PSCAD: PARAMETER SPACE CONTINUOUS ATTRACTOR DYNAMICS IN CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual learning models acquire new knowledge while preserving previously learned information. Dynamic architecture methods allocate task-specific parameters to protect previously learned knowledge from direct overwriting, but leave the growing collection of task updates without explicit coordination. Inspired by the separation of human brain between knowledge acquisition during wakefulness and memory consolidation during sleep, we introduce Parameter-Space Continuous Attractor Dynamics (PSCAD), a parameter-consolidation mechanism for rehearsal-free continual learning that complements data-driven training process. After task-specific learning, the mechanism treats accumulated task parameters as dynamical states and refines them through CANN-inspired dynamics. The dynamics combine affinity-weighted interactions among task parameters, adaptation-based regulation, and associative retrieval from a learnable Hopfield memory. We establish sufficient conditions for Lyapunov energy dissipation and convergence to the equilibrium set of the continuous-time system with fixed memory. Experiments on ImageNet-R and ImageNet-A show consistent improvements in final average accuracy across all evaluated settings, together with gains in average anytime accuracy in most settings. Ablations further support the contribution of each consolidation component. The findings demonstrate the effectiveness of parameter-space consolidation through CANN-inspired dynamics in continual learning.

1 INTRODUCTION

Continual learning requires models to acquire new knowledge from a sequence of tasks while retaining information learned from previous tasks (Wang et al., 2024). A central challenge is to prevent catastrophic forgetting, where sequential updates to shared parameters overwrite previous learned knowledge (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017; De Lange et al., 2022). Existing continual learning methods have therefore explored complementary strategies including regularization-based (Kirkpatrick et al., 2017; Zenke et al., 2017), replay-based (Rebuffi et al., 2017; Rolnick et al., 2019; Shin et al., 2017), and architecture-based approaches (Fernando et al., 2017; Mallya & Lazebnik, 2018; Serrà et al., 2018) to reduce catastrophic forgetting. In particular, architecture-based approaches allocate task-specific parameters or gate network capacity and restrict updates to parameters used by previous tasks, thereby protecting acquired knowledge. However, parameter isolation alone does not explicitly coordinate the growing collection of task-specific updates. As tasks accumulate, these parameters are learned under different data distributions and can remain largely independent after task-specific training, motivating us to investigate how task-specific parameters can be jointly refined.

This question is inspired by the complementary roles of rapid memory acquisition and subsequent consolidation in biological learning systems (McClelland et al., 1995). The knowledge acquisition during wakefulness and memory consolidation during sleep are separate processes in human brain. A substantial body of neuroscience research suggests that sleep actively supports the stabilization of newly acquired memories (Klinzing et al., 2019; Brodt et al., 2023). During sleep, previously encoded memory traces can be reactivated and reorganized, contributing to their long-term stabilization and integration with existing knowledge (Klinzing et al., 2019; Dumay & Gaskell, 2007). In spiking neural networks, sleep-like activity has also been shown to reorganize synaptic weights into a joint representation of previously learned tasks (**?**). Motivated by this mechanism, we distinguish

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.

knowledge acquisition from **parameter consolidation** in continual learning. During knowledge acquisition, the model learns task-specific parameters from the current task data. During consolidation, these learned parameters themselves become the objects of computation, allowing accumulated knowledge to be reorganized directly in parameter space without revisiting historical examples. This shifts part of continual learning process from data space to parameter space, suggesting that task-specific parameters need not remain isolated after they have been acquired. Instead, they can interact and evolve collectively to form a more organized representation.

Continuous attractor dynamics provide a natural framework for realizing such parameter-space consolidation. Continuous attractor neural networks (CANNs) support a continuum of stable states through recurrent interactions, allowing attractor states to form continuous structure (Amari, 1977; Fung et al., 2016; Khona & Fiete, 2022). Applied to task parameters, such dynamics provide a mechanism through which accumulated parameter states can interact and become collectively organized in parameter space. Besides, associative memory provides a complementary capability by storing and retrieving persistent representations. Classical attractor-based associative memory has long provided a framework for storing patterns as stable states (Hopfield, 1982), while Modern Hopfield networks extend this principle to high-dimensional continuous representations with powerful associative retrieval capabilities (Ramsauer et al., 2021). Combining CANN-inspired dynamics with associative memory therefore provides a natural basis for parameter consolidation, where task parameters are coordinated according to their mutual affinities while being guided by previously learned parameter patterns. An adaptation mechanism can further regulate the temporal evolution of these states and contribute to the dynamics (Dong et al., 2022).

Motivated by this perspective, we introduce **Parameter-Space Continuous Attractor Dynamics (PSCAD)**, a parameter-consolidation mechanism for rehearsal-free continual learning. PSCAD introduces a parameter-driven consolidation process after task-specific learning. The accumulated task parameters are refined through CANN-inspired dynamics which combine potential-weighted interactions, adaptation-based regulation, and Hopfield memory retrieval. The resulting parameter states are incorporated through a residual pathway, preserving the task-specific parameters acquired during knowledge acquisition while introducing the refinement induced by consolidation. Importantly, the consolidation process operates directly on parameter states and does not require replaying or generating historical examples. In this way, PSCAD turns the growing collection of task-specific parameters from a set of isolated updates into an active, memory-guided dynamical system.

Beyond the model design, we analyze the stability of the resulting continuous-time dynamics from an energy-based perspective. Energy functions and Lyapunov methods provide a principled framework for analyzing the stability and convergence of dynamical systems (Hopfield, 1982). We construct a Lyapunov function and establish sufficient conditions under which its energy is non-increasing along system trajectories. Under a strict condition on the system time constants, we further establish convergence to the equilibrium set of the continuous-time system with fixed memory. These results characterize the stability properties of the proposed parameter-consolidation dynamics and provide a theoretical basis for using CANN-inspired dynamics in parameter space. We evaluate PSCAD on ImageNet-R and ImageNet-A and observe consistent improvements in final average accuracy across all evaluated settings, together with gains in average anytime accuracy in most settings. Ablation studies further examine the contributions of each component to the consolidation process. The results demonstrate the effectiveness of parameter-space consolidation through CANN-inspired dynamics for rehearsal-free continual learning.

