PAI: PLASTICITY-ANCHORED INITIALIZATION FOR IMPROVED STABILITY-PLASTICITY TRADE-OFFS IN CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

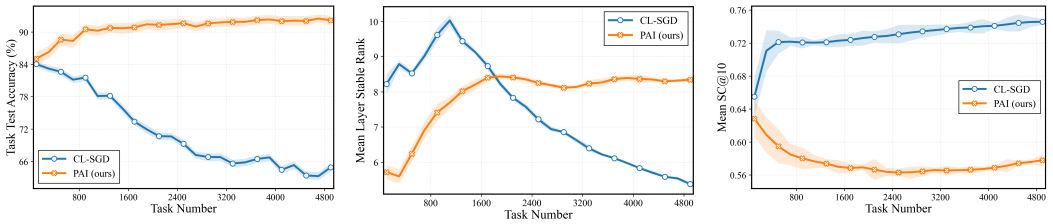
Continual learning (CL) requires models to remain adaptable to nonstationary task streams while preserving previously acquired knowledge. Yet the standard practice of warm-starting each task from the preceding task solution can progressively impair plasticity: over long task sequences, we observe increasing spectral concentration and decreasing stable rank, accompanied by deteriorating adaptation to new tasks. Existing approaches often restore plasticity by resetting or modifying parameters, but such interventions risk disrupting learned representations and exacerbating forgetting. To address this challenge, we introduce **P**lasticity-**A**nchored **I**nitialization (`PAI`), a simple task-boundary initialization strategy that exploits the learning trajectory itself to balance stability and plasticity. `PAI` maintains two complementary parameter states: a *fast-memory anchor*, given by the latest task solution, that preserves recent adaptation, and a *slow-memory anchor*, formed by averaging historical task endpoints, that aggregates information across the learning trajectory. Before each new task, `PAI` interpolates between these anchors, moderating task-specific drift while retaining learned structure and steering optimization toward a more plastic parameter region. Theoretically, we analyze how single-sample perturbations propagate across task boundaries and show that, under matched assumptions and learning rates, `PAI` admits a task-wise uniform-stability upper bound no larger than that of standard sequential SGD. Empirically, across class-incremental learning, all-data replay, and 5,000-task ImageNet sequences, `PAI` consistently improves both knowledge retention and late-stage adaptation across CNN and Transformer architectures. These results establish task-boundary initialization as a simple yet effective mechanism for sustaining plasticity in long-horizon continual learning.

1 INTRODUCTION

Continual learning (CL) requires a model to preserve acquired knowledge while remaining capable of learning from an evolving task stream (Kirkpatrick et al., 2017; Parisi et al., 2019). It must contend with both *catastrophic forgetting* of previous tasks and *plasticity loss* that impairs subsequent learning (Dohare et al., 2024). Standard sequential training initializes each task from the preceding task's endpoint. This warm start transfers learned information, but can also carry forward a parameter configuration that has become progressively less adaptable (Ash & Adams, 2020; Berariu et al., 2021; Lyle et al., 2023). The stability–plasticity trade-off therefore depends on the initialization inherited at task boundaries as well as the updates performed within each task.

Existing approaches use regularization to sustain trainability (Kumar et al., 2025; Lyle et al., 2022), or reset and modify selected parameters to restore adaptability (Nikishin et al., 2022; Shin et al., 2024; Han et al., 2026). Strong constraints can restrict adaptation, while parameter modification can disturb useful learned structure, particularly when limited replay makes historical knowledge difficult to recover. Methods combining fast and slow learning components further demonstrate the value of retaining information over different timescales (Lee et al., 2024). Building on these insights, we study how the latest task solution and accumulated task history should jointly determine the starting point for future learning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Plasticity. (b) Stable Rank. (c) Spectral Concentration @10.

Figure 1: Comparison between CL-SGD and PAI in terms of (a) task test accuracy, (b) stable rank, and (c) spectral concentration at the top 10 singular values (SC@10). As learning progresses, CL-SGD exhibits increasing spectral concentration and declining stable rank in later tasks, consistent with a few singular directions increasingly dominating the representations and reducing their effective dimensionality. This pattern accompanies deteriorating task accuracy, indicating plasticity loss. PAI maintains lower spectral concentration and higher stable rank in later tasks, alongside sustained task accuracy, consistent with better preservation of representational diversity and plasticity.

We examine this question through the spectral evolution of a CNN4 model sequentially trained on 5,000 binary classification tasks constructed from ImageNet. In later tasks, CL-SGD exhibits increasing spectral concentration in the top ten singular values and decreasing stable rank, alongside declining current-task test accuracy (Figure 1). These diagnostics indicate that a few singular directions account for a growing share of spectral energy. The joint trend motivates a mechanistic hypothesis: *repeated task specialization concentrates the spectrum in a few dominant directions, potentially narrowing the effective representation space and limiting subsequent adaptation*. Although these observations do not establish causality, they motivate preserving spectral balance when constructing the next-task initialization. Our central insight is to use the learning trajectory itself as a source of plasticity. Historical endpoints provide a learned reference that aggregates earlier adaptation. Combining this reference with the latest endpoint offers a way to moderate task-specific drift while retaining recent knowledge. We therefore ask:

Can we use learned history to recover plasticity while retaining prior knowledge?

We propose **P**lasticity-**A**nchored **I**nitialization (`PAI`), a task-boundary initialization strategy that combines two parameter states: the latest task endpoint z^t as a *fast-memory anchor*, and the arithmetic average of previous task endpoints x^t as a *slow-memory anchor*. Task t starts from

$$y_0^t = (1-\beta)z^t + \beta x^t, \qquad \beta \in [0,1].$$

The slow anchor aggregates information across tasks, while the fast anchor gives recent adaptation an explicit contribution to the initialization. The coefficient β controls their relative influence, with $\beta = 0$ recovering standard sequential warm-starting. PAI thus uses historical averaging to shape the starting point of future optimization. It requires one additional parameter state and leaves within-task training unchanged; its initialization is computed by parameter averaging without spectral decomposition.

We analyze how single-sample perturbations propagate across task transitions. Under the stated assumptions for sequential SGD, with matched learning rates and assumption constants, PAI admits task-wise algorithmic-stability bounds and an expected generalization-gap bound no larger than those derived for CL-SGD. We evaluate PAI through class-incremental learning with limited replay, expanding-data training with all-data replay, and sequences of 5,000 tasks. Across these settings, PAI improves final average accuracy, complements multiple rehearsal-based methods, and sustains current-task performance over long task streams.

We summarize our **main contributions** as follows:

- We identify increasing spectral concentration and declining stable rank as empirical signatures accompanying reduced adaptability, motivating a connection between dominant singular directions, reduced effective representational dimensionality, and plasticity loss.
- We introduce PAI, which constructs the next-task initialization by combining a fast memory, represented by the latest task endpoint, with a slow memory, maintained as an average

of historical endpoints. This task-boundary intervention integrates recent adaptation with accumulated task history to balance knowledge retention and future learning.

- We establish a comparison of perturbation propagation and the resulting stability and generalization bounds with CL-SGD under shared assumptions. Experiments across architectures and continual learning settings demonstrate improved final average accuracy and plasticity.

2 Related Work

Stability–Plasticity Tradeoff in Continual Learning. Continual learning (CL) seeks to balance adaptation to new tasks with retention of prior knowledge Wang et al. (2024). Existing approaches broadly fall into three categories: regularization-based methods constrain parameter updates Farajtabar et al. (2020); Kirkpatrick et al. (2017), rehearsal-based methods replay stored examples Chaudhry et al. (2019), and architecture-based methods allocate additional capacity to new tasks Rusu et al. (2016); Wang et al. (2022). This tradeoff also arises in continual post-training of multimodal LLMs Wei et al.; Xu et al. (2025); Yan et al. (2026), continual offline reinforcement learning Hu et al. (2025a;b), and federated continual learning Yao et al. (2024). Within parameter-based approaches, EWC Kirkpatrick et al. (2017) and SI Zenke et al. (2017) penalize changes to parameters important for previous tasks. Hare & Tortoise Lee et al. (2024) combines slow and fast learning components to balance retention and adaptation, while WSC Cho et al. (2026) couples importance-based parameter resetting with weight averaging to restore plasticity and consolidate knowledge under abundant memory. Rehearsal provides a flexible framework for combining experience replay with knowledge distillation. Experience Replay (ER) Riemer et al. (2019) jointly trains on current and buffered examples. DER Buzzega et al. (2020) additionally stores and matches historical logits, while XDER Boschini et al. (2022) extends this strategy with additional mechanisms. CLS-ER Arani et al. (2022) uses complementary fast and slow components for plasticity and stability, and BFP Gu et al. (2023) aligns feature representations through a learnable linear projection. SCoMMER Sarfraz et al. (2023) and SARL Sarfraz et al. (2025) exploit sparse representations, whereas IDER Liu et al. (2026) explicitly addresses plasticity degradation across sequential tasks.

Loss of Plasticity. Prior methods mitigate plasticity loss through weight reinitialization, such as S&P (Ash & Adams, 2020) and DASH (Shin et al., 2024); parameter regularization (Kumar et al., 2025; Elsayed et al., 2024; Lewandowski et al., 2024a); spectral or rank constraints (Kumar et al., 2021a;b; He et al., 2024); and utility-based neuron resets (Sokar et al., 2023; Dohare et al., 2024; Elsayed & Mahmood, 2024). Other studies exploit the greater plasticity of linear networks (Dohare et al., 2024; Lewandowski et al., 2024b; Park et al., 2025). While restoring trainability does not necessarily preserve prior knowledge, this tension is particularly important with limited replay. We therefore study CL with a small replay buffer, where `PAI` aims to balance new-task adaptation and knowledge retention.

3 Methodology

Problem Setup. In this work, we consider a continual learning setting with a sequence of T tasks $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T\}$. Each task t is associated with a data distribution $\mathcal{D}_t$ over the example space $\mathcal{Z}$, and the model is trained on these tasks sequentially. For each task t, the population risk is defined as:

$$F_t(w) = \mathbb{E}_{\xi \sim \mathcal{D}_t} \left[\ell(w; \xi) \right], \tag{1}$$

where $w \in \mathcal{W}$ denotes the model parameter and $\ell(w; \xi)$ is the loss evaluated on example ξ. In practice, for each task t, we only observe a finite training dataset $S_t = \{\xi_{t,j}\}_{j=1}^n, \xi_{t,j} \sim \mathcal{D}_t$, and minimize the empirical risk $\widehat{F}_t(w) = \frac{1}{n} \sum_{j=1}^n \ell(w; \xi_{t,j})$.

3.1 Plasticity in Continual Learning from a Spectral Perspective

From a spectral perspective, plasticity loss may be associated with the progressive dominance of a few singular directions. For a weight matrix $W \in \mathbb{R}^{m \times n}$ with singular values $\sigma_1(W) \geq \cdots \geq \sigma_r(W) > 0$, we quantify its spectral structure by $\mathrm{SC@}k(W) = \frac{\sum_{i=1}^k \sigma_i^2(W)}{\sum_{i=1}^r \sigma_i^2(W)}, \mathrm{StableRank}(W) = \frac{\|W\|_F^2}{\|W\|_2^2} = \frac{\sum_{i=1}^r \sigma_i^2(W)}{\sigma_1^2(W)}$, where $r = \mathrm{rank}(W)$. Higher $\mathrm{SC@}k$ indicates stronger concentration in the

Algorithm 1 Plasticity-Anchored Initialization

Require: Task sequence $\{S_1, S_2, \ldots, S_T\}$, initial model w_0, local update steps K, learning rate η, anchoring weight $\beta \in [0, 1]$.

Initialize $z^1 \leftarrow w_0$ and $x^1 \leftarrow w_0$.
for $t = 1, 2, \ldots, T$ **do**
  Receive current task dataset S_t.
  Set the task initialization $y_0^t \leftarrow (1-\beta)z^t + \beta x^t$.
  for $k = 0, 1, \ldots, K-1$ **do**
    Sample a mini-batch $B_k^t \subset S_t$.
    Compute the stochastic gradient $g_k^t = \frac{1}{|B_k^t|} \sum_{\xi \in B_k^t} \nabla \ell(y_k^t; \xi)$.
    Update the model $y_{k+1}^t \leftarrow y_k^t - \eta g_k^t$.
  end for
  Store the task endpoint $z^{t+1} \leftarrow y_K^t$.
  Update the historical anchor $x^{t+1} \leftarrow (1 - c_{t+1})x^t + c_{t+1}z^{t+1}, c_{t+1} = \frac{1}{t}$.
end for
return y_K^T.

leading singular directions, whereas higher stable rank reflects a more distributed spectrum and greater effective dimensionality. As shown in Figure 1, sequential training exhibits increasing spectral concentration and decreasing stable rank alongside deteriorating current-task performance. This joint trend motivates our central hypothesis: *repeated task specialization progressively concentrates spectral energy into a few dominant directions, narrowing the effective representation space and limiting subsequent adaptation*. While this evidence is correlational rather than causal, it motivates preserving spectral balance when constructing the initialization for future tasks.

3.2 Our Algorithm: Plasticity-Anchored Initialization

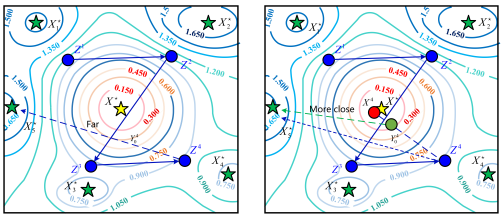
Slow-memory Anchor. `PAI` maintains two parameter states. The first one is the task endpoint z^t, which represents the model obtained after adapting to the previous task. This state contains the most recent task-specific knowledge and therefore provides stability for subsequent learning. The second one is the plasticity anchor x^t, which is average of historical task endpoints:

$$x^{t+1} = (1 - c_{t+1})x^t + c_{t+1}z^{t+1}, \tag{2}$$

where c_{t+1} is the anchor update coefficient. In our experiments, we set $c_{t+1} = \frac{1}{t}$, which makes the anchor an arithmetic average of historical endpoints: $x^{t+1} = \frac{1}{t}\sum_{s=2}^{t+1} z^s$. This averaging operation smooths out task-specific drift and produces a more stable reference point along the learning trajectory. Intuitively, x^t preserves information that is consistently useful across tasks, while reducing the influence of directions that are overly specialized to the most recent task.

Fast-memory Anchor. We retain the latest task endpoint as a fast-memory anchor, $z^{t+1} = y_K^t$, which captures recent task-specific adaptation without temporal averaging. Together with the historical anchor x^{t+1}, it determines the next-task initialization: $y_0^{t+1} = (1-\beta)z^{t+1} + \beta x^{t+1}$. This interpolation preserves a direct contribution from the most recently learned parameters while incorporating accumulated historical information.

PAI. At the beginning of the task t, `PAI` constructs the initial state as a interpolation of the fast-memory anchor z^t and the slow-memory anchor x^t: $y_0^t = (1-\beta)z^t + \beta x^t$, where $\beta \in [0, 1]$. When $\beta = 0$, it reduces to standard CL, since the next task starts directly from the previous endpoint. When $\beta = 1$, it is fully initialized from the historical anchor, which provides stronger plasticity but may weaken training efficiency. Intermediate β can effectively balance these two effects.



(a) CL-SGD (b) PAI

Figure 2: Schematic illustration of `CL-SGD` and `PAI`.

Task-wise training process. After computing the anchored initialization y_0^t, `PAI` proceeds with the standard optimization procedure on the current task S_t. Specifically, at each local step k, we

sample a mini-batch $B_k^t \subset S_t$ and compute the stochastic gradient $g_k^t = \frac{1}{|B_k^t|} \sum_{\xi \in B_k^t} \nabla \ell(y_k^t; \xi)$. The task-specific model is updated by $y_{k+1}^t = y_k^t - \eta g_k^t$, where η denotes the learning rate. This step is identical to the usual stochastic optimization update, except that the starting point y_0^t is constructed from the proposed anchoring mechanism rather than directly inherited from the previous task endpoint. After the entire task sequence has been processed, PAI returns the final anchor y_K^T as the output model. In this way, PAI preserves the standard within-task training procedure while modifying only the task-boundary initialization, the detailed pipeline is stated in Algorithm 1.

3.3 Discussion

PAI differs from regularization-based methods that sustain plasticity by continuously constraining optimization, and from reset-based methods that restore plasticity by modifying or discarding learned parameters. Instead, PAI intervenes only at task boundaries, interpolating the latest endpoint z^t with the historical anchor x^t. As illustrated in Figure 2, the fast anchor preserves recent task knowledge, while the slow anchor moderates task-specific drift and guides optimization toward a more stable and plastic region. This simple mechanism improves the stability–plasticity trade-off while leaving within-task training unchanged and introducing negligible overhead.

