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

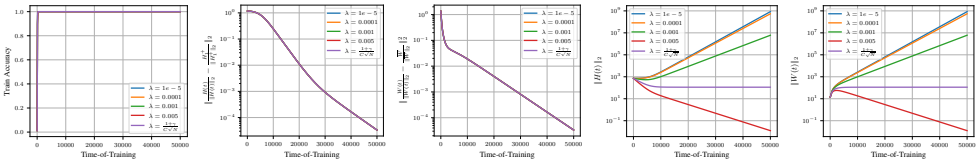
Table 2: Test accuracies on imbalanced CIFAR-10 under different explicit feature regularization.

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**



(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 (3.11) 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.



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

Figure 7: Verification of the behavior of regularized gradient descent iterates in Equation (3.14) 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 3.4. (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 3.4, $\boldsymbol{W}(t)$ 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.

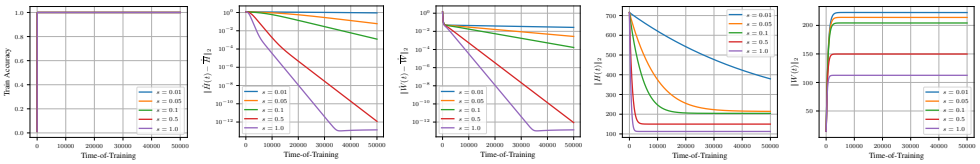
learning rate annealing Loshchilov and Hutter (2017) with $T_{\max}$ being the corresponding epochs. The initial learning rate is set to 0.1, weight decay is set to 5×10^{-4}, and batch size is set to 256. Typical data augmentations including random width/height shift and horizontal flip are applied. Moreover, to use the PAL and FNPAL (Zhou et al., 2022b) that anchors prototypes with a neural collapse solution, we remove the ReLU layer before the linear classifier in the last layer.

B.3 Imbalanced Classification

For the experiments of imbalanced learning in Tab. 1, Tab. 2, Tab. 3, Fig. 10, and Fig. 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., 2022c). 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 denote the ratio between sample sizes of the most frequent and least frequent class. Moreover, long-tailed imbalance (Cui et al., 2019) that utilizes

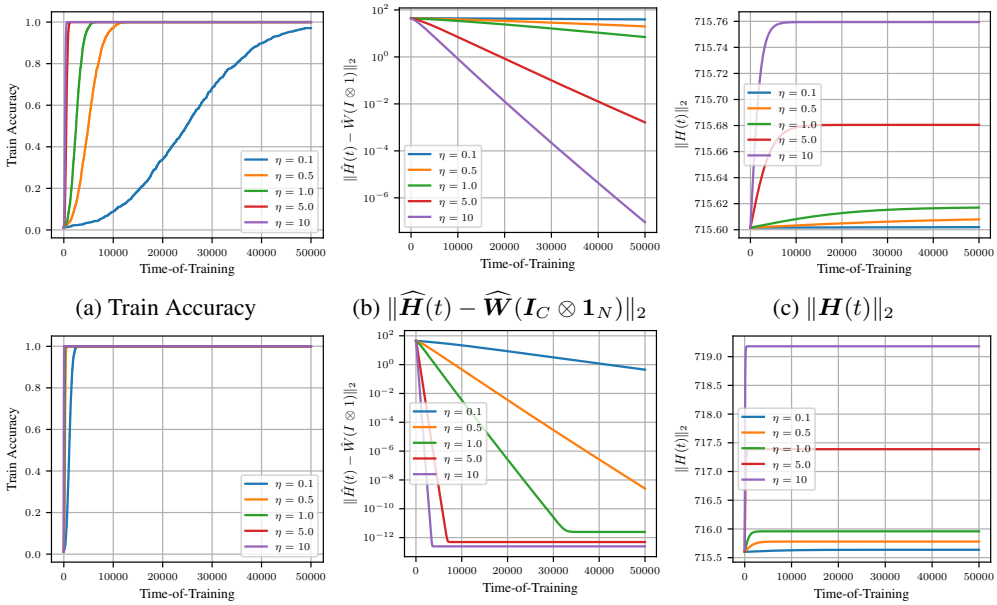
Table 3: Test accuracies on imbalanced CIFAR-100 under different explicit feature regularization.

