Towards Addressing the Plasticity-Stability Dilemma in Continual Reinforcement Learning

Qingchen Hu, Hongyao Tang, Glen Berseth

Keywords: Continual Reinforcement Learning, Plasticity Loss, Catastrophic Forgetting

Summary

Continual reinforcement learning (CRL) presents a fundamental challenge in sequential decision-making, requiring agents to continuously acquire new skills while retaining previously learned behaviours. A key difficulty in CRL is balancing plasticity, the ability to adapt to new tasks, with stability, the preservation of past knowledge. In this work, we systematically evaluate the plasticity-stability trade-off in CRL by empirically benchmarking a diverse set of existing methods based on Proximal Policy Optimization (PPO) across a sequence of MinAtar games. Our findings reveal that most existing approaches tend to favor either plasticity or stability, with no single method consistently performing the best across all scenarios. Motivated by these findings, we propose Rehearsal CRL with Marginal L2 (REMARL2), which integrates selective weight regularization with experience rehearsal. Our method achieves a better balance in the plasticity-stability trade-off, demonstrating the effectiveness of hybrid strategies that combine plasticity loss prevention and knowledge retention for improving CRL.

Contribution(s)

1. We conduct a comprehensive empirical study of CRL using PPO as the base agent on MinAtar games (Young & Tian, 2019), systematically evaluating how different existing methods behave in terms of the plasticity-stability trade-off. Our benchmarking analysis highlights the strengths and limitations of existing methods, showing that most methods tend to favor either plasticity or stability often with a sacrifice on the other aspect, struggling to achieve a good balance.

   Context: While prior studies have examined catastrophic forgetting (van de Ven et al., 2024; Hayes et al., 2020) and plasticity loss (Juliani & Ash, 2024; Lyle et al., 2023; Abbas et al., 2023), their evaluations are often constrained to specific tasks or supervised learning settings. Our work provides a unified analysis of these challenges in reinforcement learning, revealing the conditions under which different mitigation strategies succeed or fail, offering insights for designing more effective continual learning methods.

2. We propose MARGINAL L2 and REHEARSAL REGULARIZATION as complementary approaches to address the plasticity-stability trade-off in CRL. By integrating these methods, we introduce REMARL2, which achieves superior knowledge retention while maintaining strong plasticity across tasks. Empirical results demonstrate that REMARL2 provides a robust solution for CRL in dynamic environments.

   Context: Existing methods in CRL often fail to balance plasticity and stability effectively. Plasticity-oriented approaches, such as L2 REGULARIZATION (Lyle et al., 2023), perform well in plasticity but compromise knowledge retention. On the other hand, stability-oriented methods, like EWC (Kirkpatrick et al., 2016) , prioritize retaining past knowledge but limit flexibility, hindering the system's ability to effectively learn new tasks. Rescaling-based methods mitigate non-stationarity by adjusting for environmental changes but emphasize plasticity over stability, leading to performance instability. These limitations highlight the need for a more balanced approach.

Towards Addressing the Plasticity-Stability Dilemma in Continual Reinforcement Learning

Qingchen Hu[1,3†], **Hongyao Tang**[2,3†], **Glen Berseth**[2,3†]
qingchen.hu@mail.mcgill.ca, {tang.hongyao,glen.berseth}@@mila.quebec

[1]**McGill University**
[2]**Université de Montréal**
[3]**Mila – Quebec AI Institute**

† Equal contribution.

Abstract

Continual reinforcement learning (CRL) presents a fundamental challenge in sequential decision-making, requiring agents to continuously acquire new skills while retaining previously learned behaviours. A key difficulty in CRL is balancing plasticity, the ability to adapt to new tasks, with stability, the preservation of past knowledge. In this work, we systematically evaluate the plasticity-stability trade-off in CRL by empirically benchmarking a diverse set of existing methods based on Proximal Policy Optimization (PPO) across a sequence of MinAtar games. Our findings reveal that most existing approaches tend to favor either plasticity or stability, with no single method consistently performing the best across all scenarios. Motivated by these findings, we propose Rehearsal CRL with Marginal L2 (REMARL2), which integrates selective weight regularization with experience rehearsal. Our method achieves a better balance in the plasticity-stability trade-off, demonstrating the effectiveness of hybrid strategies that combine plasticity loss prevention and knowledge retention for improving CRL.