4 Theoretical Analysis

In this section, we present an algorithmic stability-based analysis of PAI in continual learning, demonstrating its theoretical advantages over the conventional sequential training paradigm.

Setup. Let $\mathcal{A}$ be a randomized continual learning algorithm. After sequentially training on $\mathcal{S}^{(T)}$, the final output model is $w_{T,K} := \mathcal{A}(\mathcal{S}^{(T)})$, where K denotes the number of updates performed on each task. To analyze stability, we consider a neighboring sequence $\widetilde{\mathcal{S}}^{(T)}$ that differs from $\mathcal{S}^{(T)}$ in exactly one example. Specifically, for some task $t^\star \in [T]$ and sample index $j^\star \in [n]$, the example $z_{t^\star,j^\star}$ is replaced by an independent copy $\widetilde{z}_{t^\star,j^\star} \sim \mathcal{D}_{t^\star}$, while all other examples remain unchanged:

$$\widetilde{S}_{t^\star} = \{z_{t^\star,1}, \ldots, \widetilde{z}_{t^\star,j^\star}, \ldots, z_{t^\star,n}\}, \qquad \widetilde{S}_t = S_t, \quad t \neq t^\star. \tag{3}$$

The corresponding output model is $\widetilde{w}_{T,K} := \mathcal{A}(\widetilde{\mathcal{S}}^{(T)})$.

Given a loss function $\ell(w; z)$, the population loss over the task sequence is defined as $\mathcal{L}_{\mathcal{D}^{(T)}}(w) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{z \sim \mathcal{D}_t} [\ell(w; z)]$, and the empirical loss over the observed continual learning sequence is $\widehat{\mathcal{L}}_{\mathcal{S}^{(T)}}(w) := \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \sum_{j=1}^{n} \ell(w; z_{t,j})$. Our goal is to bound the expected generalization gap

$$\mathcal{E}_{\text{gen}} := \mathbb{E}\left[\mathcal{L}_{\mathcal{D}^{(T)}}(w_{T,K}) - \widehat{\mathcal{L}}_{\mathcal{S}^{(T)}}(w_{T,K})\right].$$

Definition 4.1 (**Uniform Stability for Continual Learning**). *Let $\mathcal{A}$ be a randomized continual learning algorithm that maps a task sequence $\mathcal{S}^{(T)} = \{S_1, \ldots, S_T\}$ to the final model $w_T = \mathcal{A}(\mathcal{S}^{(T)})$. For any neighboring sequence $\widetilde{\mathcal{S}}^{(T)}$ differing from $\mathcal{S}^{(T)}$ in exactly one example, let $\widetilde{w}_T = \mathcal{A}(\widetilde{\mathcal{S}}^{(T)})$. We say that $\mathcal{A}$ is ϵ-uniformly stable if* $\sup_{z \in \mathcal{Z}} \mathbb{E}_{\mathcal{A}} [|\ell(w_T; z) - \ell(\widetilde{w}_T; z)|] \leq \epsilon$.

We analyze the final model's expected generalization gap via task-averaged stability bounds and show that PAI admits a bound no larger than that of CL-SGD under shared assumptions. Following prior stability analyses (Hardt et al., 2016; Lei & Ying, 2020), we adopt the following assumptions:

Assumption 4.2. *For every task $t \in [T]$ and every sample $z \in \mathcal{Z}$, the loss $\ell(\cdot; z)$ is L-smooth, i.e., $\|\nabla \ell(w; z) - \nabla \ell(w'; z)\| \leq L\|w - w'\|, \forall\, w, w' \in \mathbb{R}^d$.*

Assumption 4.3 (Uniform Terminal-Region Control)**.** *For a fixed horizon and deterministic learning-rate schedule, there exist a deterministic convex set $\mathcal{C}_T \subseteq \mathbb{R}^d$ and constants $L_G \geq 0$, $U > 0$, shared by CL-SGD and PAI for all $\beta \in [0, 1]$, such that every admissible training sequence yields final outputs in $\mathcal{C}_T$ almost surely, and $\|\nabla \ell(v; z)\| \leq L_G, 0 \leq \ell(v; z) \leq U, \forall v \in \mathcal{C}_T,\ z \in \mathcal{Z}$.*

Assumption 4.4 (Uniform Within-Task Gradient Discrepancy)**.** *For each $t \in [T]$, there exists a deterministic $\sigma_t \geq 0$ such that $\|\nabla \ell(w; z) - \nabla \ell(w; z')\| \leq 2\sigma_t, \quad \forall w \in \mathbb{R}^d,\ z, z' \in \text{supp}(\mathcal{D}_t)$.*

Before presenting the main theorem, we first introduce a key lemma. We express the two model-averaging constructions in PAI as recursive sequences along the task index t, and prove a dominance relation between their magnitudes, which serves as a crucial ingredient in our subsequent analysis.

Lemma 4.5 (**Attenuated Amplification Factor**). *Fix a task index $t \in [T]$. Suppose $\beta \in [0,1]$ and $A_u \geq 1$ for all $u = t{+}1, \ldots, T$. Define $Q_{t+1}^{(t)} := 1, Q_{u+1}^{(t)} := A_u Q_u^{(t)}, u = t{+}1, \ldots, T$. Equivalently, $Q_u^{(t)} = \prod_{r=t+1}^{u-1} A_r, u = t+1, \ldots, T+1$, where an empty product equals 1. Define the sequences $\rho_u^{(\beta;t)}$ and $\nu_u^{(\beta;t)}$ by $\rho_{t+1}^{(\beta;t)} = \frac{1}{t}, \nu_{t+1}^{(\beta;t)} = 1$, and, for $u = t+1, \ldots, T$, by $\rho_{u+1}^{(\beta;t)} = \frac{u-1+\beta A_u}{u}\rho_u^{(\beta;t)} + \frac{(1-\beta)A_u}{u}\nu_u^{(\beta;t)}, \nu_{u+1}^{(\beta;t)} = A_u\left(\beta\rho_u^{(\beta;t)} + (1-\beta)\nu_u^{(\beta;t)}\right)$. Then, for every $u = t+1, \ldots, T+1$, $0 \leq \rho_u^{(\beta;t)} \leq Q_u^{(t)}, 0 \leq \nu_u^{(\beta;t)} \leq Q_u^{(t)}$. In particular, the endpoint propagation coefficient satisfies*

$$\nu_{T+1}^{(\beta;t)} \leq Q_{T+1}^{(t)} = \prod_{u=t+1}^{T} A_u. \tag{4}$$

In fact, the LHS of Eq. (4) characterizes the cumulative scalar factor induced by PAI in the stability bound, whereas the RHS corresponds to the amplification factor accumulated under conventional CL iterations. Therefore, Lemma 4.5 provides a key comparison principle: the perturbation propagation under PAI performs a reduced propagation factor in the stability upper bound. Building on this lemma, the following theorems present the explicit forms of the stability bounds for both procedures.

Setup. Assume Assumptions 4.2–4.4. Training samples are mutually independent, initialization is data-independent, and each update independently samples one current-task example uniformly with replacement. Both algorithms use the same deterministic learning-rate schedule $\{\eta_m\}_{m=1}^{TK}$ and shared assumption constants. Define

$$A_u := \prod_{m=(u-1)K+1}^{uK} (1 + L\eta_m), \qquad d_\tau(k) := \frac{2\sigma_\tau}{nL}\left[\prod_{m=(\tau-1)K+k+1}^{\tau K} (1 + L\eta_m) - 1\right], \tag{5}$$

where $u, \tau \in [T]$, $k \in \{0, \ldots, K\}$, and empty products equal 1.

Theorem 4.6 (Generalization bound of CL-SGD). *Under the above setup, the final endpoint $w_{T,K}$ of sequential SGD satisfies*

$$\left|\mathcal{E}_{\text{gen}}^{\text{CL-SGD}}\right| \leq \frac{1}{T}\sum_{\tau=1}^{T} \min_{0\leq k\leq K}\left\{\frac{Uk}{n} + L_G d_\tau(k)\prod_{u=\tau+1}^{T} A_u\right\}, \tag{6}$$

where the minimum is over integer k.

Theorem 4.7 (Generalization bound of PAI). *Under the same setup, consider PAI with a fixed, data-independent $\beta \in [0,1]$. Its final endpoint y_K^T satisfies*

$$\left|\mathcal{E}_{\text{gen}}^{\text{PAI}}\right| \leq \frac{1}{T}\sum_{\tau=1}^{T} \min_{0\leq k\leq K}\left\{\frac{Uk}{n} + L_G d_\tau(k)\nu_{T+1}^{(\beta;\tau)}\right\}, \tag{7}$$

where the minimum is over integer k and $\nu_{T+1}^{(\beta;\tau)}$ is defined in Lemma 4.5. The right-hand side is no larger than that of equation 6, with equality when $\beta = 0$.

Why historical anchoring helps. Our analysis provides a theoretical justification for using historical information in task initialization. In standard CL-SGD, a sample perturbation introduced at task τ is passed through successive task endpoints, yielding the propagation factor $\prod_{u=\tau+1}^{T} A_u$. By mixing the latest endpoint with a running historical anchor, PAI modifies this propagation through the initialization itself. Lemma 4.5 shows that the resulting endpoint coefficient satisfies $\nu_{T+1}^{(\beta;\tau)} \leq \prod_{u=\tau+1}^{T} A_u$. Under shared assumption constants and the same learning-rate schedule, this yields task-wise stability upper bounds, and consequently an upper bound on the absolute expected generalization gap, no larger than those of CL-SGD, recovering the CL-SGD bounds when $\beta = 0$. These results support historical anchoring as a mechanism for controlling sensitivity to individual training samples across tasks. They complement our empirical evaluation of retention and adaptation, which assesses the stability–plasticity trade-off beyond the guarantees established here.

5 Experiments

Datasets and Models. We evaluate PAI on CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), Tiny-ImageNet, and ImageNet (Deng et al., 2009), across CNN-4, VGG-16 (Simonyan & Zisserman, 2014), ResNet-18 (He et al., 2016), and ViT-Tiny (Dosovitskiy et al., 2020).

Table 1: Comparison of Final Average Accuracy (FAA) across different CL methods. All experiments are repeated 5 times with different seeds.

Method	CIFAR-10		CIFAR-100		Tiny-ImageNet	
	Buffer 200	Buffer 500	Buffer 500	Buffer 2000	Buffer 500	Buffer 4000
Joint (upper bound)	91.93±0.29		71.15±0.51		59.52±0.33	
Replay Riemer et al. (2019)	44.46±2.87	58.84±3.85	23.41±1.15	40.47±0.95	10.13±0.39	25.12±0.56
EWC Kirkpatrick et al. (2017)	48.65±2.16	60.34±1.12	25.56±1.12	42.75±1.23	13.16±1.53	26.78±0.47
GEM Lopez-Paz & Ranzato (2017)	49.38±4.32	61.67±1.32	26.15±1.86	43.25±1.51	14.45±1.25	26.98±0.56
Hare & Tortoise Lee et al. (2024)	50.25±0.58	61.23±0.32	25.28±0.56	45.26±0.81	16.18±0.95	25.29±0.62
FIRE Han et al. (2026)	46.53±2.34	59.52±1.34	24.56±1.51	42.18±1.86	12.25±1.13	22.26±0.45
WSC Cho et al. (2026)	52.56±0.86	61.25±0.52	25.65±0.64	42.56±0.47	15.26±0.86	25.51±1.12
PAI (Ours)	**54.25**±1.54	**63.62**±0.51	**27.92**±0.63	**47.18**±0.23	**18.36**±0.95	**27.21**±1.36

Baselines. We compare PAI with two complementary groups of methods. **(i) Continual learning methods** include Replay (Riemer et al., 2019), EWC (Kirkpatrick et al., 2017), GEM (Lopez-Paz & Ranzato, 2017), Hare & Tortoise (Lee et al., 2024), and WSC (Cho et al., 2026), covering experience replay, parameter regularization, gradient constraints, fast–slow learning, and weight consolidation. To assess compatibility with stronger rehearsal-based approaches, we additionally incorporate PAI into BFP (Gu et al., 2023), SARL (Sarfraz et al., 2025), and IDER (Liu et al., 2026). **(ii) Plasticity-oriented methods** include Shrink & Perturb (S&P) (Ash & Adams, 2020), DASH (Shin et al., 2024), FIRE (Han et al., 2026), L2 regularization (Kumar et al., 2025), and continual backpropagation (CBP) (Dohare et al., 2024) , spanning parameter perturbation, reinitialization, regularization, and feature resetting. Standard sequential backpropagation (BP/CL-SGD) serves as the warm-start baseline. Full experimental details are provided in the Appendix.

Continual Learning Settings. We consider three complementary regimes. **(i) Limited-replay CIL.** Following Gu et al. (2023); Boschini et al. (2022), we use 5 tasks on CIFAR-10 and 10 tasks on CIFAR-100 and Tiny-ImageNet, with disjoint classes, no task identities at inference, and ResNet-18 as the backbone. **(ii) All-data replay.** Following Lewandowski et al. (2024b), we introduce new classes over 20 stages while retaining all previously observed data, isolating plasticity from limited-memory effects. **(iii) Long-sequence CL.** Following Dohare et al. (2024), we construct 5,000 binary classification tasks from ImageNet by sampling two classes per task, directly evaluating plasticity over long task sequences. Results are reported over five random seeds, with further training details provided in the Appendix.

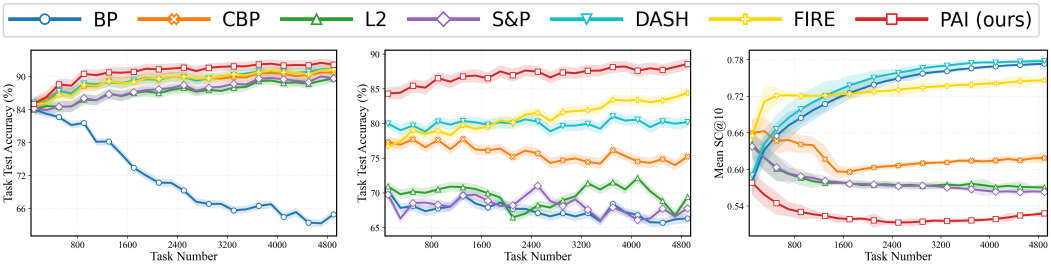
5.1 Stability–Plasticity Experiments

Comparison with Baseline Methods. Table 1 reports FAA under different replay budgets. PAI achieves the highest mean FAA across all evaluated datasets and buffer sizes, outperforming conventional CL methods and plasticity-oriented baselines. With the smaller buffers, it achieves 54.25%, 27.92%, and 18.36% on CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively, exceeding the strongest competing baseline by 1.69, 1.77, and 2.18 percentage points. Gains persist with larger buffers, supporting the effectiveness of historical-anchor initialization across replay budgets.

Compatibility with Rehearsal-based CL Methods. As shown in Table 2, incorporating PAI consistently improves the performance of Replay, BFP, SARL, and IDER across different datasets and memory budgets. The improvements are particularly pronounced for vanilla Replay: PAI improves FAA by 9.79 and 8.23 percentage points on CIFAR-10 and Tiny-ImageNet, respectively, under the smaller memory budgets. Importantly, consistent gains are also observed when PAI is combined with stronger rehearsal-based methods, including BFP and SARL, demonstrating that its effectiveness is not specific to a particular replay strategy. Furthermore, combining PAI with the recent plasticity-oriented ERID further improves its performance, with gains of 2.20 and 3.10 percentage points on CIFAR-10 and CIFAR-100, respectively, under the smaller memory budgets. These results demonstrate that PAI is a general and complementary strategy that can be readily integrated into existing CL methods to improve their stability–plasticity trade-off.

Table 2: Comparison of Final Average Accuracy (FAA) across different CL methods. All experiments are repeated 5 times with different seeds.

Method	CIFAR-10		CIFAR-100		Tiny-ImageNet	
	Buffer 200	Buffer 500	Buffer 500	Buffer 2000	Buffer 500	Buffer 4000
BFP Gu et al. (2023)	68.64±2.23	73.51±1.54	46.70±1.45	57.39±0.75	28.71±0.55	43.17±1.89
BFP + PAI	70.12±1.02	74.89±1.62	48.45±1.08	59.34±0.56	29.85±0.90	44.53±1.14
SARL Sarfraz et al. (2025)	68.87±1.37	73.98±0.46	46.69±0.79	57.06±0.48	28.44±2.30	38.83±0.81
SARL + PAI	70.43±1.41	74.56±0.78	47.58±0.56	59.26±0.50	29.67±1.78	39.71±0.95
IDER Liu et al. (2026)	71.02±1.98	74.74±0.42	44.82±0.85	56.59±0.35	29.88±1.15	43.05±1.40
IDER + PAI	73.22±1.01	76.86±0.67	47.92±0.45	58.18±0.43	30.86±0.87	44.54±1.23



(a) CNN-4, Current Task Acc. (b) ViT-Tiny, Current Task Acc. (c) ViT-Tiny, Spectral Concentration

Figure 3: Test Accuracy results on ImageNet under Long-sequence continual learning. We randomly sample two classes from the 1,000 ImageNet categories to form a binary classification task (total 5000 tasks).