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) $\|\widehat{\boldsymbol{H}}(t) - \widehat{\boldsymbol{H}}_1^+\|_2$ (c) $\|\widehat{\boldsymbol{W}}(t) - \widehat{\boldsymbol{W}}_1^+\|_2$ (d) $\|\boldsymbol{H}(t)\|_2$ (e) $\|\boldsymbol{W}(t)\|_2$

Figure 8: Verification of the behavior of regularized gradient descent iterates in Equation (3.14) 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 3.4. As can be seen, larger scale parameter s can achieve faster convergence speed.



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

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

Figure 9: Verification of the behavior of discrete gradient descent iterates in Equation (3.15) 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 3.5, the feature norm $\|\boldsymbol{H}(t)\|_2$ is non-decreasing, and the error $\|\widehat{\boldsymbol{H}}(t) - \widehat{\boldsymbol{W}}(\boldsymbol{I}_C \otimes \mathbf{1}_N^\top)\|_2$ shows exponential decrease.

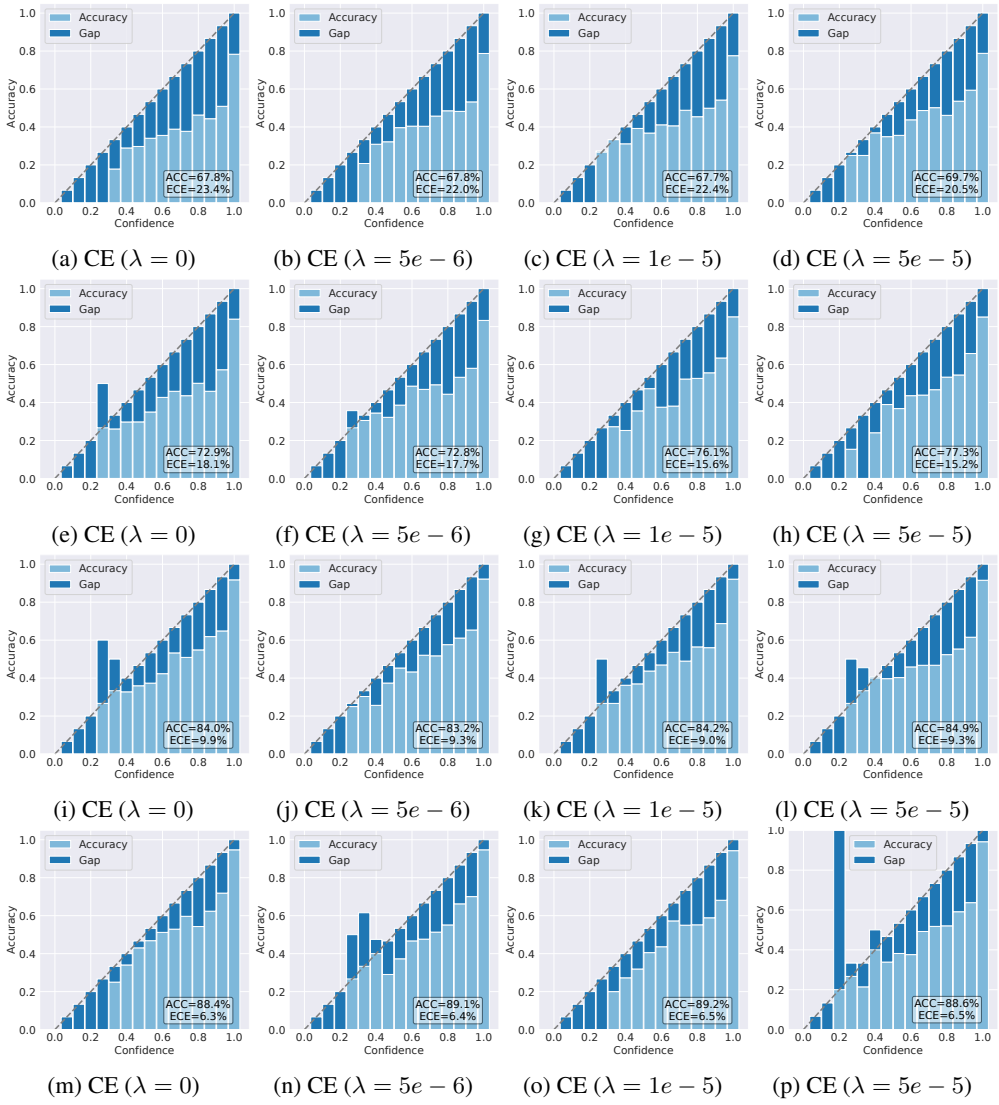
an exponential decay in samples sizes and step imbalance (Buda et al., 2018)(that sets all minority classes to have the same number of samples, as do all majority classes) are considered.