1 Introduction

Continual reinforcement learning (CRL) (Thrun, 1994; Stulp, 2012; Khetarpal et al., 2022) remains a fundamental yet unresolved challenge in sequential decision-making tasks. Unlike learning from a fixed dataset or a stationary environment, non-stationary environments introduce unique difficulties. Two primary issues arise in this setting: **plasticity loss**, where an online-trained neural network struggles to adapt to new tasks (Abbas et al., 2023; Juliani & Ash, 2024), and **catastrophic forgetting**, where previously learned knowledge is rapidly erased when new tasks are introduced (Khetarpal et al., 2022; van de Ven et al., 2024; Hayes et al., 2020).

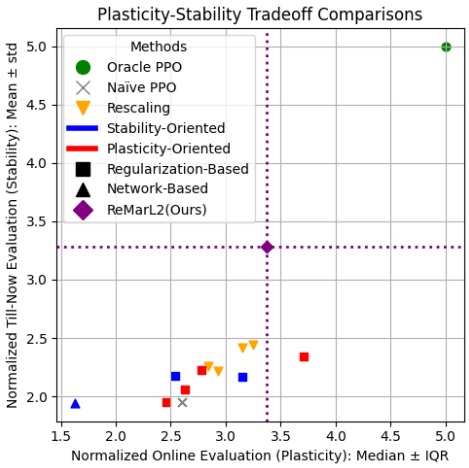


Figure 1: Plasticity-stability trade-off across methods. Higher values indicate better plasticity (x-axis) and knowledge retention (y-axis). REMARL2 achieves the best balance.

CRL requires agents to be sufficiently plastic to adapt to evolving tasks while being stable to retain prior knowledge. Consistent efforts have been devoted in the literature to achieve the desiderata from different perspectives. Regularization-based methods (Kirkpatrick et al., 2016; Elsayed et al., 2024)

mitigate forgetting but hinder plasticity, while adaptive architectures (Rusu et al., 2016; Nagabandi et al., 2019; Zhang et al., 2024) and resetting mechanisms (Schwarz et al., 2018b; Lyle et al., 2023; Kumar et al., 2024) improve plasticity at the cost of stability. Existing methods struggle to balance the plasticity-stability trade-off. How to properly balance the stability-plasticity trade-off remains underexplored, particularly in recent CRL methods, where mitigation strategies often optimize one aspect at the expense of the other (Abbas et al., 2023; van de Ven et al., 2024).

To this end, we conduct a systematic empirical investigation into the stability-plasticity trade-off across a diverse set of existing methods in CRL, using a continual configuration of MinAtar (Ceron & Castro, 2021; Gogianu et al., 2021) to better understand the strengths of different methods. Our primary objective is to identify methods that achieve the strongest plasticity while minimizing forgetting.

We begin by evaluating a standard PPO agent to establish a baseline for its plasticity and susceptibility to catastrophic forgetting. We assess existing methods designed to either enhance plasticity or prevent forgetting, observing that they predominantly favor one aspect at the expense of the other. An overview of the benchmarking results is in Figure 1. The balance between stability and plasticity methods is non-trivial. Adding to much regularization for stability hampers plasticity and visa-versa. To address this delicate balance, we propose Rehearsal CRL with Marginal L2 (REMARL2), a memory-based method that integrates selective weight regularization with experience rehearsal. Empirical results demonstrate that REMARL2 outperforms existing approaches by achieving a more effective balance between plasticity and stability, thereby demonstrating the effectiveness of hybrid strategies that combine plasticity loss prevention and knowledge retention in CRL settings.