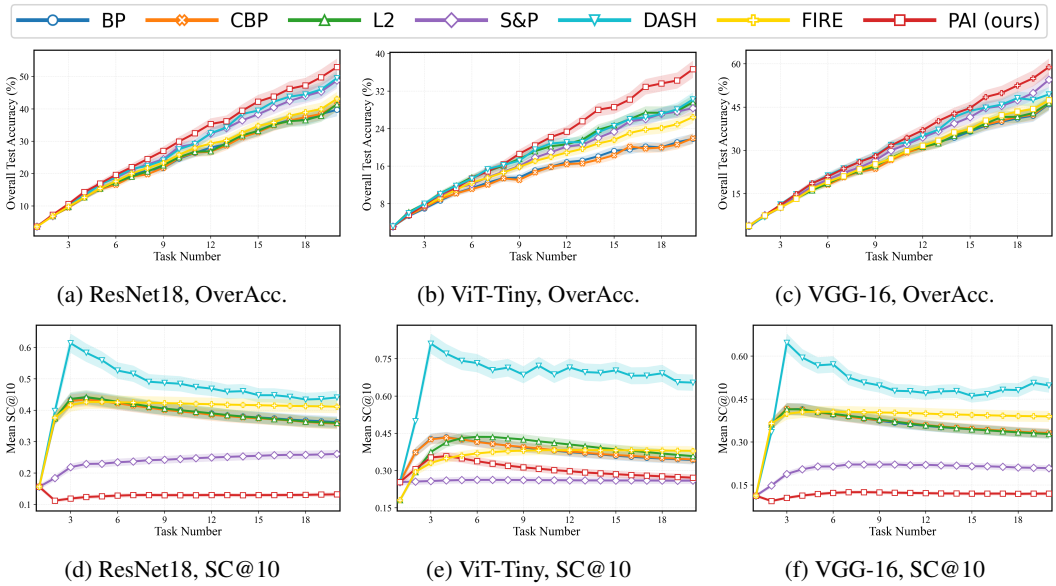
Table 3: Effect of β on 5-task CIFAR-10 with buffer 200 and 10-task CIFAR-100 with buffer 500.

	CIFAR-10								CIFAR-100							
β	0.0	0.3	0.5	0.7	0.8	0.85	0.9	1.0	0.0	0.3	0.5	0.7	0.8	0.85	0.9	1.0
Acc. (*all*)	44.46	48.62	50.23	52.36	53.62	**54.25**	54.06	44.56	23.41	24.23	25.06	26.12	27.23	**27.92**	27.65	24.56

5.2 Plasticity Experiments

Class-incremental learning with all-data replay. We evaluate PAI on CIFAR-100 over 20 phases, with five new classes introduced per phase and all previously observed data retained for replay (Rebuffi et al., 2017; Dohare et al., 2024; Lewandowski et al., 2024b). This setting removes the replay-budget constraint and evaluates whether task-boundary initialization remains beneficial when historical data are fully accessible. As shown in Figure 4, PAI consistently achieves the highest overall accuracy across ResNet-18, ViT-Tiny, and VGG-16, with increasingly pronounced gains in later phases. Meanwhile, PAI maintains substantially lower and more stable spectral concentration (SC@10), indicating that less spectral energy is concentrated in a few dominant singular directions. The consistent trend across CNN and Transformer architectures shows that the gains of PAI extend beyond limited-memory replay and are accompanied by a more balanced singular-value spectrum.

Long-sequence continual learning. We further evaluate PAI on ImageNet over 5,000 binary classification tasks, following prior studies on loss of plasticity (Lee et al., 2024). As shown in Figure 3, PAI maintains consistently high current-task accuracy on both CNN-4 and ViT-Tiny, whereas standard backpropagation progressively loses its ability to adapt to new tasks. On ViT-Tiny, PAI also maintains substantially lower and more stable spectral concentration (SC@10), in contrast to the progressive spectral concentration observed for several plasticity-oriented baselines such as FIRE and DASH. The simultaneous preservation of current-task accuracy and a less concentrated spectrum is consistent with our spectral hypothesis that limiting the dominance of a few singular directions is associated with sustained plasticity over long task sequences.



(a) ResNet18, OverAcc. (b) ViT-Tiny, OverAcc. (c) VGG-16, OverAcc.

(d) ResNet18, SC@10 (e) ViT-Tiny, SC@10 (f) VGG-16, SC@10

Figure 4: Overall accuracy on CIFAR-100 under class-incremental learning with replay, where the dataset is divided into 20 tasks (5 classes each) and each stage utilizes all previous data for training. Overall accuracy is the proportion of correctly classified test samples across all classes, including unseen classes.

Table 4: Ablation on task-boundary initialization strategies.

Initialization	**Formula**	**CIFAR-10**		**CIFAR-100**		**Tiny-ImageNet**	
		Buffer 200	Buffer 500	Buffer 500	Buffer 2000	Buffer 500	Buffer 4000
Previous endpoint	$y_K^t, y_0^{t+1} = y_K^t$	44.46±2.87	58.84±3.85	23.41±1.15	40.47±0.95	10.13±0.39	25.12±0.56
EMA initialization	$z^t = (1-\beta)y_K^t + \beta z^{t-1}, y_0^t = z^t$	49.26±1.25	61.32±1.42	26.21±1.09	45.47±0.85	16.13±0.45	26.62±0.48
SWA initialization	$z^t = (1-\frac{1}{t})z^{t-1} + \frac{1}{t}y_K^t, y_0^t = z^t$	44.56±1.31	60.56±1.27	24.56±1.31	44.25±0.56	15.67±0.25	26.45±0.67
PAI (Ours)	$y_0^t = (1-\beta)z^t + \beta x^t$	**54.25**±1.54	**63.62**±0.51	**27.92**±0.63	**47.18**±0.23	**18.36**±0.95	**27.21**±1.36

Initialization Ablation. Table 4 compares initialization strategies under identical within-task training procedures. While interpolating consecutive task endpoints improves over standard warm-starting, PAI achieves the highest mean FAA across all evaluated settings. These results support combining the latest endpoint with a historical anchor rather than only the preceding endpoint.

Effect of β. Table 3 examines the balance between the fast-memory anchor and the slow-memory anchor. At $\beta = 0$, initialization uses only fast memory, recovering standard sequential warm-starting (CL-SGD). At $\beta = 1$, it uses only slow memory (SWA initialization), which includes the latest endpoint but gives it no additional weight. Increasing β from 0 to 0.85 improves accuracy from 44.46% to 54.25% on CIFAR-10 and from 23.41% to 27.92% on CIFAR-100. However, using slow memory alone reduces accuracy to 44.56% and 24.56%, respectively. Among the evaluated values, $\beta = 0.85$ performs best on both datasets, supporting the combination of recent task knowledge in fast memory and accumulated historical knowledge in slow memory.

6 Conclusion

In this paper, we show that repeatedly warm-starting from the latest endpoint can lead to spectral concentration and reduced adaptability. Then we propose the Plasticity-Anchored Initialization (PAI), which initializes each task from an interpolation between the latest task endpoint and a historical anchor. This strategy preserves recent task knowledge while pulling optimization toward a more stable and plastic region. We further provide a stability-based analysis showing that PAI can achieve better algorithmic stability bound than conventional CL paradigms. Experiments across multiple continual learning settings demonstrate that PAI consistently improves the stability-plasticity trade-off and provides a simple, robust, and broadly applicable strategy for long-term continual learning.

REFERENCES

Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Learning fast, learning slow: A general continual learning method based on complementary learning system. *arXiv preprint arXiv:2201.12604*, 2022.

Jordan Ash and Ryan P Adams. On warm-starting neural network training. *Advances in neural information processing systems*, 33:3884–3894, 2020.

Tudor Berariu, Wojciech Czarnecki, Soham De, Jorg Bornschein, Samuel Smith, Razvan Pascanu, and Claudia Clopath. A study on the plasticity of neural networks. *arXiv preprint arXiv:2106.00042*, 2021.

Matteo Boschini, Lorenzo Bonicelli, Pietro Buzzega, Angelo Porrello, and Simone Calderara. Class-incremental continual learning into the extended der-verse. *IEEE transactions on pattern analysis and machine intelligence*, 45(5):5497–5512, 2022.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, P Dokania, P Torr, and M Ranzato. Continual learning with tiny episodic memories. In *Workshop on Multi-Task and Lifelong Reinforcement Learning*, 2019.

Dongkyu Cho, Taesup Moon, Rumi Chunara, Kyunghyun Cho, and Sungmin Cha. Forget forgetting: Continual learning in a world of abundant memory. In *International Conference on Learning Representations*, volume 2026, pp. 14441–14466, 2026.

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

Shibhansh Dohare, J Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A Rupam Mahmood, and Richard S Sutton. Loss of plasticity in deep continual learning. *Nature*, 632(8026): 768–774, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Mohamed Elsayed and A Rupam Mahmood. Addressing loss of plasticity and catastrophic forgetting in continual learning. *arXiv preprint arXiv:2404.00781*, 2024.

Mohamed Elsayed, Qingfeng Lan, Clare Lyle, and A Rupam Mahmood. Weight clipping for deep continual and reinforcement learning. *arXiv preprint arXiv:2407.01704*, 2024.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, pp. 3762–3773. PMLR, 2020.

Qiao Gu, Dongsub Shim, and Florian Shkurti. Preserving linear separability in continual learning by backward feature projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24286–24295, 2023.

Isaac Han, Sangyeon Park, Seungwon Oh, Donghu Kim, Hojoon Lee, and Kyung-Joong Kim. Fire: Frobenius-isometry reinitialization for balancing the stability-plasticity tradeoff. *arXiv preprint arXiv:2602.08040*, 2026.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 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*, pp. 770–778, 2016.

Qiang He, Tianyi Zhou, Meng Fang, and Setareh Maghsudi. Adaptive regularization of representation rank as an implicit constraint of bellman equation. *arXiv preprint arXiv:2404.12754*, 2024.

Jifeng Hu, Sili Huang, Li Shen, Zhejian Yang, Shengchao Hu, Shisong Tang, Hechang Chen, Lichao Sun, Yi Chang, and Dacheng Tao. Tackling continual offline rl through selective weights activation on aligned spaces. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Jifeng Hu, Li Shen, Sili Huang, Zhejian Yang, Hechang Chen, Lichao Sun, Yi Chang, and Dacheng Tao. Continual diffuser (cod): Mastering continual offline rl with experience rehearsal. *IEEE Transactions on Neural Networks and Learning Systems*, 2025b.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

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

Aviral Kumar, Rishabh Agarwal, Dibya Ghosh, and Sergey Levine. Implicit under-parameterization inhibits data-efficient deep reinforcement learning. In *International Conference on Learning Representations*, 2021a.

Aviral Kumar, Rishabh Agarwal, Tengyu Ma, Aaron Courville, George Tucker, and Sergey Levine. Dr3: Value-based deep reinforcement learning requires explicit regularization. *arXiv preprint arXiv:2112.04716*, 2021b.

Saurabh Kumar, Henrik Marklund, and Benjamin Van Roy. Maintaining plasticity in continual learning via regenerative regularization. In *Conference on Lifelong Learning Agents*, pp. 410–430. PMLR, 2025.

Hojoon Lee, Hyeonseo Cho, Hyunseung Kim, Donghu Kim, Dugki Min, Jaegul Choo, and Clare Lyle. Slow and steady wins the race: Maintaining plasticity with hare and tortoise networks. In *International Conference on Machine Learning*, pp. 26416–26438. PMLR, 2024.

Yunwen Lei and Yiming Ying. Fine-grained analysis of stability and generalization for stochastic gradient descent. In *International Conference on Machine Learning*, pp. 5809–5819. PMLR, 2020.

Alex Lewandowski, Michał Bortkiewicz, Saurabh Kumar, András György, Dale Schuurmans, Mateusz Ostaszewski, and Marlos C Machado. Learning continually by spectral regularization. *arXiv preprint arXiv:2406.06811*, 2024a.

Alex Lewandowski, Dale Schuurmans, and Marlos C Machado. Plastic learning with deep fourier features. *arXiv preprint arXiv:2410.20634*, 2024b.

Zhanwang Liu, Yuting Li, Haoyuan Gao, Yexin Li, Linghe Kong, Lichao Sun, and Weiran Huang. Ider: Idempotent experience replay for reliable continual learning. *arXiv preprint arXiv:2603.00624*, 2026.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Clare Lyle, Mark Rowland, and Will Dabney. Understanding and preventing capacity loss in reinforcement learning. In *International Conference on Learning Representations*, 2022.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In *International Conference on Machine Learning*, pp. 23190–23211. PMLR, 2023.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. The primacy bias in deep reinforcement learning. In *International conference on machine learning*, pp. 16828–16847. PMLR, 2022.

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019.

Sangyeon Park, Isaac Han, Seungwon Oh, and KyungJoong Kim. Activation by interval-wise dropout: A simple way to prevent neural networks from plasticity loss. In *Forty-second International Conference on Machine Learning*, 2025.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *ICLR (Poster)*, 2019.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Sparse coding in a dual memory system for lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9714–9722, 2023.

Fahad Sarfraz, Elahe Arani, and Bahram Zonooz. Semantic aware representation learning for lifelong learning. In *The Thirteenth International Conference on Learning Representations*, 2025.

Baekrok Shin, Junsoo Oh, Hanseul Cho, and Chulhee Yun. Dash: Warm-starting neural network training in stationary settings without loss of plasticity. *Advances in Neural Information Processing Systems*, 37:43300–43340, 2024.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *International Conference on Machine Learning*, pp. 32145–32168. PMLR, 2023.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 139–149, 2022.

Lai Wei, Yuting Li, Chen Wang, Yue Wang, Linghe Kong, Weiran Huang, and Lichao Sun. First sft, second rl, third upt: Continual improving multi-modal llm reasoning via unsupervised post-training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2087–2098, 2025.

Zhiling Yan, Dingjie Song, Zhe Fang, Yisheng Ji, Xiang Li, Quanzheng Li, and Lichao Sun. Livemedbench: A contamination-free medical benchmark for llms with automated rubric evaluation. *arXiv preprint arXiv:2602.10367*, 2026.

Dezhong Yao, Sanmu Li, Yutong Dai, Zhiqiang Xu, Shengshan Hu, Peilin Zhao, and Lichao Sun. Variational bayes for federated continual learning. *arXiv preprint arXiv:2405.14291*, 2024.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. PMLR, 2017.

ETHICS STATEMENT

This work studies Plasticity-Anchored Initialization (PAI) for continual learning using existing visual benchmarks: CIFAR-10, CIFAR-100, Tiny-ImageNet, and ImageNet. The experiments use benchmark datasets rather than collecting new data from users. Although improved adaptation and knowledge retention may benefit systems operating on changing data streams, benchmark performance does not establish fairness, privacy, or safety in deployment. Models may retain biases present in their training data, and replay-based implementations require appropriate safeguards for stored examples. The stability analysis characterizes sensitivity to single-example replacement under the stated assumptions; it does not provide a differential privacy guarantee. Applications in safety-critical settings require additional validation and application-specific safeguards.

REPRODUCIBILITY STATEMENT

The paper specifies the PAI initialization rule, including the fast-memory anchor, historical endpoint averaging for the slow-memory anchor, and the interpolation coefficient β. The main text and appendices state the assumptions and provide detailed proofs of the stability and generalization bounds for the final task endpoint y_K^T. These results concern the stated sequential, single-example SGD protocol; extensions to replay, mini-batch updates, momentum, adaptive optimizers, and gradient clipping require separate analyses. The experimental sections describe the datasets, model architectures, task settings, replay budgets, baseline methods, and hyperparameter configurations. Evaluations cover class-incremental learning with limited replay, all-data replay, and 5,000-task ImageNet sequences across CNN-4, VGG-16, ResNet-18, and ViT-Tiny. The paper also reports ablations of the initialization strategy and anchoring coefficient, together with spectral diagnostics. Continual visual learning results are reported over five random seeds. We will publicly release the implementation and experiment scripts upon acceptance.

AI USE STATEMENT

Large language models assisted with language editing, manuscript organization, and drafting supplementary material. They also helped identify potential inconsistencies, examine mathematical formulations, and suggest additional experiments and evaluation protocols. AI-generated content does not constitute independent evidence for theoretical or empirical claims. The authors are responsible for verifying all incorporated material and for the accuracy and integrity of the final manuscript, including its proofs, experimental results, and citations.