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*,

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

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, $\text{acc}(\cdot)$ and $\text{conf}(\cdot)$ denote the accuracy and predicted confidence of $\mathcal{S}_b$, respectively.

As shown in Tab. 2 and Tab. 3, explicit feature regularization can improve imbalanced learning on CIFAR-10/-100 in most cases.



(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 ($\lambda = 5e-5$)

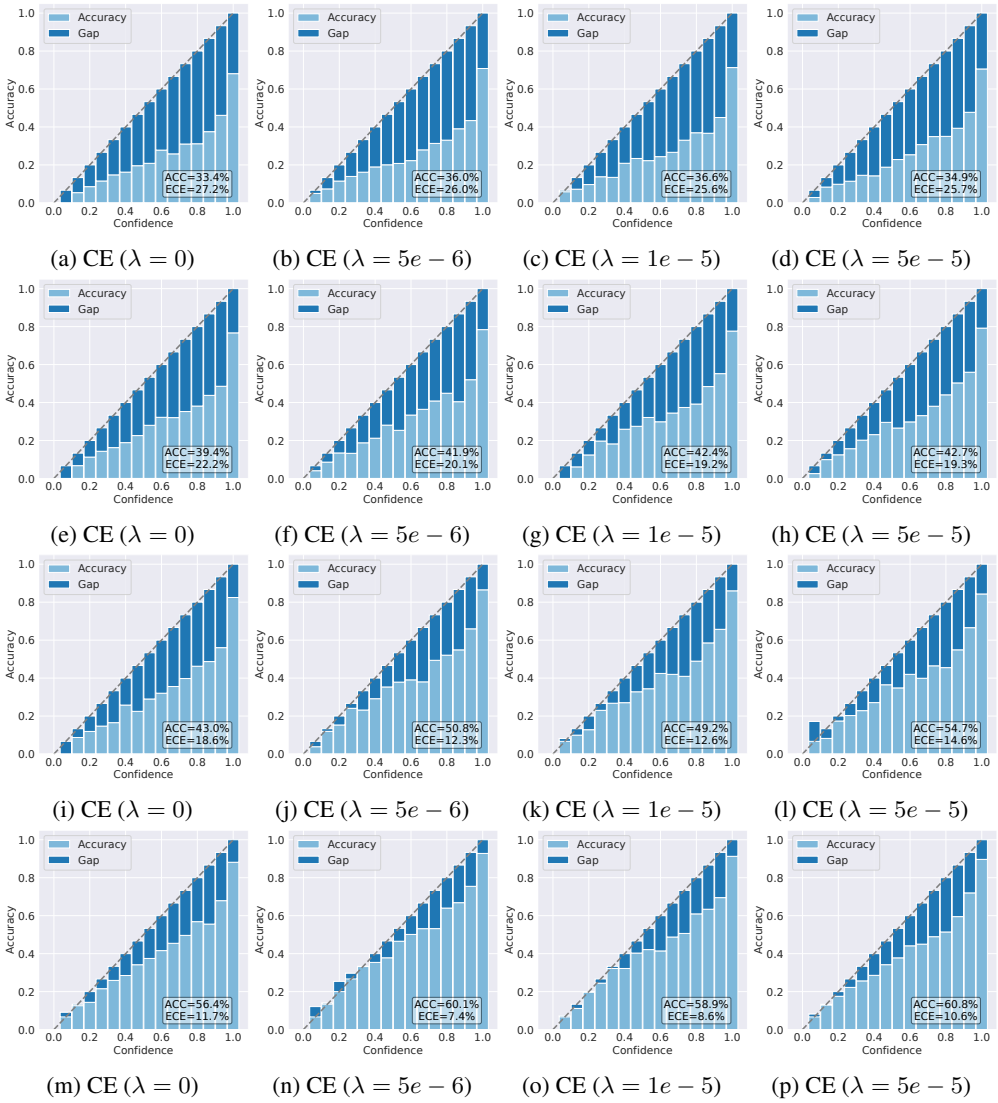
(i) CE ($\lambda = 0$) (j) CE ($\lambda = 5e-6$) (k) CE ($\lambda = 1e-5$) (l) CE ($\lambda = 5e-5$)