Our contributions are summarized as follows:

- We formulate **parameter-space consolidation** as a complementary process to knowledge acquisition in rehearsal-free continual learning, inspired by the separation between learning and memory consolidation in biological systems.
- We propose **Parameter-Space Continuous Attractor Dynamics (PSCAD)** as a rehearsal-free method for continual learning, which consolidates task parameters through CANN-inspired dynamics with potential-based interaction, adaptation and Hopfield memory retrieval.
- We establish sufficient conditions for Lyapunov energy dissipation and convergence to the equilibrium set of the continuous dynamics in parameter space and demonstrate the effec-

tiveness of PSCAD on ImageNet-R and ImageNet-A through quantitative comparisons and ablation studies.

2 Related Work

Continual Learning Continual learning aims to acquire knowledge from a sequence of tasks while retaining previously learned information (Wang et al., 2024). Representative approaches include regularization-based, replay-based, and architecture-based approaches. Among them, architecture-based approaches mitigate direct parameter interference by allocating task-specific parameters, modules, or computational paths (Fernando et al., 2017; Mallya & Lazebnik, 2018; Serrà et al., 2018). In particular, Scalable Decoupled LoRA (SD-LoRA) (Wu et al., 2025) allocates a task-specific LoRA parameter subset to each task and combine them through learned weighting coefficients. During continual learning, however, previously learned task parameters remain fixed without interaction with other tasks. Our method therefore introduces a parameter-consolidation operator that refines the accumulated task parameters without replaying data from previous tasks.

Parameter-Space Consolidation Our parameter-consolidation mechanism is related to model merging, which integrates task-specific models or parameter updates directly in weight space. Fukuda et al. (2025) consolidate task-specific adapters into a single adapter through weight averaging, together with centroid prototype mapping. Yang et al. (2025) adapt interpolation coefficients by projecting parameter differences onto task-relevant subspaces. In parameter-efficient continual learning, Qiao & Mahdavi (2025) sequentially merge task-specific LoRA updates into a single adapter. Recent work further formulates continual model merging as an ODE that moves toward a merged parameter point along constrained low-loss paths (Lin & Kang, 2026). In contrast, our method does not collapse task updates into a single merged parameterization; instead, it retains task-specific parameters as coupled states and evolves them through CANN-inspired dynamics, thereby modeling cross-task interactions.

Attractor Neural Networks Attractor neural networks represent information through stable states toward which nearby trajectories converge. Continuous attractor neural networks (CANNs), rooted in Amari's neural-field dynamics (Amari, 1977), further support continuous families of stable states (Amari, 1977; Khona & Fiete, 2022). In continual learning, Attention Attractor Networks regularize newly learned classifier weights toward attractors constructed from base-class parameters (Ren et al., 2019), while Predictive Attractor Models preserve sequence memories through context-dependent neural activation and lateral inhibition (Mounir & Sarkar, 2024). Rather than only constraining new parameters relative to learned attractors or defining attractor evolution through neuron-level activation and inhibition, our method treats task-specific LoRA parameters as dynamical states that evolve in CANN-inspired dynamic.

Hopfield Networks Hopfield networks retrieve stored patterns by evolving an initial state toward an energy minimum (Hopfield, 1982). Modern Hopfield networks extend this principle to continuous high-dimensional representations through a Log-Sum-Exp energy, with retrieval updates equivalent to attention (Ramsauer et al., 2021). For continual learning, Sparse Quantized Hopfield Networks address online-continual associative memory by using Hopfield network to encode and retrieve input samples (Alonso & Krichmar, 2023). More recently, Hopfield-inspired retrieval has been used to generate input-conditioned routing prompts (Masano et al., 2026). Our Hopfield component instead operates on task-specific LoRA parameters rather than original input samples. It retrieves parameter-space memory that acts as external input to the consolidation dynamics.

3 Methodology

We consider the class-incremental learning setting (Liang & Li, 2024; Wu et al., 2025) for continual learning, where a sequence of T classification tasks is presented sequentially. At task t, the model is trained only on the current dataset $\mathcal{D}_t = \{(\mathbf{x}, y)\}$, where $\mathbf{x}$ and y denote an input image and its corresponding label, respectively. While training on task t for continual learning, following the architecture-based paradigm, we allocate a dedicated parameter subset θ_t to each task and introduce a parameter-consolidation operator g_{Φ_t} inspired by human sleep memory consolidation. The

operation is parameterized by a memory matrix $\Phi_t = [\phi_1, \phi_2, \ldots, \phi_M] \in \mathbb{R}^{d\times M}$, where M is the memory capacity and each column $\phi_k \in \mathbb{R}^d$ represents a learnable memory pattern. Specifically, g_{Φ_t} is instantiated as a continuous-time attractor dynamical system that evolves the accumulated task-specific parameters along an internal consolidation time s, *i.e.*,

$$\tau_s \frac{\partial \theta_i}{\partial s} = -\theta_i - V_i + \Omega_i(\theta_{1:t}) + \mathcal{M}(\theta_i, \Phi_{t-1}),\ i = 1, 2, \cdots, t, \tag{1}$$

$$\tau_z \frac{\partial V_i}{\partial s} = -V_i + m\theta_i. \tag{2}$$

The four terms on the right-hand side of Equation (1) represent state decay, adaptation-based regulation, potential-based interaction among task parameters and Hopfield-guided memory retrieval external input, where $\tau_s > 0$ is the time constant of the consolidation dynamics. The adaptation-based regulation (Equation (2)) captures responses attenuation under sustained excitation, counterbalancing recurrent excitation and preventing overactive parameter dynamics, where $\tau_z > 0$ is the adaptation time constant and $m > 0$ controls the adaptation strength. These mechanisms enable the parameter-consolidation operator to coordinate task parameters according to parameter-space proximity, suppress excessive activation, and incorporate memory information.

Potential-based Interaction $\Omega_i(\theta_{1:t})$. We define the potential induced by the affinity-weighted interactions among task-specific parameter states by

$$\Psi(\theta_{1:t}) = -\rho J_0 \sum_{1\le i<j\le t} w_{ij} \exp\left(-\frac{\|\theta_i - \theta_j\|_2^2}{2a_t^2}\right), \tag{3}$$

where $\rho > 0$ controls the overall interaction scale, $J_0 > 0$ denotes the coupling strength, and $a_t > 0$ is the interaction bandwidth. The coefficient $w_{ij} = w_{ji} \in [0, 1]$ represents a fixed task affinity during one trajectory. The potential-based interaction $\Omega_i(\theta_{1:t})$ is defined as negative gradient of the potential, *i.e.*,

$$\Omega_i(\theta_{1:t}) = -\frac{\partial \Psi(\theta_{1:t})}{\partial \theta_i}, \tag{4}$$

which is jointly determined by the task affinity w_{ij} and the Gaussian proximity kernel, exerting stronger attraction between semantically related task parameters.