2 Problem Description

2.1 Preliminaries

In this study, we adopt the standard formulation of reinforcement learning (Sutton & Barto, 2018) under the Markov decision process (MDP) framework. An MDP is defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ represents the state space, $\mathcal{A}$ the action space, $\mathcal{P}(s'|s,a)$ the transition dynamics, $\mathcal{R}(s,a)$ the reward function, and $\gamma \in [0,1)$ the discount factor. At each timestep, the agent observes a state $s \in \mathcal{S}$, selects an action $a \in \mathcal{A}$, and receives a reward $r = \mathcal{R}(s,a)$. The environment then transitions to a new state $s' \sim \mathcal{P}(s'|s,a)$, and this interaction process repeats. The agent's objective is to learn a policy $\pi_\theta(a|s)$, parameterized by θ, that maximizes the expected discounted return: $J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{\infty}\gamma^t\mathcal{R}(s_t,a_t)\right]$, where $\tau = (s_0, a_0, s_1, a_1, \dots)$ denotes trajectories sampled under the policy π_θ. This formulation captures the agent's goal of optimizing long-term cumulative rewards while navigating the environment's dynamics.

2.2 Continual Reinforcement Learning

Continual learning is the process of designing learning algorithms that perform well under changes in tasks or distributions (Wang et al., 2024). In continual RL (Ring, 1997; Khetarpal et al., 2022), the agent sequentially encounters K distinct tasks, each corresponding to a unique MDP $\mathcal{M}_i = (\mathcal{S}_i, \mathcal{A}_i, \mathcal{P}_i, \mathcal{R}_i, \gamma_i)$, where $i \in \{1, \dots, K\}$. Tasks may differ in dynamics ($\mathcal{P}_i$), rewards ($\mathcal{R}_i$), or both. The agent must learn these tasks incrementally without revisiting prior task data, leading to challenges in balancing plasticity (adapting to new tasks) and stability (retaining knowledge of previous tasks).

2.2.1 Evaluation Metrics

Evaluating continual reinforcement learning methods solely by average episodic rewards (van de Ven et al., 2024) fails to capture the trade-off between learning new tasks and retaining past knowledge. To address this, we define two key evaluation metrics:

Online Evaluation (No-Lookback) This metric quantifies plasticity by measuring the agent's cumulative performance on all tasks encountered so far. For a policy π_i trained on task T_i, we record the final episodic reward $\bar{J}_i(\pi_i) = \mathbb{E}_{\tau\sim\pi_i}\left[\sum_{t=0}^{T} r_t^i\right]$ (in practice the last few episodes are used), achieved on task T_i and compute the Online Evaluation for each task as:

$$\text{Online}(\{\pi_i\}_{i=1}^{K}) = \sum_{i=1}^{K} \bar{J}_i(\pi_i). \tag{1}$$

Till-Now Evaluation (Lookback) To assess stability, we track the retention of earlier tasks as new ones are introduced using the Area Under the Curve (AUC) metric, which is averaged over global steps interpolated at certain intervals (e.g., every 10K steps in our experiments). Specifically, for each task i, the AUC of the policy's performance is computed across the entire training process:

$$\text{Till-Now}_i(\{\pi_j\}_{j=i+1}^{K}) = \sum_{j=i+1}^{K} \text{AUC}\left(\bar{J}_i(\pi_j)\right). \tag{2}$$

A decline in performance on earlier tasks indicates catastrophic forgetting, whereas stable retention suggests effective knowledge preservation.

To account for variations in reward scales across tasks, we normalize the scores using the standard approach proposed by Aitchison et al. (2023): $Z_i(x) = 1 + \max\left(0, \frac{x_i - r_i}{h_i - r_i}\right)$, where r_i, h_i denote the baseline score and the reference score respectively.

2.2.2 Key Challenges in Continual Reinforcement Learning

To better understand the underlying causes of the challenges in CRL, we highlight two key factors that significantly impact learning performance in CRL settings.

Task Reward Scale Variations Sequential decision-making tasks often exhibit significant differences in reward magnitudes, complexity, and dynamics. These discrepancies result in imbalanced gradient updates, where tasks with extreme reward values (either excessively high or low) dominate learning, suppressing the model's ability to differentiate between reward scales effectively. As a result, learning can become unstable, leading to poor adaptation across tasks (Henderson et al., 2017; Mnih et al., 2013; Hafez & Erekmen, 2024; Schaul et al., 2021; van Hasselt et al., 2016; Dann & Thangarajah, 2021; van Hasselt et al., 2016).