(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

B.4 Out-of-Distribution Detection

For the experiments of OOD detection in Fig. 3, Tab. 4, Fig. 16, Fig. 17, and Fig. 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) 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 ($\lambda = 5e-5$)

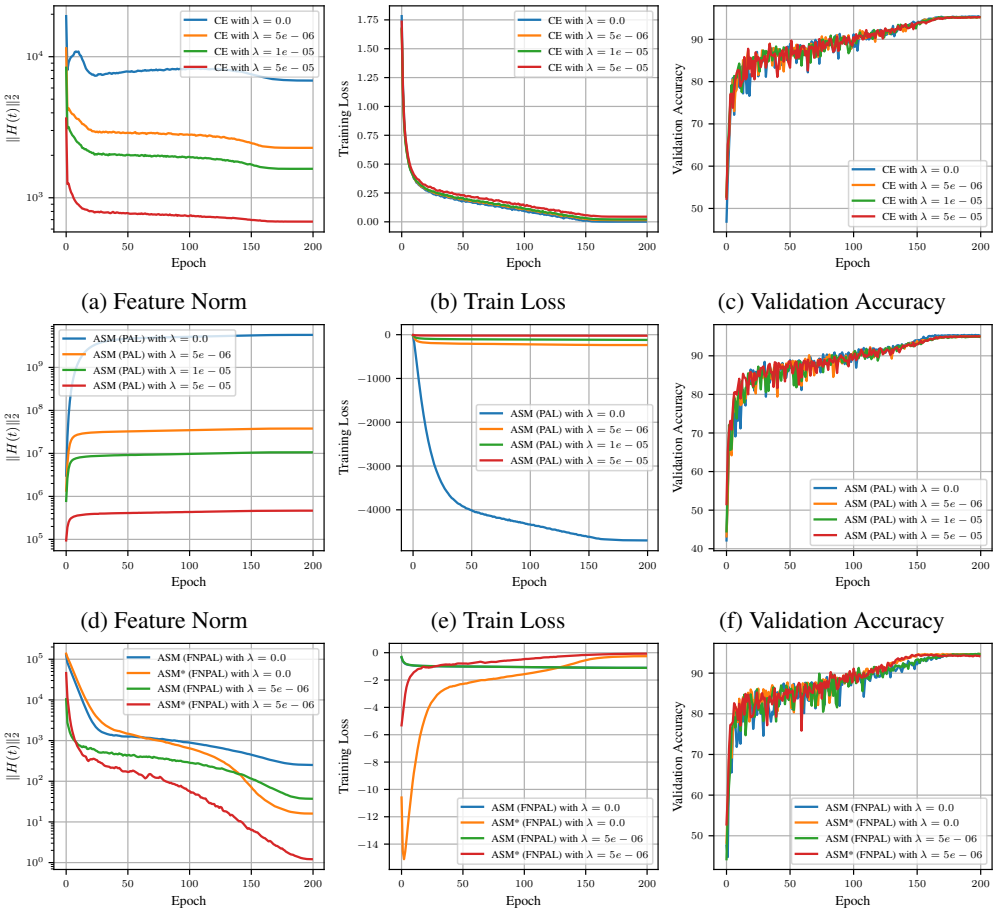
(m) CE ($\lambda = 0$) (n) CE ($\lambda = 5e-6$) (o) CE ($\lambda = 1e-5$) (p) CE ($\lambda = 5e-5$)

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.