Hopfield-Guided Memory Retrieval $\mathcal{M}(\theta_i, \Phi_{t-1})$. During consolidation for task t, we retrieve previously acquired knowledge from a compact memory matrix Φ_{t-1}, which is accumulated over tasks $1, \ldots, t-1$ and provides memory-guided signal. We define memory-dependent Hopfield energy term for task parameter θ_i as

$$E_H(\theta_i, \Phi_{t-1}) = -\frac{1}{\beta} \mathrm{LSE}(\beta \Phi_{t-1}^\top \theta_i), \tag{5}$$

where LSE denotes Log-Sum-Exp function and $\beta > 0$ controls the sharpness of memory retrieval. The corresponding memory retrieval external input is defined as

$$\mathcal{M}(\theta_i, \Phi_{t-1}) = \Phi_{t-1} \,\mathrm{Softmax}\left(\beta \Phi_{t-1}^\top \theta_i\right). \tag{6}$$

This term is the negative gradient of the memory-dependent energy term with respect to θ_i and supplies the dynamics with a retrieved representation of previously acquired knowledge.

Energy Dissipation and Stability. To characterize the stability of the proposed consolidation dynamics, we analyze the CANN-inspired dynamics through a Lyapunov energy. We consider the consolidation dynamics with an arbitrary fixed memory matrix Φ. Our analysis focuses on whether these coupled dynamics admit a scalar energy function that decreases along system trajectories and whether such dissipation guarantees convergence to stable states. To this end, we define the Lyapunov energy by

$$\mathcal{E}(\theta_{1:t}, V_{1:t}) = \Psi(\theta_{1:t}) + \sum_{i=1}^{t} \left[E_H(\theta_i, \Phi) + \frac{1+m}{2}\|\theta_i\|_2^2 + \frac{1}{2m}\|V_i - m\theta_i\|_2^2 \right]. \tag{7}$$

Besides the joint potential $\Psi(\theta_{1:t})$ and Hopfield associative energy $E_H(\theta_i, \Phi)$, the quadratic terms account for the intrinsic parameter energy and the coupling between the parameter and adaptation states.

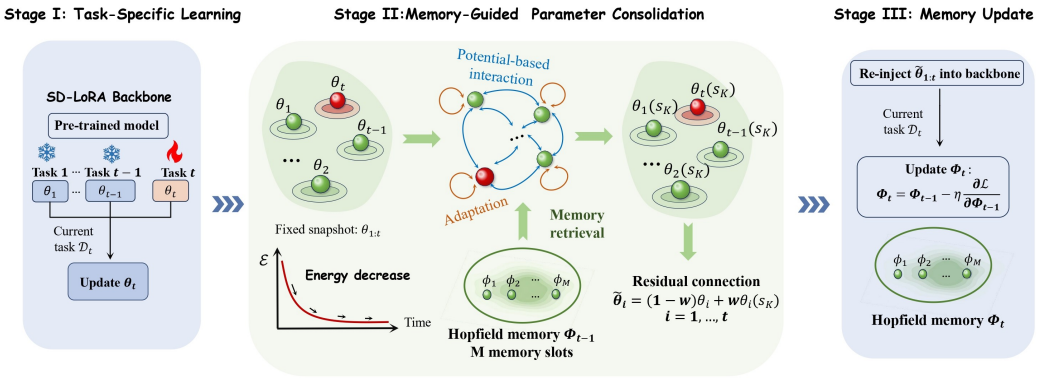

Figure 1: Overview of PSCAD framework: task-specific learning followed by parameter consolidation through memory-guided CANN-inspired dynamics.

Theorem 1. *Let $m > 0$, $\tau_z > 0$, and $\tau_s \geq m\tau_z$. Under a fixed memory matrix Φ, along any trajectory of the coupled system defined by Equation (1) and Equation (2), the Lyapunov energy satisfies*

$$\dot{\mathcal{E}} = -(\tau_s - m\tau_z) \sum_{i=1}^{t} \|\dot{\theta}_i\|_2^2 - \frac{\tau_z}{m} \sum_{i=1}^{t} \|\dot{V}_i - m\dot{\theta}_i\|_2^2 \leq 0. \tag{8}$$

Therefore, the Lyapunov energy $\mathcal{E}$ is non-increasing along every system trajectory. If $\tau_s > m\tau_z$, then $\dot{\mathcal{E}} = 0$ implies $\dot{\theta}_i = 0$ and $\dot{V}_i = 0$ for $i = 1, 2, \cdots, t$; moreover, every trajectory remains bounded if the Lyapunov energy is coercive.Consequently, by LaSalle's invariance principle, the joint state $x(s) = (\theta_{1:t}(s), V_{1:t}(s))$ approaches the equilibrium set $\mathcal{Z}$ of the coupled dynamics, namely, $\lim_{s\to\infty} \operatorname{dist}\big(x(s), \mathcal{Z}\big) = 0$.

The complete proof is provided in Appendix A.

4 Training Procedure

Inspired by knowledge acquisition during wakefulness and memory consolidation during sleep, our training framework proceeds in three stages. First, the model learns task-specific parameters from incoming data. It then performs memory-guided consolidation, jointly refining the accumulated task parameters and updating long-term memory. The procedure is shown in Figure 1.

Stage I: Task-Specific Learning At the beginning of task t, we allocate a new parameter subset θ_t to the current task. We denote the classification model after learning task t by $f(\theta_{0:t}; \mathbf{x})$, where the pretrained classifier weight matrix θ_0 and the previously consolidated parameters $\theta_{1:t-1}$ remain fixed. θ_t is optimized using the current dataset $\mathcal{D}_t$, *i.e.*,

$$\theta_t = \arg\min_{\theta} \mathcal{L}(\theta_{0:t-1}, \theta; \mathcal{D}_t) = \arg\min_{\theta} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_t} \ell\left(f(\theta_{0:t-1}, \theta; \mathbf{x}), y\right), \tag{9}$$

where $\ell(\cdot,\cdot)$ denotes a per-sample loss function, such as cross-entropy. This stage provides a data-driven learning process analogous to knowledge acquisition during wakefulness. In this paper, we adopt SD-LoRA (Wu et al., 2025) as the data-driven backbone for learning θ_t.