Non-Stationarity in Data and Objectives In CRL, agents must operate in non-stationary environments where both data distributions and learning objectives evolve over time. This continual shift challenges traditional optimization methods, which are typically designed for stationary environments. As a result, models often fail to balance plasticity with knowledge retention, leading to inconsistent learning performance when encountering new tasks or objective shifts (Berseth et al., 2018; Schwarz et al., 2018a; Abbas et al., 2023; Hafez & Erekmen, 2024; Khetarpal et al., 2020; Xie et al., 2021; Padakandla et al., 2019; Igl et al., 2020).

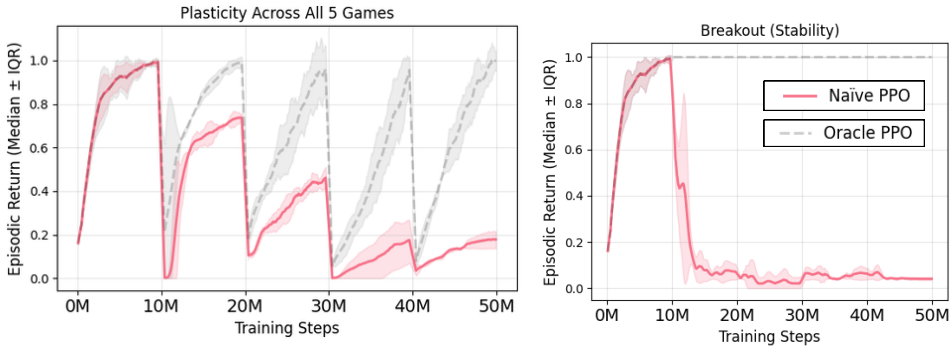
3 Forgetting and Plasticity Loss in Continual RL

In this section, we investigate the inabilities of a standard RL agent in CRL setting, highlighting the extent of catastrophic forgetting and plasticity loss. Through controlled experiments, we analyze the impact of sequential learning on performance and establish baseline calibrations for evaluating existing remedy strategies and our proposed method in later sections.

For our experiments, we consider a sequence of five Atari games from MinAtar (Young & Tian, 2019), a benchmark designed to reduce representational complexity while preserving key behavioral challenges. The five tasks included in our experimental setup are Breakout, Freeway, Asterix, Seaquest, and Space Invaders. This task order is denoted as T_i for $i = 1, 2, \ldots, 5$. It is important

to note that our experimental setup, based on Proximal Policy Optimization (PPO) (Schulman et al., 2017), is not designed to achieve state-of-the-art performance for each game. Instead, our focus is on understanding the impact of different methods on the plasticity-stability trade-off in CRL.

To establish a baseline, we define a NAÏVE PPO agent, which continuously updates model parameters across sequential tasks without additional constraints or methods applied. This corresponds to a standard PPO agent implemented using Weng et al. (2022)'s framework. The agent is initialized once at the beginning of training and learns continually across 50 million timesteps over a sequence of tasks. For comparison, we define an ORACLE PPO agent as an upper-bound benchmark for continual learning performance. This agent is fully reinitialized at the start of each task, ensuring maximum plasticity, and is assumed to have perfect knowledge retention of all previously trained environments. This assumption reflects a virtual agent with no stability constraints, achieving both perfect plasticity and stability. As illustrated in Figure 2, the ORACLE PPO agent retains its full performance on previous tasks throughout the training sequence in Breakout, serving as a theoretical benchmark for evaluating stability. A truly effective CRL method should approach the ORACLE PPO agent's plasticity while maintaining stability across previously encountered tasks.



(a) Concatenated performance in the given task order (b) Breakout

Figure 2: Performance of NAÏVE PPO and ORACLE PPO agents as normalized episodic return (median), plotted with a running window size of 10. Curves are averaged over six seeds, with shaded regions indicating the interquartile range (IQR). (a) Performance during continual training on Breakout. (b) Concatenated performance across five games in the specified task order.