Broader Impacts and Limitations. This work studies task-boundary initialization for continual learning. Improved adaptation and knowledge retention may benefit systems operating on changing data streams, but benchmark improvements do not establish reliability, fairness, or safety in deployment. Our evaluation is limited to supervised visual benchmarks and the architectures and task protocols considered in this paper. The theoretical results concern sensitivity to single-example replacement under the stated assumptions; they do not directly guarantee reduced forgetting or improved plasticity. Applications to safety-critical settings require additional validation.

APPENDIX CONTENTS

A Generalization Analysis for Sequential SGD 15

A.1 Setup, sampling protocol, and stability definition 15

A.2 Shared assumptions and terminal loss control 16

A.3 From stability to generalization 17

A.4 Perturbation creation and propagation 18

A.5 Explicit bounds for a harmonic learning rate 21

B Generalization Analysis for PAI 25

B.1 Dual-anchor recursion and final output 25

B.2 Stability of the final task endpoint 25

B.3 Harmonic learning rates and discrete optimization 29

B.4 Comparison with sequential SGD 30

B.5 Scope and boundary cases 33

C Additional Experimental Details 33

C.1 Benchmarks and evaluation protocols 33

C.2 Additional forgetting results 33

C.3 Training configuration 34

C.4 Hyperparameter search spaces 34

C.5 Selected hyperparameters: continual setting 35

C.6 Selected hyperparameters: class-incremental setting 35

C.7 Selected hyperparameters: long-sequence learning 35

Algorithm 2 Sequential SGD analyzed in this appendix

Require: initial model w^0, task sequence $\{S_1, S_2, \ldots, S_T\}$, local update steps K, learning-rate schedule $\{\eta_{t,k}\}$
Ensure: final model parameters w^T
Set w^0 to the data-independent initial model.
for $t = 1, 2, \ldots, T$ **do**
  receive current task dataset $S_t = (z_{t,j})_{j=1}^n$
  initialize task model $w_{t,0} \leftarrow w^{t-1}$
  for $k = 1, 2, \ldots, K$ **do**
    independently sample $j_{t,k} \sim \mathrm{Unif}([n])$ with replacement
    compute stochastic gradient $g_{t,k} = \nabla \ell(w_{t,k-1}; z_{t,j_{t,k}})$
    update model: $w_{t,k} = w_{t,k-1} - \eta_{t,k} g_{t,k}$
  end for
  set $w^t \leftarrow w_{t,K}$
end for
return w^T

A Generalization Analysis for Sequential SGD

A.1 Setup, sampling protocol, and stability definition

Let $[q] = \{1, \ldots, q\}$ for a positive integer q, and let $T, K \geq 1$ and $n \geq 2$ be integers. Task t has a fixed distribution $\mathcal{D}_t$ and an ordered training dataset $S_t = (z_{t,1}, \ldots, z_{t,n})$, with $z_{t,j} \sim \mathcal{D}_t$. All samples are mutually independent, including across tasks, although the task distributions may differ. Write $\mathcal{S}^{(T)} = (S_1, \ldots, S_T)$ and $\mathcal{Z} = \bigcup_{t=1}^T \mathrm{supp}(\mathcal{D}_t)$. An admissible dataset has each $z_{t,j} \in \mathrm{supp}(\mathcal{D}_t)$. All losses and algorithm outputs are assumed measurable.

Initialization is independent of the training data and sampling indices. Each update independently samples one current-task example uniformly with replacement. We use a deterministic, nonnegative global learning-rate schedule $\{\eta_m\}_{m=1}^{TK}$. Sequential SGD uses

$$w_{t,0} = w_{t-1,K}, \qquad w_{t,k} = w_{t,k-1} - \eta_{(t-1)K+k} \nabla \ell(w_{t,k-1}; z_{t,j_{t,k}}), \quad k \in [K],$$

where $w_{0,K} = w^0$. PAI uses the same global schedule, with $\eta_{t,k}^{\mathrm{PAI}} = \eta_{(t-1)K+k+1}$ for $k = 0, \ldots, K-1$, and a fixed data-independent $\beta \in [0, 1]$. Thus $\eta_{t,k+1}^{\mathrm{CL}} = \eta_{t,k}^{\mathrm{PAI}}$. The harmonic schedule is $\eta_m = c/m$, where $c > 0$.

Neighboring datasets and coupling. Two admissible sequences are neighbors in task τ if they differ only at one index (τ, κ), with both values in $\mathrm{supp}(\mathcal{D}_\tau)$. Stability statements hold for every such fixed pair. We couple the runs using the same initialization and sampling indices. Independent replacements are introduced only when passing from stability to expected generalization.

For an output $W = \mathcal{A}(\mathcal{S}^{(T)}; \omega)$, where ω denotes all data-independent algorithmic randomness, define

$$\mathcal{L}_{\mathcal{D}^{(T)}}(w) = \frac{1}{T} \sum_{t=1}^T \mathbb{E}_{z \sim \mathcal{D}_t}[\ell(w; z)],$$

$$\widehat{\mathcal{L}}_{\mathcal{S}^{(T)}}(w) = \frac{1}{Tn} \sum_{t=1}^T \sum_{j=1}^n \ell(w; z_{t,j}),$$

$$\mathcal{E}_{\mathrm{gen}} = \mathbb{E}_{\mathcal{S}^{(T)}, \omega} \left[\mathcal{L}_{\mathcal{D}^{(T)}}(W) - \widehat{\mathcal{L}}_{\mathcal{S}^{(T)}}(W) \right].$$

The analyzed outputs are $w_{T,K}$ for CL-SGD and $y_K^T = z^{T+1}$ for PAI.

Definition A.1 (Task-dependent uniform stability)**.** *Under the shared-randomness coupling, an algorithm has task-dependent stability coefficients $\epsilon_1, \ldots, \epsilon_T$ if every fixed neighboring pair differing in task t satisfies*

$$\sup_{z \in \mathcal{Z}} \mathbb{E}_\omega \left[\left| \ell(\mathcal{A}(\mathcal{S}^{(T)}; \omega); z) - \ell(\mathcal{A}(\widetilde{\mathcal{S}}^{(T)}; \omega); z) \right| \right] \leq \epsilon_t.$$

In particular, $\max_{t\in[T]} \epsilon_t$ *is a uniform-stability coefficient over all replacement tasks. The expectation of the absolute loss difference also bounds the absolute difference of expected losses.*

A.2 Shared assumptions and terminal loss control

Assumption A.2 (Sample-wise smoothness)**.** *The loss* $\ell(\cdot; z) : \mathbb{R}^d \to \mathbb{R}$ *is continuously differentiable for every* $z \in \mathcal{Z}$*, and there is a deterministic* $L > 0$ *such that*

$$\|\nabla\ell(w; z) - \nabla\ell(v; z)\| \le L\|w - v\|, \qquad w, v \in \mathbb{R}^d, \quad z \in \mathcal{Z}. \tag{8}$$

All parameter and gradient norms are Euclidean.

Assumption A.3 (Uniform within-task gradient discrepancy)**.** *For each* $t \in [T]$*, a deterministic constant* $\sigma_t \ge 0$ *satisfies*

$$\|\nabla\ell(w; z) - \nabla\ell(w; z')\| \le 2\sigma_t, \qquad w \in \mathbb{R}^d, \quad z, z' \in \mathrm{supp}(\mathcal{D}_t). \tag{9}$$

This is a bound on sample-to-sample gradient differences, not on gradient magnitudes and not merely on population gradient variance. The stronger pointwise deviation condition

$$\|\nabla\ell(w; z) - \nabla F_t(w)\| \le \sigma_t, \qquad w \in \mathbb{R}^d, \quad z \in \mathrm{supp}(\mathcal{D}_t), \tag{10}$$

also suffices whenever F_t and its gradient are well-defined: apply the triangle inequality at the same w. No interchange of differentiation and expectation is needed in the pairwise formulation. The bounds involving σ_t are separate assumptions and are not implied by the terminal-region bounds below.

Assumption A.4 (Uniform terminal-region control)**.** *For the fixed horizon and deterministic learning-rate schedule, there exist a deterministic convex set* $\mathcal{C}_T \subseteq \mathbb{R}^d$ *and deterministic constants* $L_G \ge 0$ *and* $U > 0$ *such that:*

1. *For every admissible deterministic training sequence, the final output of CL-SGD and the final output of PAI belong to* $\mathcal{C}_T$ *almost surely over algorithmic randomness. For comparisons across* β*, this same set and the same constants apply to every* $\beta \in [0, 1]$*.*
2. *For all* $v \in \mathcal{C}_T$ *and all test examples* $z \in \mathcal{Z}$*,*

$$\|\nabla\ell(v; z)\| \le L_G, \qquad 0 \le \ell(v; z) \le U. \tag{11}$$

The constants and set must be uniform over datasets, replacements, test examples, and random seeds (up to the stated probability-one qualification). A neighborhood chosen after inspecting one realized output does not suffice for uniform stability. No bounded-loss or bounded-gradient condition is imposed outside $\mathcal{C}_T$, and intermediate training states need not lie in $\mathcal{C}_T$. The region is an assumption about the unmodified algorithm; no projection or clipping step is being added.

Lemma A.5 (Loss control between terminal outputs)**.** *Under Assumptions A.2 and A.4, for every* $w, v \in \mathcal{C}_T$ *and* $z \in \mathcal{Z}$*,*

$$|\ell(w; z) - \ell(v; z)| \le L_G\|w - v\|. \tag{12}$$

In particular, for any coupled neighboring final outputs $W, \widetilde{W}$*,*

$$|\ell(W; z) - \ell(\widetilde{W}; z)| \le \min\{U, L_G\|W - \widetilde{W}\|\} \quad \textit{almost surely}. \tag{13}$$

Proof. Convexity implies $v + s(w - v) \in \mathcal{C}_T$ for $s \in [0, 1]$. The fundamental theorem of calculus and Cauchy–Schwarz give

$$\begin{aligned} |\ell(w; z) - \ell(v; z)| &= \left| \int_0^1 \langle \nabla\ell(v + s(w - v); z), w - v\rangle \, ds \right| \\ &\le \int_0^1 \|\nabla\ell(v + s(w - v); z)\| \, \|w - v\| \, ds \\ &\le L_G\|w - v\|. \end{aligned}$$

Each of the two coupled outputs is in $\mathcal{C}_T$ with probability one, so both are in that set with probability one under their joint coupling. Their losses lie in $[0, U]$, hence their absolute difference is also at most U. This proves equation 13. The argument bounds gradients along the whole connecting segment, not merely at one endpoint. □

Relation to global boundedness. Global smoothness together with $0 \le \ell(w; z) \le U$ on all of $\mathbb{R}^d$ implies a global gradient bound. Indeed, the descent lemma gives

$$0 \le \ell\left(w - \frac{1}{L}\nabla\ell(w; z); z\right) \le \ell(w; z) - \frac{1}{2L}\|\nabla\ell(w; z)\|^2,$$

so $\|\nabla\ell(w; z)\| \le \sqrt{2LU}$ globally. Integrating along segments then implies global Lipschitz continuity. Restricting both loss control and gradient control to the terminal region avoids this implicit global restriction while retaining the event-decomposition proof below.

Proposition A.6 (A concrete sufficient condition for terminal control)**.** *Suppose Assumption A.2 holds. Let $w_\circ \in \mathbb{R}^d$, $R \ge 0$, $G_\circ \ge 0$, and $M_\circ \ge 0$ be deterministic and common to the algorithms and anchoring weights under comparison. Suppose every final output lies almost surely in $\overline{B}(w_\circ, R)$ for every admissible dataset. Suppose additionally, for every $z \in \mathcal{Z}$,*

$$\|\nabla\ell(w_\circ; z)\| \le G_\circ, \qquad 0 \le \ell(w_\circ; z) \le M_\circ,$$

and $\ell(v; z) \ge 0$ on this ball. Then Assumption A.4 holds with

$$\mathcal{C}_T = \overline{B}(w_\circ, R), \qquad L_G = G_\circ + LR, \qquad U = 1 + M_\circ + G_\circ R + \frac{L}{2}R^2. \tag{14}$$

The added 1 merely ensures $U > 0$ and may be omitted when the remaining expression is positive.

Proof. For v in the ball, smoothness gives

$$\|\nabla\ell(v; z)\| \le \|\nabla\ell(w_\circ; z)\| + L\|v - w_\circ\| \le G_\circ + LR.$$

The descent lemma at $w_\circ$ gives

$$\begin{aligned}\ell(v; z) &\le \ell(w_\circ; z) + \langle\nabla\ell(w_\circ; z), v - w_\circ\rangle + \frac{L}{2}\|v - w_\circ\|^2 \\ &\le M_\circ + G_\circ R + \frac{L}{2}R^2.\end{aligned}$$

Combine this with the assumed nonnegativity and terminal containment. The ball is convex, completing the proof. These reference-point and containment conditions are sufficient conditions, not automatic consequences of a smooth loss or of convergence in expectation. □

Example: quadratic losses. Let $\ell(w; z) = \frac{1}{2}(w - z)^2$ on $\mathbb{R}$, with $z \in [-B, B]$ and $B > 0$. It is 1-smooth but is neither globally Lipschitz nor globally bounded. Its gradient discrepancy obeys $|\nabla\ell(w; z) - \nabla\ell(w; z')| = |z - z'| \le 2B$, so take $\sigma_t = B$. If $w^0 \in [-B, B]$ almost surely and $0 \le \eta_{t,k} \le 1$, every SGD update is $(1 - \eta_{t,k})w + \eta_{t,k}z \in [-B, B]$ whenever $w \in [-B, B]$. PAI's averaging and initialization are also convex combinations. Induction therefore places all endpoints of both algorithms, for every $\beta \in [0, 1]$, in $\mathcal{C}_T = [-B, B]$. On this region $|\nabla\ell(w; z)| \le 2B$ and $0 \le \ell(w; z) \le 2B^2$. Thus the shared assumptions hold with $L = 1$, $L_G = 2B$, $U = 2B^2$, and $\sigma_t = B$, despite failure of both corresponding global loss conditions.

Scope of terminal control. Terminal-region control is used only to convert the final parameter discrepancy into a loss difference and to bound the contribution of the complementary event. The update recursions use smoothness and within-task gradient discrepancy. The constants L_G and U may depend on the fixed horizon and schedule; any asymptotic interpretation must include this dependence. Comparisons use the same constants and terminal region for both algorithms and all anchoring weights under consideration.

A.3 From stability to generalization

Lemma A.7 (Task-dependent stability implies generalization)**.** *Suppose an algorithm has task-dependent stability coefficients $\epsilon_1, \ldots, \epsilon_T$ and its terminal losses satisfy Assumption A.4. Then*

$$|\mathcal{E}_{\mathrm{gen}}| \le \frac{1}{T}\sum_{t=1}^{T}\epsilon_t.$$

Proof. For each (t, j), let $z'_{t,j} \sim \mathcal{D}_t$ be an independent copy of $z_{t,j}$, and let $\mathcal{S}^{(T),(t,j)}$ replace only this sample. Writing $\mathcal{S} = \mathcal{S}^{(T)}$ for brevity gives

$$\mathcal{E}_{\text{gen}} = \frac{1}{Tn} \sum_{t=1}^{T} \sum_{j=1}^{n} \mathbb{E}\left[\ell(\mathcal{A}(\mathcal{S}; \omega); z'_{t,j}) - \ell(\mathcal{A}(\mathcal{S}; \omega); z_{t,j})\right]. \tag{15}$$

The pair $(z_{t,j}, z'_{t,j})$ is exchangeable and independent of all remaining data and ω. Consequently,

$$\mathbb{E}\left[\ell(\mathcal{A}(\mathcal{S}; \omega); z'_{t,j})\right] = \mathbb{E}\left[\ell(\mathcal{A}(\mathcal{S}^{(T),(t,j)}; \omega); z_{t,j})\right]. \tag{16}$$

Substitution yields

$$\mathcal{E}_{\text{gen}} = \frac{1}{Tn} \sum_{t=1}^{T} \sum_{j=1}^{n} \mathbb{E}\left[\ell(\mathcal{A}(\mathcal{S}^{(T),(t,j)}; \omega); z_{t,j}) - \ell(\mathcal{A}(\mathcal{S}; \omega); z_{t,j})\right]. \tag{17}$$

Conditional on the data and replacement, the two datasets are fixed neighbors and $z_{t,j} \in \mathcal{Z}$. Uniformity over test examples therefore allows the task-t stability bound to be applied to the expectation over ω. The triangle inequality gives

$$|\mathcal{E}_{\text{gen}}| \leq \frac{1}{Tn} \sum_{t=1}^{T} \sum_{j=1}^{n} \epsilon_t = \frac{1}{T} \sum_{t=1}^{T} \epsilon_t.$$

All expectations are finite by terminal loss boundedness. □

A.4 Perturbation creation and propagation

Fix a neighboring pair of continual learning sequences $\mathcal{S}^{(T)}$ and $\widetilde{\mathcal{S}}^{(T)}$ that differ in exactly one sample, namely the κ-th sample of task τ. We couple the two algorithmic runs by using the same random sampling indices $\{j_{t,k}\}$.

