A Tiny Change, A Giant Leap: Long-Tailed Class-Incremental Learning via Geometric Prototype Alignment (Supplementary Materials)

Xinyi Lai[1] Luojun Lin[1*] Weijie Chen[2,3] Yuanlong Yu[1*]
[1]Fuzhou University, China [2]Zhejiang University, China [3]Hikvision Research Institute, China
laixinyi023@gmail.com, chenweijie@zju.edu.cn, {ljlin, yu.yuanlong}@fzu.edu.cn

A. Theoretical Proofs

A.1. Proof of Theorem 1: Accelerated Convergence

Proof. Consider the cross-entropy loss $\mathcal{L}_{\text{CE}}(W)$ with Softmax-activated logits. Under $\lambda_{\min}$-strong convexity near the optimal weights W^*, the gradient descent update satisfies:

$$\|W^{(k+1)} - W^*\|^2 \leq \left(1 - \frac{\eta\lambda_{\min}}{2}\right)\|W^{(k)} - W^*\|^2, \tag{1}$$

for learning rate $\eta \leq 1/L$, where L is the Lipschitz constant of $\nabla\mathcal{L}_{\text{CE}}$. The number of iterations needed to achieve $\|W^{(k)} - W^*\| \leq \epsilon$ is bounded by:

$$k \geq \frac{2}{\eta\lambda_{\min}}\log\frac{\|W^{(0)} - W^*\|}{\epsilon}. \tag{2}$$

Let $\theta_c = \arccos(\langle W_c^{(0)}, W_c^*\rangle)$. The initial alignment error is geometrically quantified by:

$$\|W_c^{(0)} - W_c^*\|^2 = 2(1 - \cos\theta_c), \quad \text{for unit vectors } W_c^{(0)}, W_c^*. \tag{3}$$

Under hyperspherical initialization (HSI) in GPA:

$$\cos\theta_c^{\text{GPA}} = \langle\mu_c/\|\mu_c\|, W_c^*\rangle \geq \cos\theta_c^{\text{rand}}, \tag{4}$$

where $\theta_c^{\text{rand}} \sim \mathcal{U}(0, \pi/2)$ for random initialization. Empirical measurements show $\mathbb{E}[\theta_c^{\text{GPA}}] \approx 18^\circ$ vs $\mathbb{E}[\theta_c^{\text{rand}}] \approx 45^\circ$, yielding a complexity reduction factor:

$$\frac{1 - \sin\theta_{\text{rand}}}{1 - \sin\theta_{\text{GPA}}} \approx \frac{1 - \sin 45^\circ}{1 - \sin 18^\circ} \approx 2.7. \tag{5}$$

□

A.2. Proof of Theorem 2: Fisher-Optimal Direction

Proof. For Gaussian class-conditionals $\phi(x)|y = c \sim \mathcal{N}(\mu_c, \Sigma)$, the Fisher-optimal classifier between class c and background 0 maximizes:

$$J(W_c) = \frac{(W_c^T(\mu_c - \mu_0))^2}{W_c^T\Sigma W_c}. \tag{6}$$

The optimal solution is $W_c^{\text{Fisher}} \propto \Sigma^{-1}(\mu_c - \mu_0)$.

Under GPA initialization $W_c^{(0)} = \mu_c/\|\mu_c\|$. When $\Sigma = \sigma^2 I + E$ with $\|E\|_2 \leq \mathcal{O}(1/\sqrt{N_c})$:

$$W_c^{(0)} \propto \mu_c \approx \frac{\sigma^{-2}(\mu_c - \mu_0)}{1 + \mathcal{O}(1/\sqrt{N_c})}, \tag{7}$$

where we use $\mu_0 = \mathbb{E}[\phi(x)] \approx \frac{1}{|\mathcal{C}|}\sum_c \mu_c$. Hence:

$$\langle W_c^{(0)}, W_c^{\text{Fisher}}\rangle \geq 1 - \mathcal{O}\left(\frac{\sqrt{d}}{\sqrt{N_c}}\right), \tag{8}$$

proving approximate alignment when $N_c \gg d$. □

A.3. Proof of Proposition 1: Generalization Bound

Proof. Using the margin-based Rademacher complexity framework [1], for classifier $h(x) = \arg\max_c[W_c^T\phi(x) + b_c]$, the generalization error satisfies:

$$\mathcal{E} \leq \underbrace{\frac{C_1}{\sqrt{N}}}_{\text{Sample Complexity}} + \underbrace{\mathbb{E}\left[\max_{x,y}\min_{c\neq y}(W_y - W_c)^T\phi(x)\right]}_{-\text{Margin Term}} + \mathcal{O}\left(\frac{d^{3/2}}{\lambda_{\min}N}\right). \tag{9}$$

The margin term is governed by the minimal prototype distance $\delta_{\min}$:

$$\mathbb{E}\left[\min_{c\neq y}(W_y - W_c)^T\phi(x)\right] \geq \frac{\delta_{\min}}{2} - \mathcal{O}(\sqrt{\log|\mathcal{C}|/N}), \tag{10}$$

while the class imbalance ratio ρ amplifies the error through biased gradients. Hence:

$$\mathcal{E} \leq \frac{C_1}{\sqrt{N}} + C_2\rho\left(\frac{1}{\delta_{\min}} + \mathcal{O}(\sqrt{\log|\mathcal{C}|/N})\right). \tag{11}$$

GPA maximizes $\delta_{\min}$ via geometric alignment, directly reducing the dominant $C_2\rho\delta_{\min}^{-1}$ term. □

B. Overview of Integrated LT-CIL Methods

We briefly introduce the six methods integrated into our framework, highlighting their core strategies in addressing long-tailed class-incremental learning.