Through the behaviours of the two agents in Figure 2, we observe that when the naive agent is sequentially trained over five games, it rapidly and drastically forgets the first task while training on later ones. Moreover, the agent exhibits plasticity loss when encountering later games, indicating that upon experiencing a new environmental distribution shift, it fails to learn as effectively as a freshly initialized network, achieving performance as low as 20% of the original agent. Notably, when contrasted with the ideal agent, which is effectively trained in an interleaved manner across five games, the same network successfully learns all five tasks, demonstrating that catastrophic forgetting and plasticity loss is not a consequence of limited model capacity but rather a result of the training paradigm and the inability to balance plasticity and stability.

4 Benchmarking Existing Methods on the Plasticity-Stability Tradeoff

In this section, we benchmark existing methods on the plasticity-stability tradeoff. We first assess reward rescaling as a simple but straightforward solution for mitigating non-stationarity across environments. Next, we evaluate approaches aimed at enhancing plasticity and mitigating forgetting, highlighting their respective strengths and limitations.

4.1 Existing Methods

To systematically analyze existing approaches for addressing plasticity loss and catastrophic forgetting, we classify them into two primary categories based on their underlying principles: **Plasticity-oriented approaches** enhance plasticity by reducing overly restrictive regularization constraints, while **Stability-oriented strategies** constrain weight updates to mitigate catastrophic forgetting. These methods can further be classified as either TASK-AWARE or TASK-AGNOSTIC, depending on whether they require explicit awareness of distributional shifts:

- **TASK-AWARE methods**: These methods assume knowledge of when task boundaries occur, allowing for targeted adjustments. However, detecting distribution shifts is often infeasible in real-world settings.
- **TASK-AGNOSTIC methods**: These approaches are agnostic to task transitions, applying constraints uniformly throughout training without requiring shift detection.

For plasticity- and stability-oriented methods, Table 1 summarizes key approaches, their objectives, and corresponding formulations. Notably, we introduce a TASK-AWARE variation of Regenerative Loss: PERIODIC REGENERATIVE LOSS, which, rather than constraining the model to its initial weights, regularizes parameters toward those of the trained agent from the most recent task. This adaptation enhances plasticity while preserving stability, addressing limitations of the original formulation.

Table 1: Comparison of plasticity-oriented and stability-oriented methods in CRL. Here, θ_i represents network parameters after training on task i, θ_i^0 denotes initial weights, and θ_{i-1} refers to parameters from the previous task. λ is the regularization coefficient and c is the clipping threshold. $\mathbf{W}_i$ denotes the weight matrix of layer i, $\mathbf{I}$ is the identity matrix, and F_j represents the Fisher information matrix. The notation $\|\cdot\|_F$ denotes the Frobenius norm, and $\mathbf{m}$ is a pruning mask.

Method	Objective	Category	Key Formulation
Plasticity-Oriented Methods			
L2 REGULARIZATION (Lyle et al., 2023)	Prevents weight overfitting while allowing adaptation	Regularization-Based	$L_{\text{L2}} = \lambda \sum_i \|\theta_i\|_2^2$
REGENERATIVE LOSS (Kumar et al., 2024)	Encourages weight updates by penalizing deviations from initial parameters	Regularization-Based	$L_{\text{reg}} = \lambda \sum_i \|\theta_i - \theta_i^0\|_2^2$
CRELU ACTIVATIONS (Abbas et al., 2023)	Maintains non-zero gradients for better plasticity	Network-Based	$f_{\text{CReLU}}(x) = [\max(0, x), \max(0, -x)]$
WEIGHT CLIPPING (Elsayed et al., 2024)	Limits excessive weight growth while maintaining flexibility	Regularization-Based	$\theta \leftarrow \max(-c, \min(\theta, c))$
PARSEVAL REGULARIZATION (Chung et al., 2025)	Enforces weight orthogonality to improve optimization stability	Regularization-Based	$L_{\text{Parseval}} = \lambda \sum_i \|\mathbf{W}_i^\top \mathbf{W}_i - \mathbf{I}\|_F^2$
Stability-Oriented Methods			
PERIODIC REGENERATIVE LOSS	Regularizes weights toward those of the most recent task	Regularization-Based	$L_{\text{periodic}} = \lambda \sum_i \|\theta_i - \theta_{i-1}\|_2^2$
ELASTIC WEIGHT CONSOLIDATION (EWC) (Kirkpatrick et al., 2016)	Selectively constrains updates based on Fisher information	Regularization-Based	$L_{\text{EWC}} = \frac{\lambda}{2} \sum_j F_j (\theta_j - \theta_j^*)^2$
PACKNET (Mallya & Lazebnik, 2017)	Allocates separate subnetworks for different tasks through pruning	Network-Based	$\boldsymbol{\theta} \leftarrow \mathbf{m} \odot \boldsymbol{\theta}$