C Other Potential Insights

C.1 A Good Initialization of Prototypes

As depicted in Sec. 3 and Appendix A.8, the dynamics under the ASM 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.*, $\widehat{\boldsymbol{w}}_i^\top \widehat{\boldsymbol{w}}_j = \frac{-1}{C-1}, \forall i \neq j$; (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.



(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, ASM (PAL) and ASM (FNPAL) under different weight decay coefficients.

C.2 Refined Decision-makings

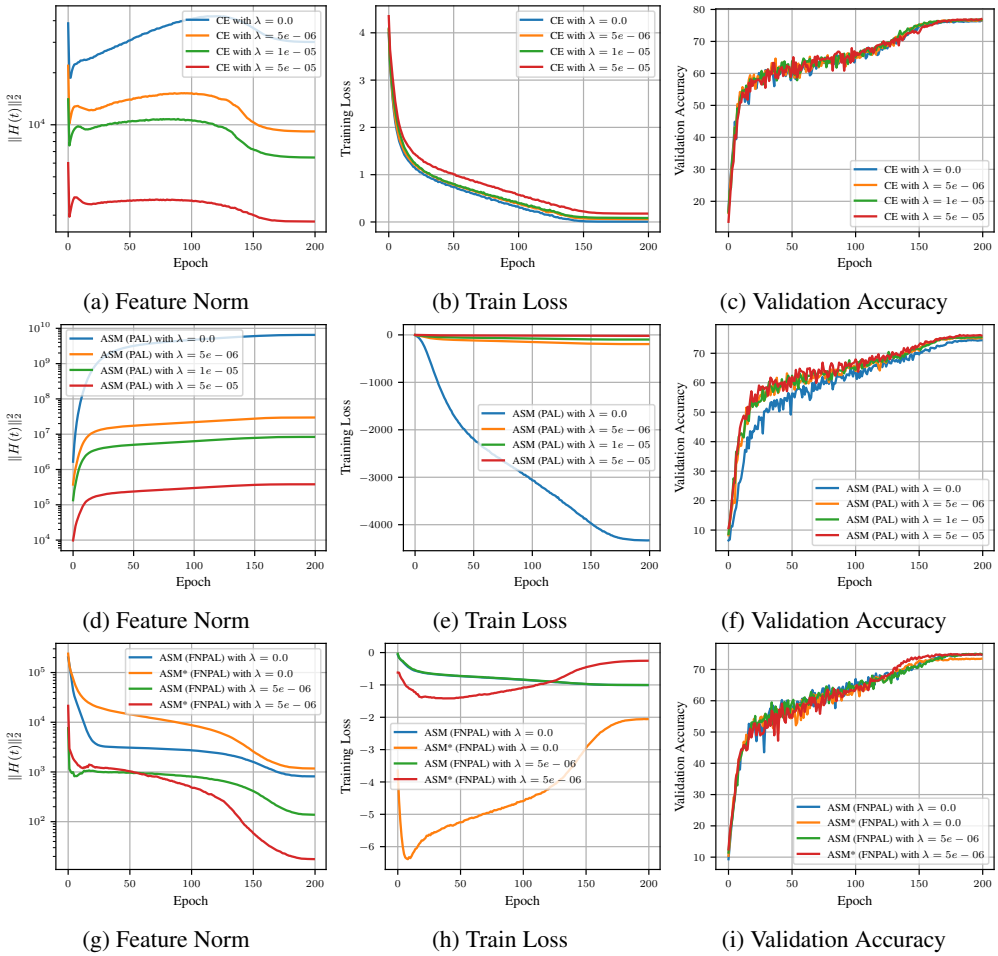
Recalling the rule—$\arg\max_{c'} \langle \boldsymbol{w}_{c'}, \boldsymbol{h} \rangle + b_{c'}$ that makes decision by selecting the class with the largest logit (where the inner product $\langle \boldsymbol{w}_{c'}, \boldsymbol{h} \rangle$ is dominant), which may not be good to directly use the learned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