Stage II: Memory-Guided Parameter Consolidation Given the task-specific parameter θ_i, We initialize the parameter states as $\theta_i(s_0) = \theta_i, V_i(s_0) = \mathbf{0}$ and implement the CANN dynamics using Euler discretization:

$$\theta_i(s_{k+1}) = \theta_i(s_k) + \frac{\Delta s}{\tau_s}\Big[-\theta_i(s_k) - V_i(s_k) + \Omega_i(\theta_{1:t}(s_k)) + \mathcal{M}(\theta_i(s_k), \Phi_{t-1})\Big], \tag{10}$$

$$V_i(s_{k+1}) = V_i(s_k) + \frac{\Delta s}{\tau_z}\Big[-V_i(s_k) + m\theta_i(s_{k+1})\Big]. \tag{11}$$

Algorithm 1 Parameter-Space Continuous Attractor Dynamics

Input: Pretrained parameters θ_0, initial memory Φ_0, task datasets $\{\mathcal{D}_t\}_{t=1}^{T}$
Output: Consolidated parameters $\{\tilde{\theta}_{1:T}\}$ and memory Φ_T
for $t = 1, \cdots, T$ **do**
Stage I: Task-Specific Learning
  Optimize θ_t on $\mathcal{D}_t$ with $\theta_0, \theta_{1:t-1}$ fixed, *i.e.*, $\theta_t = \arg\min_\theta \mathcal{L}(\theta_{0:t-1}, \theta; \mathcal{D}_t)$
  Stage II: Parameter Consolidation
  Initialize $\theta_i(s_0) \leftarrow \theta_i$ and $V_i(s_0) \leftarrow \mathbf{0}$ for $i = 1, \cdots, t$
  for $n = 0, \cdots, N-1$ **do**
    Update $\theta_i(s_{n+1})$ using the CANN dynamics (Equation (1))
    Update $V_i(s_{n+1})$ using the adaptation dynamics (Equation (2))
  $\tilde{\theta}_i \leftarrow (1-w)\theta_i + w\theta_i(s_K)$ for $i = 1, \cdots, t$
  Stage III: Memory Update
  $\Phi_t \leftarrow \Phi_{t-1} - \eta \nabla_{\Phi_{t-1}} \mathcal{L}(\tilde{\theta}_{1:t}; \mathcal{D}_t)$

For $k = 0, \cdots, K-1$, $s_k = k\Delta s$ denotes the k-th discretization point of the internal consolidation time. After K iterations, the evolved parameter $\theta_i(s_K)$ are obtained through a residual connection with original task parameter θ_i, *i.e.*,

$$\tilde{\theta}_i = (1 - w_i)\theta_i + w_i\theta_i(s_K), \tag{12}$$

where w_i controls the contribution of the parameter consolidation. The residual pathway preserves the task-wise knowledge acquired from data, while incorporating the parameter-space refinement.

Stage III: Memory Update After parameter consolidation, the resulting task loss $\mathcal{L}(\tilde{\theta}_{0:t}; \mathcal{D}_t)$ provides a learning signal for updating the long-term memory. We therefore update the memory matrix by

$$\Phi_t = \Phi_{t-1} - \eta \frac{\partial \mathcal{L}(\tilde{\theta}_{0:t}; \mathcal{D}_t)}{\partial \Phi_{t-1}}, \text{ where } \frac{\partial \mathcal{L}(\tilde{\theta}_{0:t}; \mathcal{D}_t)}{\partial \Phi_{t-1}} = \sum_{i=1}^{t} \frac{\partial \mathcal{L}(\tilde{\theta}_{0:t}; \mathcal{D}_t)}{\partial \tilde{\theta}_i} \cdot \frac{\partial \tilde{\theta}_i}{\partial \Phi_{t-1}}, \tag{13}$$

where η denotes the memory learning rate. In this sense, the consolidated parameters $\tilde{\theta}_{0:t}$ obtained from this trajectory are retained for subsequent tasks, while the updated memory Φ_t is used to guide consolidation of future task parameters.

5 Experiment

5.1 Experimental Setup

Datasets and protocols We conduct class-incremental learning experiments on ImageNet-R (Hendrycks et al., 2021a; Boschini et al., 2022) and ImageNet-A (Hendrycks et al., 2021b). ImageNet-R contains 200 classes represented by artistic renditions and is evaluated under 5, 10, and 20 sequential tasks (denoted by N), corresponding to 40, 20, and 10 classes per task. ImageNet-A contains 200 classes of naturally occurring adversarial examples and is evaluated under 10 sequential tasks with 20 classes per task. At task t, the learner has access only to the training data of the current task and the parameters retained from previous tasks. No historical training examples are replayed or generated during either task-specific learning or parameter consolidation.

Baselines We compare PSCAD with representative continual-learning methods spanning prompt-based and parameter-efficient adaptation paradigms. Full fine-tuning provides a reference under unconstrained sequential parameter updates. L2P (Wang et al., 2022b), DualPrompt (Wang et al., 2022a), and CODA-Prompt (Smith et al., 2023) represent prompt-based continual-learning approaches. InfLoRA (Liang & Li, 2024) and SD-LoRA (Wu et al., 2025) represent low-rank adaptation approaches, with SD-LoRA serving as the direct backbone of PSCAD. We further compare with recent parameter-efficient continual fine-tuning methods, including PLAN (Wang et al., 2025), HippoTune (Chen et al., 2026), PROL (Ma'sum et al., 2025), and ACMap (Fukuda et al., 2025). Baseline results are taken from the available previous publications. Experiment dataset varies across methods, as some methods report results on only one benchmark.

Implementation details We use ViT-B/16 (Dosovitskiy et al., 2021) as the foundation model, initialized from a checkpoint pretrained on ImageNet-21K and fine-tuned on ImageNet-1K. We use SGD (Bottou, 2010) to optimize the task-specific parameters in Stage I and AdamW (Loshchilov & Hutter, 2017) to optimize the memory matrix through the consolidation dynamics in Stages II–III. We train for 30 epochs on ImageNet-R and 20 epochs on ImageNet-A. The training epochs are split equally between task-specific parameter learning (Stage I) and memory learning through parameter consolidation (Stages II–III). We report final average accuracy (Acc) and average anytime accuracy (AAA). Acc measures the final performance after learning the complete task sequence, whereas AAA captures performance throughout the continual-learning process and therefore reflects both final retention and the quality of intermediate states. Our results are averaged over three independent runs with different random seeds.

5.2 Comparison with Other Continual Learning Methods

We first evaluate whether parameter-space consolidation improves continual learning over existing approaches. Tables 1 and 2 summarize the results on ImageNet-R and ImageNet-A, respectively. We make the following observations.