The non-stationarity of reward scales across tasks remains a central obstacle in CRL, leading to instability in policy updates. To address this, we explore rescaling methods that normalize rewards or TD errors, building on PPO's discount-based reward scaling (Engstrom et al., 2020). Specifically, we investigate TASK-AGNOSTIC and TASK-AWARE reward rescaling strategies, as well as scaling TD errors during policy updates (Schaul et al., 2021). These approaches are summarized in Table 2.

TASK-AGNOSTIC rescaling applies a single normalization factor across all tasks, ensuring consistent reward magnitudes and mitigating instability from varying distributions. In contrast, TASK-AWARE rescaling adjusts rewards per task, normalizing based on task-specific statistics to improve plasticity in non-stationary environments. Building on these approaches, we explored TD error scaling, which

Table 2: Comparison of Rescaling Methods: Reward and TD error are denoted by r and δ, and time step and task ID by t and i, respectively. σ_R and $\sigma_{R,i}$ represent the global and per-task standard deviations of rewards, while σ_δ and $\sigma_{\delta,i}$ represent the standard deviations of TD errors similarly.

Method	**TASK-AGNOSTIC**	**TASK-AWARE**
REWARD RESCALING	$r'_t = \frac{r_t}{\sigma_R}$	$r'_{t,i} = \frac{r_{t,i}}{\sigma_{R,i}}$
TD ERROR SCALING	$\delta'_t = \frac{\delta_t}{\sigma_\delta}$	$\delta'_{t,i} = \frac{\delta_{t,i}}{\sigma_{\delta,i}}$

normalizes policy update signals rather than rewards, ensuring stable learning without requiring explicit task knowledge (Schaul et al., 2021). Here, TASK-AGNOSTIC TD SCALING applies a fixed normalization factor across all tasks, while TASK-AWARE TD SCALING reinitializes per task to better adapt to varying dynamics.

To evaluate the effects and limitations of the above methods, we follow the evaluation protocol outlined in Section 2.2.1, conducting experiments with a fixed task order across six predefined seeds to control the pseudorandom nature of the environments. We discuss the benchmarking results in the following.

4.2 Benchmarking Results

In this section, we analyze the performance of existing methods with respect to their plasticity-stability trade-off. The results, summarized in Table 3, highlight the distinct strengths and limitations of different approaches across various tasks.

Table 3: Benchmark of Existing Methods: Online evaluation as median ± IQR; Till-Now (AUC) evaluation as mean ± std after normalization. Raw and total game scores are shown.