C.3 Adjusted Sample Margin Loss

As aforementioned in Sec. 3.3 and the proof of Theorem 3.5, we will encounter zero gradients when the cosine similarity $\widehat{\boldsymbol{w}}_y^\top \widehat{\boldsymbol{h}}$ is -1 or 1, so we can adjust the loss to avoid the issue by to $\boldsymbol{w}_y$ and accelerate convergence:

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

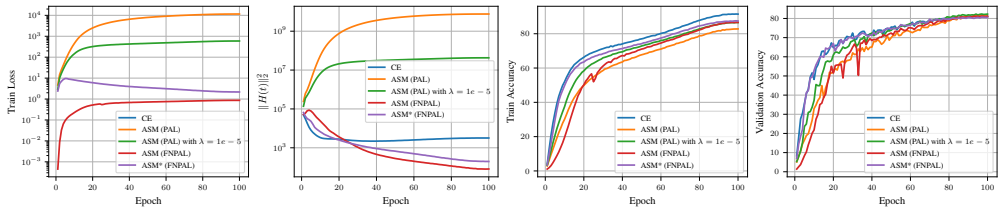


(a) Feature Norm (b) Train Loss (c) Validation Accuracy

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

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

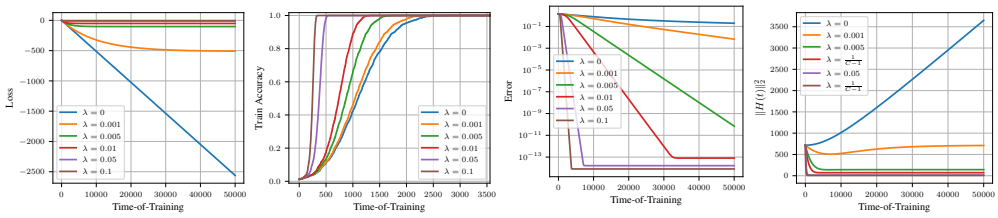
Figure 13: Behavior of visual classification on CIFAR-100 with CE, ASM (PAL) and ASM (FNPAL) under different weight decay coefficients.



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

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

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

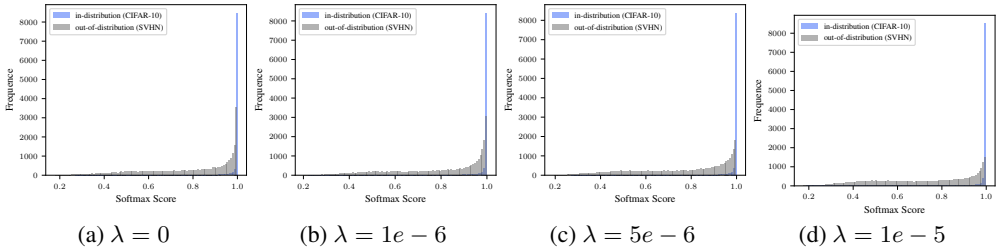


(a) Train Loss (b) Train Accuracy) (c) $\|\widehat{\boldsymbol{H}}(t) - \widehat{\boldsymbol{W}}\widehat{\boldsymbol{M}}\|_2$ (d) Feature Norm

Figure 15: Behavior of gradient descent iterates of the ASM (PAL) loss in Theorem 4.1 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 4.1, the error $\|\widehat{\boldsymbol{H}}(t) - \widehat{\boldsymbol{W}}\widehat{\boldsymbol{M}}\|_2$ decreases as an exponential rate $O(e^{-\lambda\eta t})$, and a larger λ can accelerate the convergence.

Table 4: 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.

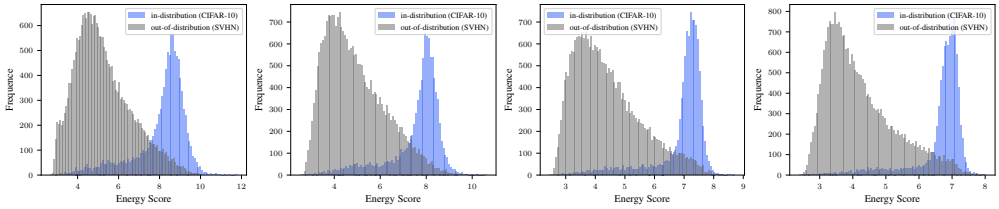
Dataset $\mathcal{D}_{\text{in}}^{\text{test}}$	λ	**FPR95** ↓	**AUROC** ↑	**AUPR** ↑
		Softmax-based / Energy-based / Feature Norm-based		
CIFAR-10	0	52.09 / 43.04 / 52.10	91.67 / 91.94 / 89.54	84.11 / 82.80 / 77.06
	1e-6	54.00 / 43.72 / **51.45**	91.44 / **92.12** / 89.08	82.31 / 81.77 / 74.16
	5e-6	**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**
		Softmax-based / Energy-based / Feature Norm-based		
CIFAR-100	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**