On ImageNet-R, PSCAD consistently improves over its SD-LoRA backbone across all three task configurations. The gains in Acc/AAA are $0.83/0.99$, $0.64/1.29$, and $0.17/0.47$ percentage points for $N = 5$, 10, and 20, respectively. These consistent improvements demonstrate the benefit of parameter-space consolidation beyond the underlying parameter-efficient adaptation strategy. Across different task granularities, PSCAD maintains positive gains as the number of sequential tasks increases from 5 to 20, indicating that parameter-space consolidation remains effective as task-specific parameters accumulate.

On ImageNet-A, PSCAD achieves the highest Acc of 56.63, improving over SD-LoRA and ACMap by 0.67 and 0.44 points, respectively, while its AAA of 64.61 is slightly below SD-LoRA and ACMap. This difference indicates that the benefit of parameter consolidation can vary across datasets and metrics, with a clearer improvement in final accuracy on ImageNet-A. The discrepancy between Acc and AAA further highlights that final retention and trajectory-level performance capture different aspects of continual learning. Overall, the results support parameter-space consolidation as a complementary mechanism for parameter-efficient continual learning.

5.3 Lyapunov Energy Dissipation

Our theoretical analysis establishes non-increasing Lyapunov energy for the continuous-time parameter dynamics under fixed memory. To examine whether this property is reflected in the numerical dynamics, we track the total Lyapunov energy over 1,000 internal consolidation timesteps and plot its trajectory and logarithmic gap relative to the final recorded energy in Figure 2. We observe that the energy decreases rapidly in the early stage and gradually stabilizes as consolidation proceeds. The logarithmic energy gap decreases approximately linearly in the later stage, indicating an approximately exponential decay of the energy gap over this regime. These observations are consistent with the theoretical energy dissipation and convergence behavior of the continuous-time dynamics under the stated stability conditions.

5.4 Ablation Study

We next examine the contributions of the three PSCAD components: potential-weighted interaction, adaptation, and Hopfield-guided memory. All variants use the same SD-LoRA backbone and training protocol, with only the corresponding component removed. As shown in Table 3, removing any component consistently reduces both Acc and AAA. Although the performance differences are modest, their consistent direction indicates that the three components provide complementary contributions. Among them, removing Hopfield memory causes the largest drop in AAA (-0.19), while removing adaptation causes the largest drop in Acc (-0.24). This suggests that associative memory and adaptation contribute differently to the consolidated parameter states. Overall, the ablation results show that removing any component leads to a performance decrease, while the full model achieves the best results on both metrics.

Table 1: Results on ImageNet-R across different task distributions.

Method	ImageNet-R ($N = 5$)		ImageNet-R ($N = 10$)		ImageNet-R ($N = 20$)	
	Acc↑	AAA↑	Acc↑	AAA↑	Acc↑	AAA↑
Full Fine-Tuning	64.92	75.57	60.57	72.31	49.95	65.32
L2P	73.04	76.94	71.26	76.13	68.97	74.16
DualPrompt	69.99	72.24	68.22	73.81	65.23	71.30
CODA-Prompt	76.63	80.30	74.05	78.14	69.38	73.95
InfLoRA	76.95	81.81	74.75	80.67	69.89	76.68
SD-LoRA	79.15	83.01	77.34	82.04	75.26	80.22
PLAN	77.79	81.93	75.25	80.41	71.06	77.93
HippoTune	77.16	81.04	74.85	79.92	74.06	79.33
PSCAD	**79.98**	**84.00**	**77.98**	**83.33**	**75.43**	**80.69**

Table 2: Results on ImageNet-A ($N = 10$)

Method	ImageNet-A	
	Acc↑	AAA↑
Full Fine-Tuning	16.31	30.04
L2P	42.94	51.40
DualPrompt	45.49	54.68
CODA-Prompt	45.36	57.03
InfLoRA	49.20	60.92
SDLoRA	55.96	64.95
PROL	47.72	58.96
ACMap	56.19	**65.19**
PSCAD	**56.63**	64.61

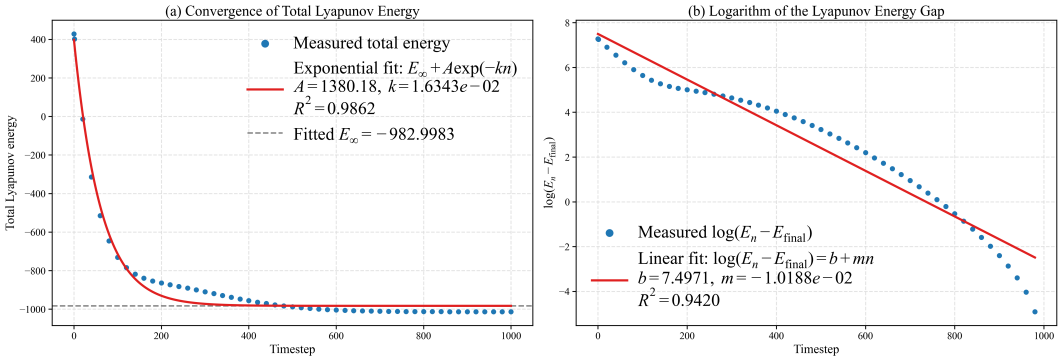

Figure 2: Evolution of the total Lyapunov energy. Left: measured energy and an exponential fit. Right: logarithmic energy gap relative to the final recorded energy and a linear fit.

5.5 Efficiency Analysis

The parameter-consolidation operator of PSCAD introduces an additional computational stage and stores task-level memory parameters. We therefore compare PSCAD with its SD-LoRA backbone in terms of computational cost, trainable parameter count, and peak GPU memory during training, as reported in Table 4. GFLOPs are measured with an input resolution of 224×224 and a batch size of 32. The trainable parameter count includes all parameters updated during the current task. Peak GPU memory is measured under the same batch size, numerical precision, and software environment after the initial warm-up iterations. Compared with SD-LoRA, PSCAD incurs additional computational and memory costs due to parameter consolidation and memory storage. We observe that the differences between PSCAD and SD-LoRA backbone remain limited, indicating that the ad-

Table 3: Ablation Study on ImageNet-R ($N = 5$)

Method	ImageNet-R ($N = 5$) Acc↑	AAA↑
Full Model	79.98	84.00
Without Interaction	79.76	83.97
Without Adaptation	79.74	83.90
Without Hopfield	79.75	83.81

Table 4: Efficiency Analysis on ImageNet-R ($N = 5$)

Method	ImageNet-R ($N = 5$) GFLOPs/batch ↓	Trainable parameters/M ↓	Peak GPU memory/GiB ↓
SD-LoRA	2336.46	0.52	5.41
PSCAD	4673.36	2.00	5.41

ditional resource requirements of the parameter-consolidation operator are practically manageable and worthwhile.