Method	**Metrics**	**Breakout**	**Freeway**	**Asterix**	**Seaquest**	**SpaceInvaders**	**Raw Total**	**Normalized Total**
Baseline for Comparison								
ORACLE PPO	ONLINE	12.71±0.61	43.84±1.00	4.14±0.05	12.69±2.14	100.15±10.90	173.54±14.41	5.00±0.36
	AUC	11.64±2.14	21.22±14.84	1.30±0.91	1.06±0.69	3.75±2.81	38.97±16.19	5.00±2.44
NAÏVE PPO	ONLINE	12.71±0.61	32.32±0.96	2.07±0.10	2.33±0.79	17.93±3.35	67.36±5.62	2.60±0.19
	AUC	5.50±2.12	6.87±3.08	0.70±0.28	0.17±0.25	1.69±1.46	14.93±11.04	1.95±0.79
Rescaling-Based Methods								
TASK-AGNOSTIC TD SCALING	ONLINE	10.57±1.89	31.62±1.43	3.18±0.85	4.53±0.88	47.01±23.51	96.91±28.84	3.15±0.69
	AUC	4.35±3.39	9.58±7.37	0.83±0.52	0.40±0.29	2.16±1.20	17.32±10.30	2.42±1.08
TASK-AWARE TD SCALING	ONLINE	11.93±2.75	**36.19±3.70**	2.79±0.66	3.66±0.30	20.61±10.30	75.17±18.24	2.93±0.61
	AUC	4.78±3.91	8.34±5.47	0.77±0.39	0.31±0.22	1.98±1.47	16.18±11.29	2.22±0.89
TASK-AGNOSTIC RESCALING	ONLINE	12.09±3.44	27.97±0.72	1.96±0.15	3.88±0.98	47.68±19.03	93.59±24.51	2.84±0.59
	AUC	5.45±4.06	7.32±6.56	0.72±0.30	0.32±0.23	2.23±1.15	16.04±8.08	2.26±0.90
TASK-AWARE RESCALING	ONLINE	12.23±2.93	36.04±2.93	2.45±0.22	4.81±1.38	**49.39±24.70**	**104.93±32.44**	3.25±0.71
	AUC	4.98±3.87	7.97±5.49	0.77±0.38	0.35±0.29	**2.66±1.99**	16.73±11.59	**2.44±1.14**
Plasticity-Oriented Methods								
L2 REGULARIZATION	ONLINE	9.85±1.73	34.73±2.70	**5.45±0.84**	6.18±0.40	34.40±3.46	90.61±9.14	3.71±0.47
	AUC	4.09±3.17	8.68±6.25	0.93±0.78	0.44±0.38	1.69±1.04	15.83±10.98	2.34±1.23
REGENERATIVE LOSS	ONLINE	12.68±0.66	29.91±1.08	2.43±0.29	6.04±1.69	3.50±1.75	54.54±5.34	2.78±0.30
	AUC	5.52±4.25	8.28±4.12	0.73±0.35	**0.45±0.36**	1.43±0.74	16.41±9.34	2.23±0.93
CRELU	ONLINE	11.71±2.62	34.89±2.83	**11.25±3.28**	2.26±1.13	26.04±13.02	86.15±23.62	**4.87±1.28**
	AUC	4.57±3.85	9.15±8.13	**1.45±1.85**	0.15±0.12	1.29±0.81	16.61±10.21	2.42±1.61
WEIGHTCLIPPING	ONLINE	10.45±2.09	29.43±0.75	2.08±0.18	2.59±0.33	26.17±13.09	70.71±16.49	2.46±0.38
	AUC	4.50±3.47	6.16±584	0.72±0.32	0.22±0.19	1.93±1.65	13.53±9.65	1.95±0.78
PARSEVAL REGULARIZATION	ONLINE	12.52±2.91	29.34±1.15	2.07±0.33	5.02±1.59	25.50±12.75	74.46±18.93	2.80±0.59
	AUC	5.09±4.18	**10.59±7.26**	0.72±0.33	0.41±0.30	1.57±0.96	**18.38±8.62**	2.30±0.95
Stability-Oriented Methods								
PERIODIC REGENERATIVE LOSS	ONLINE	12.68±0.66	30.49±1.15	3.03±0.66	**7.83±2.38**	10.32±5.16	64.34±10.03	3.15±0.48
	AUC	5.35±4.34	8.29±6.81	0.82±0.48	0.34±0.29	1.38±1.02	16.18±9.90	2.17±0.92
EWC	ONLINE	**12.71±0.61**	28.03±0.37	1.63±0.14	3.61±1.21	22.32±1.57	68.30±3.71	2.54±0.20
	AUC	**6.69±3.90**	7.25±4.6	0.68±0.25	0.29±0.23	1.75±1.63	16.66±10.14	2.18±0.85
PACKNET	ONLINE	5.75±0.38	13.75±6.88	1.74±0.13	3.84±0.47	14.05±5.52	39.13±13.47	1.63±0.31
	AUC	4.74±3.88	3.81±2.28	0.68±0.24	0.41±0.24	1.67±1.22	11.31±6.14	1.94±0.77