Fix an observation point $k_0 \in \{0, 1, \ldots, K\}$ inside the perturbed task τ. Define the first hitting time of the perturbed sample as

$$I_{\tau,\kappa} := \min\{k \in [K] : j_{\tau,k} = \kappa\},$$

with the convention $I_{\tau,\kappa} = +\infty$ if the perturbed sample is never selected. Define the event

$$\mathcal{E}_{\tau,\kappa,k_0} := \{I_{\tau,\kappa} > k_0\}.$$

That is, the perturbed sample is not selected during the first k_0 stochastic updates of task τ.

Since each stochastic update samples one example uniformly with replacement from S_τ, we have

$$\mathbb{P}\left(\mathcal{E}^c_{\tau,\kappa,k_0}\right) = 1 - \left(1 - \frac{1}{n}\right)^{k_0} \leq \frac{k_0}{n}. \tag{18}$$

Lemma A.8 (Decomposition for continual learning)**.** *Under Assumptions A.2–A.4, we consider two neighboring continual learning sequences $\mathcal{S}^{(T)}$ and $\widetilde{\mathcal{S}}^{(T)}$ that differ in exactly one sample, namely the κ-th sample of task τ. Let*

$$w_{T,K} := \mathcal{A}(\mathcal{S}^{(T)}), \qquad \widetilde{w}_{T,K} := \mathcal{A}(\widetilde{\mathcal{S}}^{(T)})$$

be the corresponding output models generated under coupled randomness. Then, for every test sample $z \in \mathcal{Z}$ and every $k_0 \in \{0, 1, \ldots, K\}$,

$$\mathbb{E}\big[\big|\ell(w_{T,K}; z) - \ell(\widetilde{w}_{T,K}; z)\big|\big] \leq \frac{Uk_0}{n} + L_G\, \delta_{\tau,k_0 \to T},$$

where

$$\delta_{\tau,k_0 \to T} := \mathbb{E}[\|w_{T,K} - \widetilde{w}_{T,K}\| \,|\, \mathcal{E}_{\tau,\kappa,k_0}].$$

The expectation is taken over the randomness of the algorithm and the sampling process.

Proof. For a fixed test sample $z \in \mathcal{Z}$, define

$$\Delta_z := \big|\ell(w_{T,K}; z) - \ell(\widetilde{w}_{T,K}; z)\big|.$$

By conditioning on the event $\mathcal{E}_{\tau,\kappa,k_0}$, we have

$$\mathbb{E}[\Delta_z] = \mathbb{E}\big[\mathbf{1}_{\mathcal{E}_{\tau,\kappa,k_0}} \Delta_z\big] + \mathbb{E}\Big[\mathbf{1}_{\mathcal{E}^c_{\tau,\kappa,k_0}} \Delta_z\Big].$$

By Assumption A.4, both final outputs belong to $\mathcal{C}_T$ almost surely, and their losses lie in $[0, U]$. This is a terminal-region bound, not a global loss bound. Therefore,

$$\Delta_z \le U.$$

Hence,

$$\mathbb{E}\Big[\mathbf{1}_{\mathcal{E}^c_{\tau,\kappa,k_0}} \Delta_z\Big] \le U\,\mathbb{P}\left(\mathcal{E}^c_{\tau,\kappa,k_0}\right) \le \frac{Uk_0}{n}.$$

On the event $\mathcal{E}_{\tau,\kappa,k_0}$, Lemma A.5 applies to the two final outputs, and gives

$$\Delta_z \le L_G \|w_{T,K} - \widetilde{w}_{T,K}\|.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\big[\mathbf{1}_{\mathcal{E}_{\tau,\kappa,k_0}} \Delta_z\big] &\le L_G \mathbb{E}\big[\mathbf{1}_{\mathcal{E}_{\tau,\kappa,k_0}} \|w_{T,K} - \widetilde{w}_{T,K}\|\big] \\
&= L_G \mathbb{P}\left(\mathcal{E}_{\tau,\kappa,k_0}\right) \mathbb{E}[\|w_{T,K} - \widetilde{w}_{T,K}\| \,|\, \mathcal{E}_{\tau,\kappa,k_0}] \\
&\le L_G \delta_{\tau,k_0\to T}.
\end{aligned}$$

Combining the two bounds gives

$$\mathbb{E}\big[\big|\ell(w_{T,K}; z) - \ell(\widetilde{w}_{T,K}; z)\big|\big] \le \frac{Uk_0}{n} + L_G\,\delta_{\tau,k_0\to T}.$$

This completes the proof. □

Lemma A.9 (Same-sample recursion)**.** *Under Assumptions A.2–A.4, we consider two coupled runs of the continual learning algorithm. If at step (t, k) the two runs use the same sample $z \in \mathcal{Z}$, then*

$$\|w_{t,k} - \widetilde{w}_{t,k}\| \le (1 + \eta_{t,k} L)\|w_{t,k-1} - \widetilde{w}_{t,k-1}\|.$$

Consequently, the same inequality holds after taking expectation or conditional expectation with respect to any event on which the coupled step uses the same sample.

Proof. If the same sample z is used in both runs at step (t, k), then

$$\begin{aligned}
\|w_{t,k} - \widetilde{w}_{t,k}\| &= \big\|w_{t,k-1} - \eta_{t,k}\nabla\ell(w_{t,k-1}; z) - \widetilde{w}_{t,k-1} + \eta_{t,k}\nabla\ell(\widetilde{w}_{t,k-1}; z)\big\| \\
&\le \|w_{t,k-1} - \widetilde{w}_{t,k-1}\| + \eta_{t,k}\|\nabla\ell(w_{t,k-1}; z) - \nabla\ell(\widetilde{w}_{t,k-1}; z)\| \\
&\le (1 + \eta_{t,k} L)\|w_{t,k-1} - \widetilde{w}_{t,k-1}\|,
\end{aligned}$$

The expectation and conditional-expectation statements follow from the same pointwise inequality. □

Lemma A.10 (Different-sample recursion)**.** *Under Assumptions A.2–A.4, consider two coupled runs of the continual learning algorithm. If at step (t, k) the two runs use two different samples $z, \widetilde{z} \in \mathrm{supp}(\mathcal{D}_t)$ from the same task t, then*

$$\|w_{t,k} - \widetilde{w}_{t,k}\| \le (1 + \eta_{t,k} L)\|w_{t,k-1} - \widetilde{w}_{t,k-1}\| + 2\eta_{t,k}\sigma_t.$$

Consequently, the same inequality holds after taking expectation or conditional expectation.

Proof. Let $G_{\eta,z}(w) = w - \eta\nabla\ell(w; z)$. Insert $G_{\eta,z}(v)$ between the two updated states. By Assumptions A.2 and A.3, for arbitrary w, v,

$$\begin{aligned}
\|G_{\eta,z}(w) - G_{\eta,\widetilde{z}}(v)\| &\le \|G_{\eta,z}(w) - G_{\eta,z}(v)\| + \eta\|\nabla\ell(v; z) - \nabla\ell(v; \widetilde{z})\| \\
&\le (1 + \eta L)\|w - v\| + 2\eta\sigma_t.
\end{aligned}$$

This is a pointwise inequality. Substituting the two current iterates and taking expectations proves the lemma, including conditional expectations on any positive-probability event. Repeated use of training examples does not require independence between the examples and the current states. No loss Lipschitzness, loss bound, or population-variance argument is used in this step. □

Theorem A.11 (Task-dependent stability of CL-SGD)**.** *Under Assumptions A.2–A.4 and the stated sampling protocol, let $\eta_m \geq 0$ be deterministic. Define*

$$A_u = \prod_{m=(u-1)K+1}^{uK} (1 + L\eta_m), \qquad P_{t,T} = \prod_{u=t+1}^{T} A_u,$$

and, for $k \in \{0, \ldots, K\}$, define

$$d_t(k) = \frac{2\sigma_t}{nL} \left[\prod_{m=(t-1)K+k+1}^{tK} (1 + L\eta_m) - 1 \right].$$

Empty products equal one. For every neighboring pair differing in task t, the coupled final endpoints satisfy

$$\sup_{z \in \mathcal{Z}} \mathbb{E}_{\mathcal{A}} \left[|\ell(w_{T,K}; z) - \ell(\widetilde{w}_{T,K}; z)| \right] \leq \varepsilon_t^{\text{CL-SGD}}(k), \tag{19}$$

where

$$\varepsilon_t^{\text{CL-SGD}}(k) := \frac{Uk}{n} + L_G d_t(k) P_{t,T}. \tag{20}$$

Consequently,

$$|\mathcal{E}_{\text{gen}}^{\text{SGD}}| \leq \frac{1}{T} \sum_{t=1}^{T} \min_{k \in \{0,\ldots,K\}} \varepsilon_t^{\text{CL-SGD}}(k). \tag{21}$$

If $\eta_m = c/m$ with $c > 0$, then for $k \in [K]$,

$$\varepsilon_t^{\text{CL-SGD}}(k) \leq \frac{Uk}{n} + \frac{2L_G\sigma_t}{nL} \left[\left(\frac{TK}{(t-1)K + k} \right)^{cL} - 1 \right]. \tag{22}$$

Proof. Fix neighbors differing at (τ, κ) and $k_0 \in \{0, \ldots, K\}$. Condition on $\mathcal{E} = \mathcal{E}_{\tau,\kappa,k_0}$, which has positive probability because $n \geq 2$. Let $\delta_{t,k} = \mathbb{E}_{\mathcal{A}}[\|w_{t,k} - \widetilde{w}_{t,k}\| \mid \mathcal{E}]$. The runs agree before task τ and through its first k_0 updates on this event, so $\delta_{\tau,k_0} = 0$.

For $k > k_0$, the shared sampling index is independent of all preceding indices and the shared initialization. In particular, conditional on the past, its probability of equaling κ is $1/n$. The same-sample and different-sample inequalities give, pathwise,

$$\|w_{\tau,k} - \widetilde{w}_{\tau,k}\| \leq (1 + L\eta_{\tau,k})\|w_{\tau,k-1} - \widetilde{w}_{\tau,k-1}\| + 2\eta_{\tau,k}\sigma_\tau \mathbf{1}_{\{j_{\tau,k} = \kappa\}}.$$

Here $\eta_{\tau,k} = \eta_{(\tau-1)K+k}$. Taking conditional expectations and adding $2\sigma_\tau/(nL)$ yields

$$\delta_{\tau,k} + \frac{2\sigma_\tau}{nL} \leq (1 + L\eta_{\tau,k}) \left(\delta_{\tau,k-1} + \frac{2\sigma_\tau}{nL} \right). \tag{23}$$

Iteration from the zero discrepancy at k_0 gives

$$\delta_{\tau,K} \leq d_\tau(k_0). \tag{24}$$

For each later task $u > \tau$, both datasets and sampled indices agree, so its updates amplify the discrepancy by at most A_u. Thus

$$\delta_{T,K} \leq d_\tau(k_0) P_{\tau,T}. \tag{25}$$

Lemma A.8 proves equation 19. Applying Lemma A.7 and minimizing separately over the finite set of observation points proves equation 21.

For the harmonic schedule and $k_0 \geq 1$, set $m_0 = (\tau - 1)K + k_0 \geq 1$. Since $a(b-1) \leq ab - 1$ for $a, b \geq 1$,

$$d_\tau(k_0) P_{\tau,T} \leq \frac{2\sigma_\tau}{nL} \left[\prod_{m=m_0+1}^{TK} \left(1 + \frac{cL}{m} \right) - 1 \right]. \tag{26}$$

Using $1 + x \leq e^x$ and an integral bound gives

$$\prod_{m=m_0+1}^{TK} \left(1 + \frac{cL}{m} \right) \leq \exp\left(cL \sum_{m=m_0+1}^{TK} \frac{1}{m} \right) \leq \exp\left(cL \int_{m_0}^{TK} \frac{dx}{x} \right) = \left(\frac{TK}{m_0} \right)^{cL}.$$

Combining the last two displays proves equation 22. □

A.5 Explicit bounds for a harmonic learning rate

Theorem A.12 (Final generalization bound of CL-SGD under globally decaying learning rate)**.** *Under Assumptions A.2–A.4, we consider standard continual SGD with the globally decaying learning rate*

$$\eta_{t,k} = \frac{c}{(t-1)K+k}, \qquad t \in [T],\ k \in [K],$$

for some constant $c > 0$. Then, for any task-dependent choice of observation points

$$k_{0,1}, \dots, k_{0,T} \in [K],$$

the expected generalization gap satisfies

$$\left|\mathcal{E}_{\mathrm{gen}}^{\mathrm{SGD}}\right| \le \frac{1}{T}\sum_{t=1}^{T} \bar{\varepsilon}_t^{\mathrm{SGD}}(k_{0,t}), \tag{27}$$

where

$$\bar{\varepsilon}_t^{\mathrm{SGD}}(k) := \frac{Uk}{n} + \frac{2L_G\sigma_t}{nL}\left[\left(\frac{TK}{(t-1)K+k}\right)^{cL} - 1\right], \qquad k \in [K]. \tag{28}$$

Optimizing these bounds over observation points gives

$$\left|\mathcal{E}_{\mathrm{gen}}^{\mathrm{SGD}}\right| \le \frac{1}{T}\sum_{t=1}^{T} \min_{k\in[K]}\left\{\frac{Uk}{n} + \frac{2L_G\sigma_t}{nL}\left[\left(\frac{TK}{(t-1)K+k}\right)^{cL} - 1\right]\right\}. \tag{29}$$

In particular, if one uses a common observation point $k_0 \in [K]$ for all tasks, then

$$\left|\mathcal{E}_{\mathrm{gen}}^{\mathrm{SGD}}\right| \le \frac{Uk_0}{n} + \frac{2L_G}{nLT}\sum_{t=1}^{T} \sigma_t\left[\left(\frac{TK}{(t-1)K+k_0}\right)^{cL} - 1\right]. \tag{30}$$

Proof. For each task $t \in [T]$, Theorem A.11 implies that for any neighboring continual learning sequences $\mathcal{S}^{(T)}$ and $\widetilde{\mathcal{S}}^{(T)}$ differing in exactly one sample from task t, and for any observation point $k_{0,t} \in [K]$, we have

$$\sup_{z\in\mathcal{Z}} \mathbb{E}_{\mathcal{A}}\left[\left|\ell\big(\mathcal{A}(\mathcal{S}^{(T)}); z\big) - \ell\big(\mathcal{A}(\widetilde{\mathcal{S}}^{(T)}); z\big)\right|\right] \le \bar{\varepsilon}_t^{\mathrm{SGD}}(k_{0,t}),$$

where $\bar{\varepsilon}_t^{\mathrm{SGD}}(k_{0,t})$ is defined in equation 28.

Thus, CL-SGD satisfies task-dependent symmetric stability with

$$\epsilon_t = \bar{\varepsilon}_t^{\mathrm{SGD}}(k_{0,t}).$$

Applying Lemma A.7 with these task-dependent stability coefficients gives

$$\left|\mathcal{E}_{\mathrm{gen}}^{\mathrm{SGD}}\right| \le \frac{1}{T}\sum_{t=1}^{T} \bar{\varepsilon}_t^{\mathrm{SGD}}(k_{0,t}),$$

which proves equation 27.

Since the above inequality holds for every choice of $k_{0,t} \in [K]$, we can optimize the observation point separately for each task. Taking the minimum over $k \in [K]$ for each task yields equation 29.

Finally, setting

$$k_{0,1} = \cdots = k_{0,T} = k_0$$

in equation 27 gives

$$\left|\mathcal{E}_{\mathrm{gen}}^{\mathrm{SGD}}\right| \le \frac{1}{T}\sum_{t=1}^{T}\left[\frac{Uk_0}{n} + \frac{2L_G\sigma_t}{nL}\left[\left(\frac{TK}{(t-1)K+k_0}\right)^{cL} - 1\right]\right],$$

which is exactly equation 30. This completes the proof. □

Corollary A.13 (Task-wise optimized observation points)**.** *Under the conditions of Theorem A.12, define*

$$p := cL, \qquad B_t := \frac{2L_G\sigma_t}{nL}.$$

For each task $t \in [T]$, introduce the shifted variable

$$q := (t-1)K + k, \qquad k \in \{1, \dots, K\},$$

so that

$$q \in [(t-1)K + 1,\ tK].$$

Consider the relaxed task-wise upper-bound objective

$$\phi_t(q) := \frac{U}{n}\big(q - (t-1)K\big) + B_t\left[\left(\frac{TK}{q}\right)^p - 1\right], \qquad q > 0.$$

If $B_t > 0$, then ϕ_t is strictly convex on $(0, \infty)$, and its unique stationary point is

$$q_t^\star = \left(\frac{2cL_G\sigma_t}{U}(TK)^{cL}\right)^{\frac{1}{1+cL}}. \tag{31}$$

If $B_t = 0$, define $q_t^\star = (t-1)K + 1$; then the loss-discrepancy term vanishes and the relaxed objective is minimized at the smallest admissible observation point.