6 Discussion and Conclusion

In this paper, we presented PSCAD as a parameter-space perspective on rehearsal-free continual learning. Rather than treating task-specific parameters as isolated solutions that are simply retained after training, PSCAD regards them as a collective dynamical system that can be reorganized and consolidated after task learning. This separation between knowledge acquisition and parameter consolidation provides a complementary view of continual learning, in which learned parameters themselves become the substrate for subsequent memory processing.

An important consequence of this perspective is that consolidation operates directly on learned parameters rather than historical data. Thus, PSCAD does not require replaying or generating past examples during consolidation, while still allowing accumulated task parameters to interact and be guided by persistent memory. At the same time, the current formulation retains task-specific parameters and therefore incurs increasing parameter and computational costs as the number of tasks grows. In addition, the consolidation dynamics and memory update are currently separated in time, leaving open the possibility of jointly evolving parameter states and memory during consolidation.

Note that our theoretical and empirical results support parameter-space consolidation as a complementary mechanism for continual learning. Beyond the specific PSCAD formulation, this perspective suggests that the parameters produced during sequential learning need not remain passive records of past tasks, but can participate in an additional dynamical process. Future work could investigate more scalable consolidation mechanisms, adaptive memory capacity, and jointly evolving parameter-memory dynamics for larger-scale and more diverse continual learning settings.

AI use statement

In this work, we used generative AI tools for assisting in the writing of proofs and assisting with translation. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, designing or providing feedback on research methodology or experiments, implementing methods, supporting qualitative and thematic data analysis and interpreting results, and generating synthetic data sets, cleaning and reformatting dataset and proposing or refining hypotheses are not applicable to this work. Additionally, we used generative AI tools for creating or editing software code, sourcing/searching for information and editing a research paper to improve readability. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We checked the authenticity of LLM-generated translation and information collected by LLM through a manual literature survey, and reviewed the language polishing for appropriateness.

We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Nick Alonso and Jeff Krichmar. A sparse quantized hopfield network for online-continual memory. *CoRR*, abs/2307.15040, 2023. doi: 10.48550/ARXIV.2307.15040. URL https://doi.org/10.48550/arXiv.2307.15040.

Shun-ichi Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87, 1977. URL https://api.semanticscholar.org/CorpusID:2811608.

Matteo Boschini, Lorenzo Bonicelli, Angelo Porrello, Giovanni Bellitto, Matteo Pennisi, Simone Palazzo, Concetto Spampinato, and Simone Calderara. Transfer without forgetting. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXIII*, volume 13683 of *Lecture Notes in Computer Science*, pp. 692–709. Springer, 2022. doi: 10.1007/978-3-031-20050-2_40. URL https://doi.org/10.1007/978-3-031-20050-2_40.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *International Conference on Computational Statistics*, 2010. URL https://api.semanticscholar.org/CorpusID:115963355.

Svenja Brodt, Marion Inostroza, Niels Niethard, and Jan Born. Sleep–a brain-state serving systems memory consolidation. *Neuron*, 111(7):1050–1075, 2023. doi: 10.1016/j.neuron.2023.03.005.

Yanxi Chen, Xiuxing Li, Yuyang Han, Zhuo Wang, Qing Li, Ziyu Li, Xiang Li, Chen Wei, and Xia Wu. Hippotune: A hippocampal associative loop–inspired fine-tuning method for continual learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=MtDiLnnYgm.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385, 2022. doi: 10.1109/TPAMI.2021.3057446.

Xingsi Dong, Zilong Ji, Tianhao Chu, Tiejun Huang, Wenhao Zhang, and Si Wu. Adaptation accelerating sampling-based bayesian inference in attractor neural networks. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/870c1e0589822bf37590b84984c345c4-Abstract-Conference.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Nicolas Dumay and M. Gareth Gaskell. Sleep-associated changes in the mental representation of spoken words. *Psychological Science*, 18:35 – 39, 2007. URL https://api.semanticscholar.org/CorpusID:6017785.

Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A. Rusu, Alexander Pritzel, and Daan Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *CoRR*, abs/1701.08734, 2017. URL http://arxiv.org/abs/1701.08734.

Takuma Fukuda, Hiroshi Kera, and Kazuhiko Kawamoto. Adapter merging with centroid prototype mapping for scalable class-incremental learning. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4884–4893, 2025. URL https://api.semanticscholar.org/CorpusID:274992202.

C. C. Alan Fung, K. Y. Michael Wong, Yuanyuan Mi, Wen-Hao Zhang, and Si Wu. Continuous attractor neural networks: Candidate of a canonical model for neural information representation. *F1000Research*, 5, 2016. URL https://api.semanticscholar.org/CorpusID:31349693.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 8320–8329. IEEE, 2021a. doi: 10.1109/ICCV48922.2021.00823. URL https://doi.org/10.1109/ICCV48922.2021.00823.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 15262–15271. Computer Vision Foundation / IEEE, 2021b. doi: 10.1109/CVPR46437.2021.01501. URL https://openaccess.thecvf.com/content/CVPR2021/html/Hendrycks_Natural_Adversarial_Examples_CVPR_2021_paper.html.

John J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79(8): 2554–2558, 1982. URL https://api.semanticscholar.org/CorpusID:784288.

Mikail Khona and Ila Rani Fiete. Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, 23:744 – 766, 2022. doi: https://doi.org/10.1038/s41583-022-00642-0. URL https://api.semanticscholar.org/CorpusID:244954629.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Jens G. Klinzing, Niels Niethard, and Jan Born. Mechanisms of systems memory consolidation during sleep. *Nature Neuroscience*, 22(10):1598–1610, 2019. doi: 10.1038/s41593-019-0467-3.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 23638–23647. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02231. URL https://doi.org/10.1109/CVPR52733.2024.02231.

Lihong Lin and Haidong Kang. Unlocking the potential of continual model merging: An ODE perspective. *CoRR*, abs/2605.19409, 2026. doi: 10.48550/ARXIV.2605.19409. URL https://doi.org/10.48550/arXiv.2605.19409.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 7765–7773. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00810. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Mallya_PackNet_Adding_Multiple_CVPR_2018_paper.html.

Alessio Masano, Giovanni Bellitto, Dipam Goswani, Joost Van de Weijer, and Concetto Spampinato. Routing without forgetting, 2026. URL https://arxiv.org/abs/2603.09576.