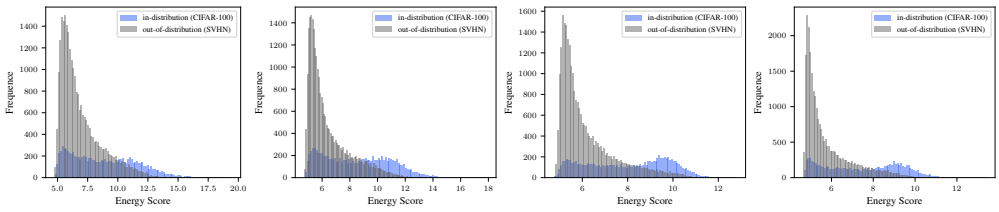
(a) $\lambda = 0$ (b) $\lambda = 1e-6$ (c) $\lambda = 5e-6$ (d) $\lambda = 1e-5$

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

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

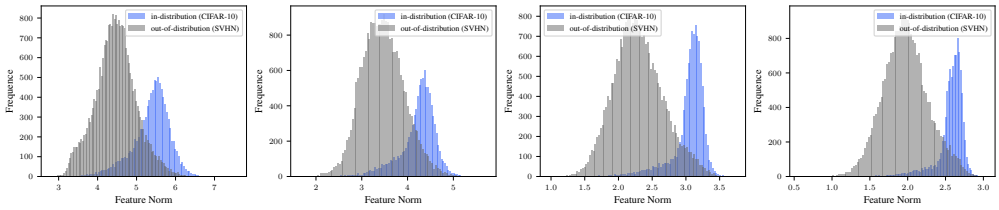


(a) $\lambda = 0$ (b) $\lambda = 1e-6$ (c) $\lambda = 5e-6$ (d) $\lambda = 1e-5$

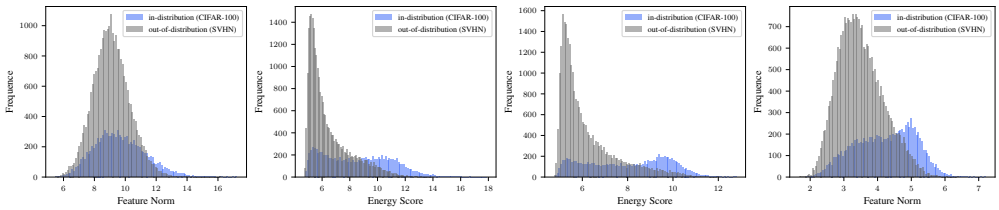


(e) $\lambda = 0$ (f) $\lambda = 1e-6$ (g) $\lambda = 5e-6$ (h) $\lambda = 1e-5$

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

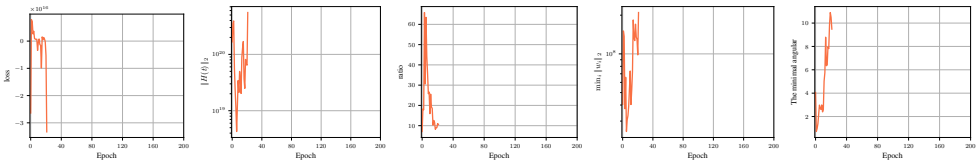


(a) $\lambda = 0$ (b) $\lambda = 1e-6$ (c) $\lambda = 5e-6$ (d) $\lambda = 1e-5$



(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 ASM loss 2.1 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: $\arccos \max_{i \neq j} \widehat{\boldsymbol{w}}_i^\top \widehat{\boldsymbol{w}}_j$. 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.