To obtain a valid discrete observation point, define

$$\widehat{q}_t := \begin{cases} \min\{tK, \max\{(t-1)K + 1,\ q_t^\star\}\}, & B_t > 0, \\ (t-1)K + 1, & B_t = 0, \end{cases}$$

and let

$$\mathcal{N}_t := \Big(\{\lfloor \widehat{q}_t \rfloor, \lceil \widehat{q}_t \rceil\} \cap [(t-1)K + 1, tK] \cap \mathbb{Z}\Big).$$

Choose

$$q_t^{\mathrm{disc}} \in \arg\min_{q \in \mathcal{N}_t} \phi_t(q),$$

and set

$$\widehat{k}_{0,t} := q_t^{\mathrm{disc}} - (t-1)K. \tag{32}$$

Then $\widehat{k}_{0,t} \in [K]$, and the generalization gap of CL-SGD satisfies

$$\big|\mathcal{E}_{\mathrm{gen}}^{\mathrm{SGD}}\big| \le \frac{1}{T}\sum_{t=1}^{T}\left\{\frac{U\widehat{k}_{0,t}}{n} + \frac{2L_G\sigma_t}{nL}\left[\left(\frac{TK}{(t-1)K + \widehat{k}_{0,t}}\right)^{cL} - 1\right]\right\}. \tag{33}$$

Moreover, consider the continuous relaxation in which the observation point is allowed to be real-valued. If

$$B_t > 0, \qquad q_t^\star \in [(t-1)K + 1,\ tK], \qquad \forall t \in [T],$$

then the relaxed task-wise observation point is

$$k_{0,t}^\star = q_t^\star - (t-1)K.$$

In this interior case, define

$$\bar{\varepsilon}_t(k) := \frac{Uk}{n} + \frac{2L_G\sigma_t}{nL}\left[\left(\frac{TK}{(t-1)K + k}\right)^{cL} - 1\right].$$

The relaxed stationary-point value satisfies

$$\bar{\varepsilon}_t\left(q_t^\star - (t-1)K\right) \le \frac{1+cL}{cL}\frac{U}{n}\left(\frac{2cL_G\sigma_t}{U}\right)^{\frac{1}{1+cL}}(TK)^{\frac{cL}{1+cL}}. \tag{34}$$

Consequently, choosing the feasible floor/ceiling minimizers defined above yields the valid discrete bound

$$\big|\mathcal{E}_{\mathrm{gen}}^{\mathrm{SGD}}\big| \le \frac{U}{n} + \frac{1+cL}{cL}\frac{U}{nT}\sum_{t=1}^{T}\left(\frac{2cL_G\sigma_t}{U}\right)^{\frac{1}{1+cL}}(TK)^{\frac{cL}{1+cL}}. \tag{35}$$

Proof. The discrete bound equation 33 follows directly from Theorem A.12 by choosing

$$k_{0,t} = \widehat{k}_{0,t}, \qquad t = 1, \dots, T.$$

We now analyze the relaxed objective. By definition,

$$\phi_t(q) = \frac{U}{n}\big(q - (t-1)K\big) + B_t(TK)^p q^{-p} - B_t.$$

Its first derivative is

$$\phi_t'(q) = \frac{U}{n} - pB_t(TK)^p q^{-p-1},$$

and its second derivative is

$$\phi_t''(q) = p(p+1)B_t(TK)^p q^{-p-2}.$$

When $B_t > 0$, we have

$$\phi_t''(q) > 0, \qquad \forall q > 0.$$

Therefore, ϕ_t is strictly convex on $(0, \infty)$, and any stationary point is the unique global minimizer of the relaxed problem.

Setting $\phi_t'(q) = 0$ gives

$$\frac{U}{n} = pB_t(TK)^p q^{-p-1}.$$

Equivalently,

$$q^{p+1} = \frac{pnB_t}{U}(TK)^p.$$

Substituting

$$p = cL, \qquad B_t = \frac{2L_G\sigma_t}{nL},$$

we obtain

$$q^{1+cL} = \frac{2cL_G\sigma_t}{U}(TK)^{cL}.$$

Thus

$$q_t^\star = \left(\frac{2cL_G\sigma_t}{U}(TK)^{cL}\right)^{\frac{1}{1+cL}},$$

which proves equation 31. If $B_t = 0$, then ϕ_t is minimized at the smallest feasible q, equivalently $k = 1$.

If

$$B_t > 0, \qquad q_t^\star \in [(t-1)K + 1,\ tK],$$

then the relaxed minimizer lies inside the feasible interval, and hence

$$k_{0,t}^\star = q_t^\star - (t-1)K.$$

It remains to bound the stationary-point value. From the stationarity condition,

$$pB_t(TK)^p (q_t^\star)^{-p-1} = \frac{U}{n}.$$

Multiplying both sides by $q_t^\star/p$ gives

$$B_t\left(\frac{TK}{q_t^\star}\right)^p = \frac{U}{np}q_t^\star.$$

Therefore,

$$\begin{aligned}
\bar{\varepsilon}_t\left(q_t^\star - (t-1)K\right) &= \frac{U}{n}\big(q_t^\star - (t-1)K\big) + B_t\left[\left(\frac{TK}{q_t^\star}\right)^p - 1\right] \\
&= \frac{U}{n}\big(q_t^\star - (t-1)K\big) + \frac{U}{np}q_t^\star - B_t \\
&\le \frac{U}{n}q_t^\star + \frac{U}{np}q_t^\star \\
&= \frac{1+p}{p}\frac{U}{n}q_t^\star.
\end{aligned}$$

Substituting $p = cL$ and the expression for $q_t^\star$ gives

$$\bar{\varepsilon}_t\left(q_t^\star - (t-1)K\right) \le \frac{1+cL}{cL}\frac{U}{n}\left(\frac{2cL_G\sigma_t}{U}\right)^{\frac{1}{1+cL}}(TK)^{\frac{cL}{1+cL}},$$

which proves equation 34.

Finally, in the interior case, choose the feasible integer $q_t^{\text{ceil}} := \lceil q_t^\star \rceil$ and set $k_{0,t}^{\text{ceil}} := q_t^{\text{ceil}} - (t-1)K$. Then $q_t^{\text{ceil}} \le q_t^\star + 1$ and $q_t^{\text{ceil}} \ge q_t^\star$. Therefore,

$$\begin{aligned}
\phi_t(q_t^{\text{ceil}}) &= \frac{U}{n}\big(q_t^{\text{ceil}} - (t-1)K\big) + B_t\left[\left(\frac{TK}{q_t^{\text{ceil}}}\right)^p - 1\right] \\
&\le \frac{U}{n}\big(q_t^\star - (t-1)K\big) + \frac{U}{n} + B_t\left[\left(\frac{TK}{q_t^\star}\right)^p - 1\right] \\
&= \bar{\varepsilon}_t\left(q_t^\star - (t-1)K\right) + \frac{U}{n}.
\end{aligned}$$

Applying Theorem A.12 with $k_{0,t}^{\text{ceil}}$ and averaging over t yields equation 35. □

Algorithm 3 PAI with single-example SGD

Require: initial model w^0, task sequence $\{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_T\}$, local update steps K, learning rates $\{\eta_{t,k}\}$, anchoring parameter $\beta \in [0, 1]$.

Initialize model parameters $z^1 = x^1 = w^0$.
for $t = 1, 2, \ldots, T$ **do**
  receive current task dataset $S_t = (z_{t,1}, \ldots, z_{t,n})$ sampled from $\mathcal{D}_t$
  set $y_0^t \leftarrow (1-\beta)z^t + \beta x^t$
  for $k = 0, 1, \ldots, K-1$ **do**
    independently sample $j_{t,k} \sim \mathrm{Unif}([n])$ with replacement
    compute stochastic gradient $g_k^t = \nabla \ell(y_k^t; z_{t,j_{t,k}})$
    update model: $y_{k+1}^t = y_k^t - \eta_{t,k} g_k^t$
  end for
  set $z^{t+1} = y_K^t$
  set $x^{t+1} \leftarrow \dfrac{t-1}{t} x^t + \dfrac{1}{t} z^{t+1}$
end for
return y_K^T

B Generalization Analysis for PAI

B.1 Dual-anchor recursion and final output

We use the data model, coupling protocol, and shared assumptions from Section A. PAI starts from $z^1 = x^1 = w^0$ and uses

$$\begin{aligned} y_0^t &= (1-\beta)z^t + \beta x^t, \\ y_{k+1}^t &= y_k^t - \eta_{(t-1)K+k+1} \nabla \ell(y_k^t; z_{t,j_{t,k}}), \quad k = 0, \ldots, K-1, \\ z^{t+1} &= y_K^t, \\ x^{t+1} &= \frac{t-1}{t} x^t + \frac{1}{t} z^{t+1}. \end{aligned}$$

The local notation $\eta_{t,k}$ below denotes $\eta_{(t-1)K+k+1}$. The historical state satisfies $x^{t+1} = t^{-1} \sum_{s=1}^{t} y_K^s$ by induction. The analyzed output is $W = y_K^T = z^{T+1}$, and $\mathcal{E}_{\mathrm{gen}}^{\mathrm{PAI}}$ is defined using this output. The auxiliary average x^{T+1} is not used as the evaluation model. All comparisons use the same L, $\{\sigma_t\}$, L_G, U, and terminal region for both algorithms.

B.2 Stability of the final task endpoint

Neighboring sequences and coupling. Fix a task $\tau \in [T]$ and an index $\kappa \in [n]$. Let $\widetilde{\mathcal{S}}^{(T)}$ differ from $\mathcal{S}^{(T)}$ only by replacing $z_{\tau,\kappa}$ with another sample from $\mathrm{supp}(\mathcal{D}_\tau)$. Couple the two runs using the same initialization and the same indices $\{j_{t,k}\}$. Use tildes for the second run, and write

$$\widetilde{W} := \widetilde{y}_K^T = \widetilde{z}^{T+1}.$$

For $k_0 \in \{0, \ldots, K\}$, define

$$\mathcal{E}_{\tau,\kappa,k_0} := \{j_{\tau,0} \neq \kappa, \ldots, j_{\tau,k_0-1} \neq \kappa\},$$

with $\mathcal{E}_{\tau,\kappa,0} = \Omega$. Then

$$\mathbb{P}_{\mathcal{A}}(\mathcal{E}_{\tau,\kappa,k_0}^c) \leq \frac{k_0}{n}. \tag{36}$$

The exact probability is $1 - (1 - 1/n)^{k_0}$, including $k_0 = 0$. Since $n \geq 2$, the conditioning event has positive probability for every $k_0 \in \{0, \ldots, K\}$.

All empty scalar products below equal 1, and empty matrix products equal the identity.

Lemma B.1 (Decomposition for the final endpoint)**.** *Under Assumption A.4, for every* $k_0 \in \{0, \ldots, K\}$ *and every test sample* $z \in \mathcal{Z}$*,*

$$\mathbb{E}_{\mathcal{A}}\left[|\ell(W; z) - \ell(\widetilde{W}; z)|\right] \leq \frac{U k_0}{n} + L_G \, \mathbb{E}_{\mathcal{A}}\left[\|W - \widetilde{W}\| \mid \mathcal{E}_{\tau,\kappa,k_0}\right]. \tag{37}$$

Proof. Let $\mathcal{E} = \mathcal{E}_{\tau,\kappa,k_0}$ and $\Delta_z = |\ell(W; z) - \ell(\widetilde{W}; z)|$. By Lemma A.5, the two final losses obey both $\Delta_z \le U$ and $\Delta_z \le L_G\|W - \widetilde{W}\|$ almost surely. Hence

$$\begin{aligned}\mathbb{E}_{\mathcal{A}}[\Delta_z] &= \mathbb{E}_{\mathcal{A}}[\mathbf{1}_{\mathcal{E}^c}\Delta_z] + \mathbb{E}_{\mathcal{A}}[\mathbf{1}_{\mathcal{E}}\Delta_z] \\ &\le U\mathbb{P}_{\mathcal{A}}(\mathcal{E}^c) + L_G\mathbb{E}_{\mathcal{A}}[\mathbf{1}_{\mathcal{E}}\|W - \widetilde{W}\|] \\ &= U\mathbb{P}_{\mathcal{A}}(\mathcal{E}^c) + L_G\mathbb{P}_{\mathcal{A}}(\mathcal{E})\mathbb{E}_{\mathcal{A}}[\|W - \widetilde{W}\| \mid \mathcal{E}] \\ &\le \frac{Uk_0}{n} + L_G\mathbb{E}_{\mathcal{A}}[\|W - \widetilde{W}\| \mid \mathcal{E}].\end{aligned}$$

□

Lemma B.2 (Local expansion and perturbation creation)**.** *Under Assumptions A.2 and A.3, for any $\eta \ge 0$, the same-sample update satisfies*

$$\|u - \eta\nabla\ell(u; z) - (v - \eta\nabla\ell(v; z))\| \le (1 + \eta L)\|u - v\|. \tag{38}$$

Fix τ, κ, k_0 with $\mathbb{P}_{\mathcal{A}}(\mathcal{E}_{\tau,\kappa,k_0}) > 0$, and set

$$d_{\tau,k} := \mathbb{E}_{\mathcal{A}}\left[\|y_k^\tau - \widetilde{y}_k^\tau\| \mid \mathcal{E}_{\tau,\kappa,k_0}\right], \qquad k = k_0, \dots, K.$$

Then $d_{\tau,k_0} = 0$, and

$$d_{\tau,k+1} \le (1 + \eta_{\tau,k}L)d_{\tau,k} + \frac{2\eta_{\tau,k}\sigma_\tau}{n}. \tag{39}$$

Consequently,

$$d_{\tau,K} \le \frac{2\sigma_\tau}{nL}\left[\prod_{k=k_0}^{K-1}(1 + \eta_{\tau,k}L) - 1\right]. \tag{40}$$

Proof. By smoothness and the triangle inequality,

$$\begin{aligned}\|u - v - \eta(\nabla\ell(u; z) - \nabla\ell(v; z))\| &\le \|u - v\| + \eta\|\nabla\ell(u; z) - \nabla\ell(v; z)\| \\ &\le (1 + \eta L)\|u - v\|,\end{aligned}$$

which proves equation 38.

Before task τ, both runs are identical. On $\mathcal{E}_{\tau,\kappa,k_0}$, they also use identical samples during the first k_0 updates of task τ. Thus $y_{k_0}^\tau = \widetilde{y}_{k_0}^\tau$ and $d_{\tau,k_0} = 0$.

For a differing-sample update, add and subtract $\nabla\ell(v; z_{\tau,\kappa})$. Using equation 38 and equation 9, we obtain

$$\begin{aligned}&\|u - \eta\nabla\ell(u; z_{\tau,\kappa}) - (v - \eta\nabla\ell(v; \widetilde{z}_{\tau,\kappa}))\| \\ &\quad\le \|u - v - \eta(\nabla\ell(u; z_{\tau,\kappa}) - \nabla\ell(v; z_{\tau,\kappa}))\| \\ &\qquad + \eta\|\nabla\ell(v; z_{\tau,\kappa}) - \nabla\ell(v; \widetilde{z}_{\tau,\kappa})\| \\ &\quad\le (1 + \eta L)\|u - v\| + 2\eta\sigma_\tau.\end{aligned}$$

Therefore, pathwise,

$$\|y_{k+1}^\tau - \widetilde{y}_{k+1}^\tau\| \le (1 + \eta_{\tau,k}L)\|y_k^\tau - \widetilde{y}_k^\tau\| + 2\eta_{\tau,k}\sigma_\tau\mathbf{1}_{\{j_{\tau,k}=\kappa\}}.$$

For $k \ge k_0$, the current sampling index is independent of the preceding indices, including the event $\mathcal{E}_{\tau,\kappa,k_0}$, and hence

$$\mathbb{P}_{\mathcal{A}}(j_{\tau,k} = \kappa \mid \mathcal{E}_{\tau,\kappa,k_0}) = \frac{1}{n}.$$

Taking conditional expectations proves equation 39.