M. A. Ma'sum, Mahardhika Pratama, Savitha Ramasamy, Lin Liu, Habibullah Habibullah, and Ryszard Kowalczyk. Prol: Rehearsal free continual learning in streaming data via prompt online learning. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2471–2481, 2025. URL https://api.semanticscholar.org/CorpusID:280274946.

James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419–457, 1995. URL https://api.semanticscholar.org/CorpusID:2832081.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pp. 109–165. Elsevier, 1989. doi: 10.1016/S0079-7421(08)60536-8.

Ramy Mounir and Sudeep Sarkar. Predictive attractor models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/5df4313ecd4875931fbdacc486cc1fcf-Abstract-Conference.html.

Fuli Qiao and Mehrdad Mahdavi. Merge before forget: A single lora continual learning via continual merging. *CoRR*, abs/2512.23017, 2025. doi: 10.48550/ARXIV.2512.23017. URL https://doi.org/10.48550/arXiv.2512.23017.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David P. Kreil, Michael K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5533–5542. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.587. URL https://doi.org/10.1109/CVPR.2017.587.

Mengye Ren, Renjie Liao, Ethan Fetaya, and Richard S. Zemel. Incremental few-shot learning with attention attractor networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 5276–5286, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/e833e042f509c996b1b25324d56659fb-Abstract.html.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Gregory Wayne. Experience replay for continual learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 348–358, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/fa7cdfad1a5aaf8370ebeda47a1ff1c3-Abstract.html.

Joan Serrà, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4555–4564. PMLR, 2018. URL http://proceedings.mlr.press/v80/serra18a.html.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio,

Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 2990–2999, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/0efbe98067c6c73dba1250d2beaa81f9-Abstract.html.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogério Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 11909–11919. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01146. URL https://doi.org/10.1109/CVPR52729.2023.01146.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(8):5362–5383, 2024. doi: 10.1109/TPAMI.2024.3367329. URL https://doi.org/10.1109/TPAMI.2024.3367329.

Xiequn Wang, Zhan Zhuang, and Yu Zhang. PLAN: proactive low-rank allocation for continual learning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 2909–2918. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00279. URL https://doi.org/10.1109/ICCV51701.2025.00279.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. Dualprompt: Complementary prompting for rehearsal-free continual learning. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVI*, volume 13686 of *Lecture Notes in Computer Science*, pp. 631–648. Springer, 2022a. doi: 10.1007/978-3-031-19809-0_36. URL https://doi.org/10.1007/978-3-031-19809-0_36.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. Learning to prompt for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 139–149. IEEE, 2022b. doi: 10.1109/CVPR52688.2022.00024. URL https://doi.org/10.1109/CVPR52688.2022.00024.

Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. Sd-lora: Scalable decoupled low-rank adaptation for class incremental learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=5U1rlpX68A.

Enneng Yang, Anke Tang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, and Jie M. Zhang. Continual model merging without data: Dual projections for balancing stability and plasticity. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/37d9f19150fce07bced2a81fc87d47a6-Abstract-Conference.html.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995. PMLR, 2017. URL http://proceedings.mlr.press/v70/zenke17a.html.

A Appendix

A.1 Lyapunov Energy Dissipation

Proof of Theorem 1. Define the joint parameter energy

$$G(\theta_{1:t}) = \Psi(\theta_{1:t}) + \sum_{i=1}^{t} E_H(\theta_i, \Phi). \tag{14}$$

The parameter dynamics in equation 1 can be written as

$$\tau_s \dot{\theta}_i = -\theta_i - V_i - \nabla_{\theta_i} G(\theta_{1:t}). \tag{15}$$

Rearranging gives

$$\nabla_{\theta_i} G(\theta_{1:t}) = -\theta_i - V_i - \tau_s \dot{\theta}_i. \tag{16}$$

Differentiating the total energy with respect to θ_i yields

$$\nabla_{\theta_i} \mathcal{E} = \nabla_{\theta_i} G + (1+m)\theta_i - (V_i - m\theta_i). \tag{17}$$

Substituting the gradient of the joint parameter energy $G(\theta_{1:t})$, we obtain

$$\begin{aligned} \nabla_{\theta_i} \mathcal{E} = & -\theta_i - V_i - \tau_s \dot{\theta}_i + (1+m)\theta_i - V_i + m\theta_i \\ = & -\tau_s \dot{\theta}_i + 2(m\theta_i - V_i). \end{aligned} \tag{18}$$

From equation 2,

$$m\theta_i - V_i = \tau_z \dot{V}_i, \tag{19}$$

and hence

$$\nabla_{\theta_i} \mathcal{E} = -\tau_s \dot{\theta}_i + 2\tau_z \dot{V}_i. \tag{20}$$

Similarly, differentiating with respect to V_i gives

$$\begin{aligned} \nabla_{V_i} \mathcal{E} &= \frac{1}{m}(V_i - m\theta_i) \\ &= -\frac{\tau_z}{m} \dot{V}_i. \end{aligned} \tag{21}$$

The time derivative of the Lyapunov energy is therefore

$$\begin{aligned} \dot{\mathcal{E}} &= \sum_{i=1}^{t} \left[\nabla_{\theta_i} \mathcal{E}^\top \dot{\theta}_i + \nabla_{V_i} \mathcal{E}^\top \dot{V}_i \right] \\ &= \sum_{i=1}^{t} \left[-\tau_s \|\dot{\theta}_i\|_2^2 + 2\tau_z \dot{V}_i^\top \dot{\theta}_i - \frac{\tau_z}{m} \|\dot{V}_i\|_2^2 \right]. \end{aligned} \tag{22}$$

Completing the square gives

$$\begin{aligned} \dot{\mathcal{E}} = & -(\tau_s - m\tau_z) \sum_{i=1}^{t} \|\dot{\theta}_i\|_2^2 \\ & - \frac{\tau_z}{m} \sum_{i=1}^{t} \|\dot{V}_i - m\dot{\theta}_i\|_2^2. \end{aligned} \tag{23}$$

Under the strict condition $\tau_s > m\tau_z$, equality requires $\dot{\theta}_i = 0$ for every i. The second term then further requires $\dot{V}_i = 0$. Thus, zero energy dissipation occurs if and only if the system is at equilibrium. □