LUCIR [5] LUCIR (Learning a Unified Classifier Incrementally with Rebalancing) tackles catastrophic forgetting by combining knowledge distillation with inter-class separation constraints. It jointly optimizes for retaining consistent features for old classes while ensuring that new classes are well separated in the feature space.

PODNET [2] PODNET (Pooled Output Distillation Network) leverages prototype distillation by extracting and storing class prototypes from previous tasks. By aligning current task features with these stored prototypes, PODNET effectively reduces interference from new tasks and balances the learning of both old and new classes.

Finetune The Finetune strategy involves pre-training the model on an initial dataset and subsequently fine-tuning it on incremental data. Although straightforward, this approach is prone to severe forgetting of previously learned classes, especially under long-tailed distributions, and is typically used as a baseline for comparison.

L2P [9] L2P (Learning to Prompt) employs prompt tuning by incorporating learnable prompts into the visual model. These prompts serve as auxiliary inputs that enable the model to adapt to new tasks with minimal modification of the pre-trained parameters, thus mitigating the forgetting effect.

DualPrompt [8] DualPrompt enhances prompt tuning by introducing dual prompts at both the input level and within internal network layers. This dual-prompt mechanism captures task-specific information more granularly, ensuring better knowledge retention across incremental stages and balancing performance between old and new tasks.

CODA-Prompt [7] CODA-Prompt combines contextual information with prompt tuning to build context-aware prompt modules. By exploiting inter-task contextual relationships, it facilitates effective knowledge transfer and generalization, thereby reducing catastrophic forgetting in incremental learning scenarios.

GradRew [3] GradRew employs a two-stage gradient reweighting strategy for long-tailed class-incremental learning: it first boosts underrepresented classes by scaling their cross-entropy gradients according to historical magnitudes, then decouples plasticity and stability with separate reweighting of cross-entropy and distillation losses, using a distribution-aware adjustment to further emphasize rare classes.

DynaPrompt [4] Dynamically Anchored Prompting (DAP) keeps a single "general" prompt and, for each task, learns a task-specific "boosting" anchor and updates the general prompt by aligning it to both this anchor and a "stabilizing" anchor summarizing past tasks, with a task-size–based coefficient that flexibly trades off plasticity and stability without any prompt pool or rehearsal.

EASE [10] Expandable Subspace Ensemble (EASE) freezes the backbone and adds a lightweight adapter per task to create task-specific subspaces, extracts prototypes in each subspace, synthesizes missing old-class prototypes via semantic similarity in the co-occurrence space, and ensembles across all subspaces—reweighting logits to emphasize the current task—without exemplars and within a fixed parameter budget.

RanPAC [6] RanPAC projects pre-trained features into a higher-dimensional random (optionally nonlinear) basis to approximate an isotropic Gaussian, then updates class prototypes via a closed-form, ridge-regularized solution using additive Gram and prototype matrix updates, yielding a rehearsal-free and parameter-fixed continual learner.

Each of these methods addresses the challenges of long-tailed class-incremental learning from different perspectives. In our work, we integrate the Geometric Prototype Alignment (GPA) module into these frameworks to enhance weight initialization and gradient flow, leading to significant improvements in overall performance.

References

[1] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *JMLR*, 3(Nov):463–482, 2002.

[2] Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV*, pages 86–102, 2020.

[3] Jiangpeng He. Gradient reweighting: Towards imbalanced class-incremental learning. In *CVPR*, pages 16668–16677, 2024.

[4] Chenxing Hong, Yan Jin, Zhiqi Kang, Yizhou Chen, Mengke Li, Yang Lu, and Hanzi Wang. Dynamically anchored prompting for task-imbalanced continual learning. *IJCAI*, 2024.

[5] Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *CVPR*, pages 831–839, 2019.

[6] Mark D McDonnell, Dong Gong, Amin Parvaneh, Ehsan Abbasnejad, and Anton Van den Hengel. Ranpac: Random projections and pre-trained models for continual learning. *NeurIPS*, 36:12022–12053, 2023.

[7] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *CVPR*, pages 11909–11919, 2023.

[8] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *ECCV*, pages 631–648, 2022.

[9] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *CVPR*, pages 139–149, 2022.

[10] Da-Wei Zhou, Hai-Long Sun, Han-Jia Ye, and De-Chuan Zhan. Expandable subspace ensemble for pre-trained model-based class-incremental learning. In *CVPR*, pages 23554–23564, 2024.