Finally,

$$d_{\tau,k+1} + \frac{2\sigma_\tau}{nL} \le (1 + \eta_{\tau,k}L)\left(d_{\tau,k} + \frac{2\sigma_\tau}{nL}\right).$$

Iterating from k_0 to $K - 1$ and using $d_{\tau,k_0} = 0$ proves equation 40. The argument also covers $k_0 = K$ through the empty-product convention. □

Theorem B.3 (Task-dependent stability of the PAI endpoint)**.** *Under Assumptions A.2–A.4, let* $\{\eta_m\}_{m=1}^{TK}$ *be a deterministic nonnegative schedule. Define*

$$A_t := \prod_{m=(t-1)K+1}^{tK} (1 + L\eta_m), \qquad t \in [T]. \tag{41}$$

For each perturbed task τ*, define*

$$\rho_{\tau+1}^{(\beta;\tau)} = \frac{1}{\tau}, \qquad \nu_{\tau+1}^{(\beta;\tau)} = 1,$$

and, for $t = \tau + 1, \dots, T$*, let*

$$\rho_{t+1}^{(\beta;\tau)} = \frac{t - 1 + \beta A_t}{t} \rho_t^{(\beta;\tau)} + \frac{(1-\beta)A_t}{t} \nu_t^{(\beta;\tau)}, \tag{42}$$

$$\nu_{t+1}^{(\beta;\tau)} = \beta A_t \rho_t^{(\beta;\tau)} + (1-\beta)A_t \nu_t^{(\beta;\tau)}. \tag{43}$$

For $k_0 \in \{0, \dots, K\}$*, set*

$$d_\tau(k_0) := \frac{2\sigma_\tau}{nL} \left[\prod_{m=(\tau-1)K+k_0+1}^{\tau K} (1 + L\eta_m) - 1 \right]. \tag{44}$$

Then any neighboring pair differing in one sample of task τ *satisfies*

$$\sup_{z \in \mathcal{Z}} \mathbb{E}_{\mathcal{A}} \left[|\ell(y_K^T; z) - \ell(\widetilde{y}_K^T; z)| \right] \le \varepsilon_\tau^{\mathrm{PAI},\beta}(k_0), \tag{45}$$

where

$$\varepsilon_\tau^{\mathrm{PAI},\beta}(k_0) := \frac{Uk_0}{n} + L_G d_\tau(k_0) \nu_{T+1}^{(\beta;\tau)}. \tag{46}$$

Equivalently, with $\mathbf{e}_2 = (0,1)^\top$*,*

$$\nu_{T+1}^{(\beta;\tau)} = \mathbf{e}_2^\top \left(M_T^{(\beta)} \cdots M_{\tau+1}^{(\beta)} \right) \begin{bmatrix} 1/\tau \\ 1 \end{bmatrix}, \tag{47}$$

where

$$M_t^{(\beta)} = \begin{bmatrix} \dfrac{t - 1 + \beta A_t}{t} & \dfrac{(1-\beta)A_t}{t} \\ \beta A_t & (1-\beta)A_t \end{bmatrix}.$$

In particular, $\nu_{T+1}^{(\beta;T)} = 1$*.*

Proof. Fix the neighboring pair and the observation point k_0. Write $\mathcal{E} = \mathcal{E}_{\tau,\kappa,k_0}$. Its probability is positive because $n \ge 2$.

Step 1: perturbation created in task τ**.** Conditional on the no-hit event, the states agree through its observation point. Subsequent indices remain independent and uniform given the past. Thus the pointwise sample-pair bound produces a fresh mismatch term with conditional mean $2\eta_{\tau,k}\sigma_\tau/n$, exactly as in the CL-SGD history argument. All pre-final parameter estimates use smoothness and gradient discrepancy only.

For $t \ge \tau + 1$, define

$$a_t := \mathbb{E}_{\mathcal{A}}[\|x^t - \widetilde{x}^t\| \mid \mathcal{E}], \qquad b_t := \mathbb{E}_{\mathcal{A}}[\|z^t - \widetilde{z}^t\| \mid \mathcal{E}].$$

Lemma B.2 and $z^{\tau+1} = y_K^\tau$ give

$$b_{\tau+1} \le d_\tau(k_0).$$

Since $x^\tau = \widetilde{x}^\tau$ and

$$x^{\tau+1} - \widetilde{x}^{\tau+1} = \frac{1}{\tau}(z^{\tau+1} - \widetilde{z}^{\tau+1}),$$

we also have

$$a_{\tau+1} = \frac{1}{\tau} b_{\tau+1} \le \frac{d_\tau(k_0)}{\tau}.$$

Step 2: propagation through subsequent tasks. For every $t > \tau$, both datasets and sampled indices agree. At initialization,

$$\mathbb{E}_{\mathcal{A}}[\|y_0^t - \widetilde{y}_0^t\| \mid \mathcal{E}] \le \beta a_t + (1-\beta) b_t.$$

Applying equation 38 through the K updates of task t yields

$$b_{t+1} \le A_t\big(\beta a_t + (1-\beta) b_t\big). \tag{48}$$

The averaging update then gives

$$\begin{aligned} a_{t+1} &\le \frac{t-1}{t} a_t + \frac{1}{t} b_{t+1} \\ &\le \frac{t-1+\beta A_t}{t} a_t + \frac{(1-\beta)A_t}{t} b_t. \end{aligned} \tag{49}$$

Step 3: comparison with the coefficient recursion. All entries of $M_t^{(\beta)}$ are nonnegative. The initial bounds and equation 48–equation 49 therefore imply, by induction, that

$$a_t \le d_\tau(k_0)\rho_t^{(\beta;\tau)}, \qquad b_t \le d_\tau(k_0)\nu_t^{(\beta;\tau)}, \qquad t = \tau+1, \dots, T+1.$$

For completeness, the induction step follows from

$$\begin{bmatrix} a_{t+1} \\ b_{t+1} \end{bmatrix} \le M_t^{(\beta)} \begin{bmatrix} a_t \\ b_t \end{bmatrix} \le d_\tau(k_0) M_t^{(\beta)} \begin{bmatrix} \rho_t^{(\beta;\tau)} \\ \nu_t^{(\beta;\tau)} \end{bmatrix},$$

where inequalities are componentwise.

Because the final output is $y_K^T = z^{T+1}$,

$$\mathbb{E}_{\mathcal{A}}[\|y_K^T - \widetilde{y}_K^T\| \mid \mathcal{E}] = b_{T+1} \le d_\tau(k_0)\nu_{T+1}^{(\beta;\tau)}. \tag{50}$$

Step 4: conversion to loss stability. Applying Lemma B.1 to equation 50 gives, uniformly in z,

$$\mathbb{E}_{\mathcal{A}}[|\ell(y_K^T; z) - \ell(\widetilde{y}_K^T; z)|] \le \frac{Uk_0}{n} + L_G d_\tau(k_0)\nu_{T+1}^{(\beta;\tau)}.$$

This proves the stability bound. The matrix expression follows by unrolling the coefficient recursion and selecting its second coordinate. □

Corollary B.4 (Task-averaged generalization)**.** *Under the conditions of Theorem B.3,*

$$\left|\mathcal{E}_{\mathrm{gen}}^{\mathrm{PAI}}\right| \le \frac{1}{T}\sum_{t=1}^{T} \min_{k\in\{0,\dots,K\}} \left\{ \frac{Uk}{n} + L_G d_t(k)\nu_{T+1}^{(\beta;t)} \right\}. \tag{51}$$

The right-hand side can also be capped by U.

Proof. Let $\mathcal{A}(\mathcal{S};\xi) = y_K^T$, where ξ is the random seed, independent of the data. For each (t,j), take an independent copy $z'_{t,j} \sim \mathcal{D}_t$, and let $\mathcal{S}^{(t,j)}$ replace $z_{t,j}$ by $z'_{t,j}$. Exchangeability of $z_{t,j}$ and $z'_{t,j}$ gives

$$\mathbb{E}[\ell(\mathcal{A}(\mathcal{S};\xi); z'_{t,j})] = \mathbb{E}[\ell(\mathcal{A}(\mathcal{S}^{(t,j)};\xi); z_{t,j})].$$

Consequently,

$$\mathcal{E}_{\mathrm{gen}}^{\mathrm{PAI}} = \frac{1}{Tn}\sum_{t=1}^{T}\sum_{j=1}^{n} \mathbb{E}\left[\ell(\mathcal{A}(\mathcal{S}^{(t,j)};\xi); z_{t,j}) - \ell(\mathcal{A}(\mathcal{S};\xi); z_{t,j})\right].$$

For each fixed neighboring pair, the stability theorem is uniform over the test sample. Thus it applies to $z_{t,j}$ after conditioning on the data and its replacement. For arbitrary deterministic choices $k_t \in \{0,\dots,K\}$, the triangle inequality gives

$$|\mathcal{E}_{\mathrm{gen}}^{\mathrm{PAI}}| \le \frac{1}{Tn}\sum_{t=1}^{T}\sum_{j=1}^{n} \varepsilon_t^{\mathrm{PAI},\beta}(k_t) = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_t^{\mathrm{PAI},\beta}(k_t).$$

The choices can be minimized separately over their finite sets, proving equation 51. Finally, terminal containment implies that both risks evaluated at the output lie in $[0,U]$, so the absolute expected gap is at most U. □

B.3 HARMONIC LEARNING RATES AND DISCRETE OPTIMIZATION

For $\eta_m = c/m$ and $k \in [K]$, the same integral estimate as in Theorem A.11, applied only through task t, gives

$$\varepsilon_t^{\mathrm{PAI},\beta}(k) \le \frac{Uk}{n} + \frac{2L_G\sigma_t}{nL}\left[\left(\frac{tK}{(t-1)K+k}\right)^{cL} - 1\right]\nu_{T+1}^{(\beta;t)}. \tag{52}$$

Minimizing over $k \in [K]$ and averaging over tasks gives the harmonic-schedule generalization bound in the main text. The following alternative bound includes $k = 0$ and admits an explicit discrete choice of observation point.

Corollary B.5 (Discrete optimization of an explicit upper bound)**.** *Assume the conditions of Theorem B.3 and the harmonic schedule $\eta_m = c/m$, with $c > 0$. Set*

$$p := cL, \qquad C_t^{(\beta)} := \frac{2L_G\sigma_t}{nL}\nu_{T+1}^{(\beta;t)}, \qquad a_t := (t-1)K + 1, \qquad b_t := tK + 1.$$

Here a_t, b_t are used only as interval endpoints in this corollary. For every integer $k \in \{0, \dots, K\}$,

$$\varepsilon_t^{\mathrm{PAI},\beta}(k) \le \bar{\varepsilon}_t^{\mathrm{PAI},\beta}(k) := \frac{Uk}{n} + C_t^{(\beta)}\left[e^p\left(\frac{b_t}{a_t+k}\right)^p - 1\right]. \tag{53}$$

Define, for $q \in [a_t, b_t]$,

$$\phi_t(q) := \frac{U}{n}(q - a_t) + C_t^{(\beta)}\left[e^p\left(\frac{b_t}{q}\right)^p - 1\right].$$

If $C_t^{(\beta)} > 0$, let

$$\begin{aligned} q_t^\star &:= \left(\frac{pnC_t^{(\beta)}e^p}{U}b_t^p\right)^{1/(1+p)} \\ &= \left(\frac{2cL_G\sigma_t}{U}e^{cL}\nu_{T+1}^{(\beta;t)}(tK+1)^{cL}\right)^{1/(1+cL)}. \end{aligned} \tag{54}$$

If $C_t^{(\beta)} = 0$, set $q_t^\star = a_t$. Define

$$\widehat{q}_t := \min\{b_t, \max\{a_t, q_t^\star\}\},$$

and select

$$q_t^{\mathrm{disc}} \in \arg\min_{q \in \{\lfloor \widehat{q}_t \rfloor, \lceil \widehat{q}_t \rceil\}} \phi_t(q).$$

Both candidates are feasible integers because a_t, b_t are integers and $\widehat{q}_t \in [a_t, b_t]$. Then

$$\widehat{k}_{0,t} := q_t^{\mathrm{disc}} - a_t \in \{0, \dots, K\},$$

and

$$|\mathcal{E}_{\mathrm{gen}}^{\mathrm{PAI}}| \le \frac{1}{T}\sum_{t=1}^{T}\phi_t(q_t^{\mathrm{disc}}). \tag{55}$$

If $C_t^{(\beta)} > 0$ and $q_t^\star \in [a_t, b_t] \cap \mathbb{Z}$ for every task, then

$$|\mathcal{E}_{\mathrm{gen}}^{\mathrm{PAI}}| \le \frac{1+cL}{cL}\frac{U}{nT}\sum_{t=1}^{T}\left(\frac{2cL_G\sigma_t}{U}e^{cL}\nu_{T+1}^{(\beta;t)}(tK+1)^{cL}\right)^{1/(1+cL)}. \tag{56}$$

Proof. Let $q = a_t + k$ and $N = tK = b_t - 1$. For $q \le N$,

$$\sum_{m=q}^{N}\frac{1}{m} \le \frac{1}{q} + \int_q^N \frac{dx}{x} \le 1 + \log\left(\frac{N+1}{q}\right).$$

For $q = N + 1$, the sum is empty and the same upper bound remains valid. Therefore,

$$\prod_{m=q}^{N} \left(1 + \frac{p}{m}\right) \leq \exp\left(p \sum_{m=q}^{N} \frac{1}{m}\right) \leq e^p \left(\frac{b_t}{q}\right)^p,$$

which proves equation 53.

For $C_t^{(\beta)} > 0$,

$$\phi_t'(q) = \frac{U}{n} - pC_t^{(\beta)} e^p b_t^p q^{-p-1},$$

and

$$\phi_t''(q) = p(p+1)C_t^{(\beta)} e^p b_t^p q^{-p-2} > 0.$$

Thus $q_t^\star$ is the unique unconstrained stationary point, and $\widehat{q}_t$ is the minimizer on $[a_t, b_t]$. The function decreases before its minimizer and increases after it. Hence its minimum over the feasible integers is attained at the floor or ceiling of $\widehat{q}_t$. If $C_t^{(\beta)} = 0$, ϕ_t is increasing and is minimized at a_t. Applying Corollary B.4 at $\widehat{k}_{0,t}$ proves equation 55.

For a feasible integer stationary point, the stationarity equation implies

$$C_t^{(\beta)} e^p \left(\frac{b_t}{q_t^\star}\right)^p = \frac{U}{np} q_t^\star.$$

Consequently,

$$\begin{aligned}\phi_t(q_t^\star) &= \frac{U}{n}(q_t^\star - a_t) + \frac{U}{np} q_t^\star - C_t^{(\beta)} \\ &\leq \frac{1+p}{p} \frac{U}{n} q_t^\star.\end{aligned}$$

Substitution and averaging give equation 56. Without the stated feasibility and integrality conditions, one must use the clipped and rounded expression equation 55. □

B.4 Comparison with sequential SGD

Lemma B.6 (PAI propagation is dominated by CL-SGD propagation)**.** *Under the conditions of Theorem B.3, fix a perturbed task $t \in [T]$ and define*

$$Q_{t+1}^{(t)} := 1, \qquad Q_{u+1}^{(t)} := A_u Q_u^{(t)}, \qquad u = t+1, \ldots, T.$$

Then, for every $u = t+1, \ldots, T+1$,

$$0 \leq \rho_u^{(\beta;t)} \leq Q_u^{(t)}, \qquad 0 \leq \nu_u^{(\beta;t)} \leq Q_u^{(t)}.$$

In particular, with

$$P_{t,T} := \prod_{u=t+1}^{T} A_u,$$

we have

$$\nu_{T+1}^{(\beta;t)} \leq P_{t,T}.$$

Thus, for every $k \in \{0, \ldots, K\}$, the derived PAI stability bound satisfies

$$\varepsilon_t^{\mathrm{PAI},\beta}(k) \leq \varepsilon_t^{\text{CL-SGD}}(k) := \frac{Uk}{n} + L_G d_t(k) P_{t,T}. \tag{57}$$

Proof. Since $\eta_{u,k} \geq 0$, we have $A_u \geq 1$. Initially,

$$\rho_{t+1}^{(\beta;t)} = \frac{1}{t} \leq 1 = Q_{t+1}^{(t)}, \qquad \nu_{t+1}^{(\beta;t)} = 1 = Q_{t+1}^{(t)}.$$

Assume both inequalities hold at u. The coefficient recursions give

$$\nu_{u+1}^{(\beta;t)} \leq A_u\big(\beta Q_u^{(t)} + (1-\beta) Q_u^{(t)}\big) = Q_{u+1}^{(t)}$$

and

$$\rho_{u+1}^{(\beta;t)} \le \frac{u-1+A_u}{u} Q_u^{(t)} \le A_u Q_u^{(t)} = Q_{u+1}^{(t)},$$

where the second inequality uses $A_u \ge 1$. Nonnegativity follows from the nonnegative initial values and coefficients. Induction proves the claim. Multiplication by $L_G d_t(k) \ge 0$ and addition of Uk/n give equation 57.