A.2 Lower Boundedness of the Lyapunov Energy

The monotonic decrease in Theorem 1 does not imply that the energy decreases without bound. Assume that the attractor memory is fixed during the internal consolidation dynamics and define

$$R_\Phi = \max_{1\le\mu\le M} \|\phi_\mu\|_2. \tag{24}$$

Using the Log-Sum-Exp bound

$$\mathrm{LSE}(\mathbf{z}) \le \max_\mu z_\mu + \log M \tag{25}$$

and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\mathrm{LSE}\left(\beta\Phi^\top\theta_i\right) &\le \log M + \beta \max_\mu \phi_\mu^\top\theta_i \\ &\le \log M + \beta R_\Phi \|\theta_i\|_2.\end{aligned} \tag{26}$$

Since $\beta > 0$, multiplying both sides by $-1/\beta$ reverses the inequality and gives

$$E_H(\theta_i;\Phi) \ge -\frac{\log M}{\beta} - R_\Phi\|\theta_i\|_2. \tag{27}$$

Combining this bound with the quadratic state energy and completing the square yields

$$\begin{aligned}&E_H(\theta_i;\Phi) + \frac{1+m}{2}\|\theta_i\|_2^2 \\ &\quad\ge -\frac{\log M}{\beta} + \frac{1+m}{2}\left(\|\theta_i\|_2 - \frac{R_\Phi}{1+m}\right)^2 - \frac{R_\Phi^2}{2(1+m)} \\ &\quad\ge -\frac{\log M}{\beta} - \frac{R_\Phi^2}{2(1+m)}.\end{aligned} \tag{28}$$

Moreover, since $0 < K(\theta_i,\theta_j) \le 1$, the potential term satisfies

$$\Psi(\theta_{1:t}) \ge -\frac{\rho J_0 t(t-1)}{2}, \tag{29}$$

while the adaptation energy is non-negative:

$$\frac{1}{2m}\sum_{i=1}^{t}\|V_i - m\theta_i\|_2^2 \ge 0. \tag{30}$$

Consequently, the total energy admits the finite lower bound

$$\mathcal{E}(\theta_{1:t}, V_{1:t}) \ge -\frac{t\log M}{\beta} - \frac{tR_\Phi^2}{2(1+m)} - \frac{\rho J_0 t(t-1)}{2}. \tag{31}$$

A.3 Compactness of Sublevel Sets and Convergence

We next verify the compactness condition required by LaSalle's invariance principle. Let

$$x = (\theta_1,\dots,\theta_t, V_1,\dots,V_t) \tag{32}$$

denote the joint state of the dynamics. For each task, the quadratic part of the Lyapunov function can be written as

$$\begin{aligned}Q_i(\theta_i,V_i) &= \frac{1+m}{2}\|\theta_i\|_2^2 + \frac{1}{2m}\|V_i - m\theta_i\|_2^2 \\ &= \frac{1}{2}\begin{bmatrix}\theta_i \\ V_i\end{bmatrix}^\top \left(\begin{bmatrix}1+2m & -1 \\ -1 & 1/m\end{bmatrix}\otimes I_d\right)\begin{bmatrix}\theta_i \\ V_i\end{bmatrix}.\end{aligned} \tag{33}$$

For $m > 0$, the matrix

$$A_m = \begin{bmatrix}1+2m & -1 \\ -1 & 1/m\end{bmatrix} \tag{34}$$

is positive definite because

$$1 + 2m > 0, \qquad \det(A_m) = \frac{1+m}{m} > 0. \tag{35}$$

Therefore, there exists a constant $c_m > 0$ such that

$$\sum_{i=1}^{t} Q_i(\theta_i, V_i) \geq c_m \sum_{i=1}^{t} \left(\|\theta_i\|_2^2 + \|V_i\|_2^2 \right) = c_m \|x\|_2^2. \tag{36}$$

Moreover, the Hopfield energy satisfies

$$\sum_{i=1}^{t} E_H(\theta_i; \Phi) \geq -\frac{t \log M}{\beta} - R_\Phi \sum_{i=1}^{t} \|\theta_i\|_2, \tag{37}$$

and the potential satisfies

$$\Psi(\theta_{1:t}) \geq -\frac{\rho J_0 t(t-1)}{2}. \tag{38}$$

Using

$$\sum_{i=1}^{t} \|\theta_i\|_2 \leq \sqrt{t} \left(\sum_{i=1}^{t} \|\theta_i\|_2^2 \right)^{1/2} \leq \sqrt{t} \|x\|_2, \tag{39}$$

we obtain

$$\mathcal{E}(x) \geq c_m \|x\|_2^2 - R_\Phi \sqrt{t} \|x\|_2 - C, \tag{40}$$

where

$$C = \frac{t \log M}{\beta} + \frac{\rho J_0 t(t-1)}{2}. \tag{41}$$

The quadratic term in equation 40 dominates the linear term as $\|x\|_2 \to \infty$. Hence,

$$\mathcal{E}(x) \to \infty \qquad \text{as } \|x\|_2 \to \infty, \tag{42}$$

which shows that $\mathcal{E}$ is radially unbounded.

For an initial state $x(0)$, define the sublevel set

$$\Omega_{x(0)} = \{x : \mathcal{E}(x) \leq \mathcal{E}(x(0))\}. \tag{43}$$

Since $\mathcal{E}$ is continuous and radially unbounded, $\Omega_{x(0)}$ is closed and bounded, and is therefore compact in the finite-dimensional state space. Furthermore, Theorem 1 gives $\dot{\mathcal{E}} \leq 0$, so

$$\mathcal{E}(x(s)) \leq \mathcal{E}(x(0)), \qquad s \geq 0. \tag{44}$$

Thus, every trajectory starting from $x(0)$ remains in $\Omega_{x(0)}$, which makes this sublevel set positively invariant.

Let

$$\mathcal{Z} = \left\{ x \in \Omega_{x(0)} : \dot{\mathcal{E}}(x) = 0 \right\}. \tag{45}$$

Under the strict condition $\tau_s > m\tau_z$, equation 23 implies

$$\dot{\mathcal{E}}(x) = 0 \quad \Longleftrightarrow \quad \dot{\theta}_i = 0 \text{ and } \dot{V}_i = 0, \quad i = 1, \ldots, t. \tag{46}$$

Therefore, $\mathcal{Z}$ coincides with the equilibrium set of the coupled dynamics within $\Omega_{x(0)}$. By LaSalle's invariance principle, every trajectory starting from $x(0)$ approaches the largest invariant subset of $\mathcal{Z}$. Since every point in $\mathcal{Z}$ is an equilibrium, this largest invariant subset is $\mathcal{Z}$ itself. Consequently,

$$\operatorname{dist}(x(s), \mathcal{Z}) \longrightarrow 0 \qquad \text{as } s \to \infty. \tag{47}$$

Hence, the buffer state converges to the equilibrium set of the dynamics.