The CL-SGD expression is exactly the factorized certificate in Theorem A.11. Since the comparison holds for every k, taking task-wise minima and then averaging preserves the ordering of the derived generalization upper bounds. □

Corollary B.7 (CL-SGD as a special case of PAI)**.** *When $\beta = 0$, PAI with output y_K^T coincides with standard warm-start CL-SGD using the same initialization, learning rates, and sampled indices. Moreover,*

$$\nu_{T+1}^{(0;t)} = \prod_{u=t+1}^{T} A_u = P_{t,T},$$

and hence

$$\varepsilon_t^{\mathrm{PAI},0}(k) = \varepsilon_t^{\text{CL-SGD}}(k).$$

Proof. When $\beta = 0$, $y_0^t = z^t$, so every task starts at the preceding task's endpoint. The auxiliary average x^t does not affect any local update. The endpoint coefficient satisfies

$$\nu_{u+1}^{(0;t)} = A_u \nu_u^{(0;t)}, \qquad \nu_{t+1}^{(0;t)} = 1.$$

Unrolling gives the stated product and equality of bounds. □

Lemma B.8 (Monotonicity in β for the endpoint stability bound)**.** *Under the conditions of Theorem B.3, fix $t \in [T]$. Both $\rho_u^{(\beta;t)}$ and $\nu_u^{(\beta;t)}$ are non-increasing in $\beta \in [0,1]$ for every $u = t+1, \dots, T+1$. Consequently, for each fixed k,*

$$\varepsilon_t^{\mathrm{PAI},\beta}(k) = \frac{Uk}{n} + L_G d_t(k) \nu_{T+1}^{(\beta;t)}$$

is minimized at $\beta = 1$. The task-averaged bound after minimizing over k is also minimized at $\beta = 1$. These minimizers need not be unique.

Proof. Set $\alpha = 1 - \beta$, and, for this proof, write

$$r_u(\alpha) := \rho_u^{(1-\alpha;t)}, \qquad v_u(\alpha) := \nu_u^{(1-\alpha;t)}, \qquad \delta_u(\alpha) := v_u(\alpha) - r_u(\alpha).$$

The recursions become

$$r_{u+1}(\alpha) = \left(1 + \frac{A_u - 1}{u}\right) r_u(\alpha) + \frac{\alpha A_u}{u} \delta_u(\alpha), \tag{58}$$

$$v_{u+1}(\alpha) = A_u r_u(\alpha) + \alpha A_u \delta_u(\alpha), \tag{59}$$

$$\delta_{u+1}(\alpha) = \frac{u-1}{u} \left[(A_u - 1) r_u(\alpha) + \alpha A_u \delta_u(\alpha)\right]. \tag{60}$$

Initially,

$$r_{t+1}(\alpha) = \frac{1}{t}, \qquad \delta_{t+1}(\alpha) = 1 - \frac{1}{t} \ge 0.$$

Since $A_u \ge 1$, the recursions imply that both $r_u(\alpha)$ and $\delta_u(\alpha)$ remain nonnegative.

Fix $0 \le \alpha_1 \le \alpha_2 \le 1$. We prove by induction that

$$r_u(\alpha_1) \le r_u(\alpha_2), \qquad \delta_u(\alpha_1) \le \delta_u(\alpha_2).$$

The initial values are independent of α, so the base case holds. If the inequalities hold at u, then

$$\alpha_2 \delta_u(\alpha_2) - \alpha_1 \delta_u(\alpha_1) = \alpha_2 \big(\delta_u(\alpha_2) - \delta_u(\alpha_1)\big) + (\alpha_2 - \alpha_1)\delta_u(\alpha_1) \ge 0.$$

Subtracting equation 58 at α_1 from the same equation at α_2 gives a sum of nonnegative terms. The same is true for equation 60. This proves the induction.

Thus r_u and δ_u are non-decreasing in α. Since $v_u = r_u + \delta_u$, v_u is also non-decreasing in α. Both original coefficients are therefore non-increasing in β.

The quantities $d_t(k)$ are independent of β and nonnegative. Hence each stability bound is non-increasing in β. Taking a minimum over the finite set of k preserves this ordering, as does averaging over tasks. Therefore $\beta = 1$ is a minimizer of these bounds. □

Corollary B.9 (Optimal endpoint bound and comparison with CL-SGD)**.** *Define*

$$B_{t,T} := \nu_{T+1}^{(1;t)}.$$

Then

$$B_{t,T} = \begin{cases} \dfrac{A_T}{t} \displaystyle\prod_{u=t+1}^{T-1} \left(1 + \frac{A_u - 1}{u}\right), & 1 \le t < T, \\ 1, & t = T. \end{cases} \tag{61}$$

The $\beta = 1$ endpoint stability bound is

$$\varepsilon_t^{\mathrm{PAI},1}(k) = \frac{Uk}{n} + L_G d_t(k) B_{t,T}. \tag{62}$$

For every earlier task $t < T$,

$$B_{t,T} \le \frac{1}{t} P_{t,T}, \tag{63}$$

whereas

$$B_{T,T} = P_{T,T} = 1.$$

In particular,

$$\left|\mathcal{E}_{\mathrm{gen}}^{\mathrm{PAI}}\right|_{\beta=1} \le \frac{1}{T} \sum_{t=1}^{T} \min_{k \in \{0,\dots,K\}} \left\{ \frac{Uk}{n} + L_G d_t(k) B_{t,T} \right\}. \tag{64}$$

The right-hand side is no larger than the corresponding derived CL-SGD generalization upper bound.

Proof. For $\beta = 1$, the coefficient recursions reduce to

$$\rho_{u+1}^{(1;t)} = \left(1 + \frac{A_u - 1}{u}\right) \rho_u^{(1;t)}, \qquad \nu_{u+1}^{(1;t)} = A_u \rho_u^{(1;t)}.$$

If $t < T$, unrolling only up to the beginning of the final task gives

$$\rho_T^{(1;t)} = \frac{1}{t} \prod_{u=t+1}^{T-1} \left(1 + \frac{A_u - 1}{u}\right).$$

The final task then gives

$$\nu_{T+1}^{(1;t)} = A_T \rho_T^{(1;t)} = \frac{A_T}{t} \prod_{u=t+1}^{T-1} \left(1 + \frac{A_u - 1}{u}\right).$$

If $t = T$, there is no later propagation, and the initial endpoint coefficient is $\nu_{T+1}^{(1;T)} = 1$. This proves equation 61.

Since $A_u \ge 1$,

$$1 + \frac{A_u - 1}{u} \le A_u.$$

Therefore, for $t < T$,

$$B_{t,T} \le \frac{A_T}{t} \prod_{u=t+1}^{T-1} A_u = \frac{1}{t} P_{t,T}.$$

For $t = T$, both propagation factors equal 1. These inequalities give a pointwise comparison of the stability bounds for each k. Taking task-wise minima and averaging proves the comparison of the derived generalization upper bounds. □

B.5 Scope and boundary cases

These comparisons concern the derived stability upper bounds and the absolute expected generalization gap $|\mathcal{E}_{\mathrm{gen}}|$. They do not order the actual generalization gaps or population risks, and do not provide a bound on the expectation of an absolute random gap. Uniform stability over replacement tasks uses a maximum of the task-dependent coefficients, whereas expected generalization uses their average. All comparisons hold with fixed learning rates and shared assumption constants; dependence of L_G or U on the horizon must be retained in asymptotic interpretations.

The minimum at $\beta = 1$ concerns the stability certificate, not test accuracy: changes in empirical fitting can favor an intermediate anchoring weight. The analysis does not directly bound forgetting, spectral concentration, or plasticity. It covers the stated current-task single-example SGD protocol; replay, mini-batch updates, momentum, adaptive optimizers, and gradient clipping require separate analyses.

Boundary cases. When $\beta = 0$, PAI coincides with sequential SGD under matched randomness. For a perturbation in the final task, $\nu_{T+1}^{(\beta;T)} = 1$ for every β; the $1/t$ comparison in equation 63 applies only to $t < T$. Also, $x^2 = z^2 = y_K^1$, so $y_0^2 = z^2$. Consequently, PAI and sequential SGD have identical iterates through the second task under the matched protocol, and the first possible change in initialization occurs at task three.

C Additional Experimental Details

C.1 Benchmarks and evaluation protocols

The continual visual learning results are reported over five random seeds.

Limited-replay class-incremental learning. We use ResNet-18 on CIFAR-10 (5 tasks), CIFAR-100 (10 tasks), and Tiny-ImageNet (10 tasks), with disjoint classes and no task identities at inference. The replay budgets are 200 and 500 examples for CIFAR-10, 500 and 2,000 for CIFAR-100, and 500 and 4,000 for Tiny-ImageNet.

All-data replay. CIFAR-100 is divided into 20 phases with 5 new classes per phase. All previously observed training data remain available. The evaluated architectures are ResNet-18, ViT-Tiny, and VGG-16. Overall accuracy is the proportion of correctly classified test samples across all classes, including unseen classes; the test set is fixed throughout this sequence. This metric is distinct from current-task accuracy in the long-sequence experiment.

Long-sequence continual learning. We construct 5,000 binary classification tasks by sampling two of the 1,000 ImageNet classes per task, and evaluate CNN-4 and ViT-Tiny. Current-task test accuracy measures adaptation to the active task.

C.2 Additional forgetting results

Table 5: Final Forgetting (FF, %; ↓) of rehearsal-based continual learning methods with and without PAI. Results are reported over five random seeds. Lower values indicate less forgetting.

Method	CIFAR-10		CIFAR-100		Tiny-ImageNet	
	Buffer 200	Buffer 500	Buffer 500	Buffer 2000	Buffer 500	Buffer 4000
BFP Gu et al. (2023)	22.53 ± 5.00	16.81 ± 1.11	35.32 ± 3.94	19.76 ± 0.87	30.02 ± 4.37	27.19 ± 6.53
BFP + PAI	20.06 ± 2.56	14.54 ± 1.54	30.21 ± 1.65	16.26 ± 0.65	28.26 ± 3.57	24.52 ± 2.25
SARL Sarfraz et al. (2025)	16.36 ± 0.91	12.81 ± 0.48	24.15 ± 1.37	11.17 ± 1.21	42.90 ± 0.54	18.87 ± 1.08
SARL + PAI	15.34 ± 0.98	11.56 ± 0.43	23.45 ± 1.23	10.45 ± 1.09	40.45 ± 0.53	16.43 ± 0.86
IDER Liu et al. (2026)	15.28 ± 2.41	11.93 ± 0.49	29.98 ± 2.52	17.46 ± 1.04	36.63 ± 3.37	22.46 ± 1.86
IDER + PAI	14.45 ± 1.78	10.76 ± 0.56	27.34 ± 2.12	16.34 ± 1.00	32.45 ± 2.45	20.67 ± 1.45

Effect on Catastrophic Forgetting. Table 5 evaluates the effect of PAI on final forgetting when integrated with three rehearsal-based methods: BFP, SARL, and IDER. Across all 18 method–dataset–buffer configurations, adding PAI consistently reduces mean final forgetting, with absolute reductions ranging from 0.70 to 5.11 percentage points. For example, on CIFAR-100 with a buffer size of 500, PAI reduces the forgetting of BFP from 35.32% to 30.21%, a reduction of 5.11 percentage points. On Tiny-ImageNet with a buffer size of 500, it reduces forgetting from 42.90% to 40.45% for SARL and from 36.63% to 32.45% for IDER. These results suggest that historical anchoring at task boundaries can complement rehearsal-based training in mitigating catastrophic forgetting across the evaluated datasets and memory budgets.

C.3 Training configuration

Table 6: Reported training settings for continual visual learning.

Parameter	**Value**
Optimizer	AdamW / SGD
Learning Rate	10^{-3} (AdamW); $0.1, 0.03, 0.01$ (SGD)
Learning Rate Scheduler	Warmup
Gradient norm clipping	0.5
Batch Size	256
Epochs per Task	100
Data augmentation	None

Table 6 summarizes the training settings. Most hyperparameters follow Lee et al. (2024).

C.4 Hyperparameter search spaces

Table 7 lists the baseline search spaces. Tables 8, 9, and 10 report the selected values for the corresponding experimental settings.

Table 7: Baseline hyperparameter search spaces.

Experiment	**Method**	**Hyperparameters**	**Search Space**
Long-sequence continual learning	S&P	λ	$0.2, 0.4, 0.6, 0.8$
	DASH	α	$0.1, 0.3$
		λ	$0.05, 0.1, 0.3$
	L2 Reg.	λ	$10^{-3}, 10^{-4}, 10^{-5}$
	CBP	ρ	$10^{-4}, 10^{-5}$
		m	$100, 1000$
	FIRE	iter	10
Continual Learning	S&P	λ	$0.2, 0.4, 0.6, 0.8$
	DASH	α	$0.1, 0.3$
		λ	$0.05, 0.1, 0.3$
	L2 Reg.	λ	$10^{-3}, 10^{-4}, 10^{-5}$
	CBP	ρ	$10^{-4}, 10^{-5}$
		m	$100, 1000$
	FIRE	iter	10
Class-Incremental Learning	S&P	λ	$0.2, 0.4, 0.6, 0.8$
	DASH	α	$0.1, 0.3$
		λ	$0.05, 0.1, 0.3$
	L2 Reg.	λ	$10^{-3}, 10^{-4}, 10^{-5}$
	CBP	ρ	$10^{-4}, 10^{-5}$
		m	$100, 1000$
	FIRE	iter	10

C.5 SELECTED HYPERPARAMETERS: CONTINUAL SETTING

Table 8: Hyperparameters for Continual Setting.

Dataset	Method	Value
CIFAR-100 ResNet-18	S&P	$\lambda = 0.8$
	DASH	$\alpha = 0.3$, $\lambda = 0.05$
	L2 Reg.	$\lambda = 10^{-3}$
	CBP	$\rho = 10^{-5}, m = 1000$
	FIRE	iter $= 10$
	PAI	$\beta = 0.85$
CIFAR-100 ViT-Tiny	S&P	$\lambda = 0.8$
	DASH	$\alpha = 0.1$, $\lambda = 0.05$
	L2 Reg.	$\lambda = 10^{-5}$
	CBP	$\rho = 10^{-5}, m = 1000$
	FIRE	iter $= 10$
	PAI	$\beta = 0.85$
CIFAR-100 VGG-16	S&P	$\lambda = 0.8$
	DASH	$\alpha = 0.1$, $\lambda = 0.1$
	L2 Reg.	$\lambda = 10^{-4}$
	CBP	$\rho = 10^{-5}, m = 1000$
	FIRE	iter $= 10$
	PAI	$\beta = 0.85$

C.6 SELECTED HYPERPARAMETERS: CLASS-INCREMENTAL SETTING

Table 9: Hyperparameters for Class-Incremental Setting.

Dataset	Method	Value
CIFAR-100 (ResNet-18)	S&P	$\lambda = 0.8$
	DASH	$\alpha = 0.3$, $\lambda = 0.3$
	L2 Reg.	$\lambda = 10^{-5}$
	CBP	$\rho = 10^{-5}, m = 1000$
	FIRE	iter $= 10$
CIFAR-100 (ViT-Tiny)	S&P	$\lambda = 0.8$
	DASH	$\alpha = 0.3$, $\lambda = 0.3$
	L2 Reg.	$\lambda = 10^{-5}$
	CBP	$\rho = 10^{-5}, m = 1000$
	FIRE	iter $= 10$
CIFAR-100 (VGG-16)	S&P	$\lambda = 0.8$
	DASH	$\alpha = 0.3$, $\lambda = 0.1$
	L2 Reg.	$\lambda = 10^{-3}$
	CBP	$\rho = 10^{-4}, m = 1000$
	FIRE	iter $= 10$

C.7 SELECTED HYPERPARAMETERS: LONG-SEQUENCE LEARNING

Table 10: Hyperparameters for Long-sequence continual learning.

Dataset	**Method**	**Value**
ImageNet (CNN-4)	S&P	$\lambda = 0.8$
	DASH	$\alpha = 0.3$, $\lambda = 0.3$
	L2 Reg.	$\lambda = 10^{-5}$
	CBP	$\rho = 10^{-5}, m = 1000$
	FIRE	iter $= 10$
ImageNet (ViT-Tiny)	S&P	$\lambda = 0.8$
	DASH	$\alpha = 0.3$, $\lambda = 0.3$
	L2 Reg.	$\lambda = 10^{-5}$
	CBP	$\rho = 10^{-5}, m = 1000$
	FIRE	iter $= 10$