Rescaling-Based Methods Mitigate Non-Stationarity but Prioritize Plasticity Over Stability Reward rescaling methods mitigate the issue of non-stationary reward scales by dynamically normalizing reward distributions, which may aid with both plasticity and stability. In our results, TASK-AWARE RESCALING is particularly effective in enhancing plasticity, as it adapts quickly to varying reward scales and facilitates rapid policy updates. This advantage is reflected in its superior performance across highly adaptive tasks. However, despite its good plasticity, its stability remains only moderate, as its dynamic scaling does not fully prevent long-term forgetting. Conversely, TASK-AGNOSTIC RESCALING enforces more consistent reward normalization across tasks, which improves retention in games like Breakout, but comes at the cost of reduced plasticity, leading to lower overall AUC scores. These results suggest that while rescaling methods effectively address non-stationarity—tackling a major cause of both plasticity loss and forgetting—they primarily enhance plasticity, with only partial success in stabilizing long-term learning.

Plasticity-Oriented Methods Prioritize Adaptation but Compromise Knowledge Retention Among plasticity-oriented approaches, L2 REGULARIZATION is the most effective, achieving a high Normalized Total Reward with a balanced trade-off across tasks. Unlike CRELU, which excels in a single environment due to its extreme plasticity, L2 REGULARIZATION maintains plasticity while reducing weight drift, ensuring consistent performance. REGENERATIVE LOSS and PARSEVAL REGULARIZATION perform slightly worse, enhancing plasticity but lacking stability. WEIGHT CLIPPING, with stricter constraints, prevents catastrophic forgetting but hinders learning. While plasticity-oriented methods enable rapid adaptation, they do so at the cost of long-term retention. CRELU exemplifies this trade-off, achieving the highest overall reward but relying heavily on a single-task advantage rather than balanced learning. Among these methods, only L2 REGULARIZATION mitigates the instability of emphasizing plasticity, yet no approach fully resolves the inherent challenge of continual adaptation in RL.

Stability-Oriented Methods Struggle with Limited Learning Flexibility Stability-oriented approaches effectively prevent catastrophic forgetting but often impose rigid constraints that hinder adaptation. EWC maintains strong retention but relies on weight consolidation, restricting flexibility and leading to poor performance in adaptation-heavy tasks. PERIODIC REGENERATIVE LOSS, while slightly more plastic, constrains learning dynamics, preventing optimal adaptation across diverse environments. Both methods perform competitively in the initial task but fail to balance stability with sufficient plasticity, resulting in declining performance over time, particularly in later tasks where plasticity is crucial. PACKNET suffers the most due to its aggressive weight pruning, which severely limits future learning. While it preserves early knowledge, its near-complete freezing of task representations leaves little capacity for new information, leading to catastrophic underperformance in later environments. This underscores a fundamental limitation of stability-oriented methods in continual reinforcement learning: their emphasis on retention comes at the cost of long-term plasticity.

Our evaluation of existing methods designed to enhance plasticity or prevent forgetting reveals that they predominantly prioritize one objective at the expense of the other. Rescaling-based approaches address non-stationarity, improving plasticity while offering only moderate stability. Plasticity-oriented methods enable rapid adaptation but suffer from forgetting, whereas stability-oriented methods aim to mitigate forgetting yet are constrained in terms of their learning flexibility. These findings underscore the inherent difficulty of achieving a well-balanced trade-off between plasticity and stability in CRL.

5 REMARL2: Balancing the Trade-off with Rehearsal and Marginal L2

Building on key insights from existing CRL methods, we introduce REMARL2, a novel method that integrates REHEARSAL REGULARIZATION and MARGINAL L2 to improve the balance between plasticity and stability. By addressing specific limitations of prior approaches, our method enhances the agent's ability to learn new tasks while retaining previously acquired knowledge.

To encourage plasticity with less constrain, MARGINAL L2 combines L2 REGULARIZATION and WEIGHT CLIPPING to impose a controlled constraint on parameter updates. Instead of hard clipping, which can introduce instability, this component selectively applies L2 regularization to weights exceeding a computed threshold, encouraging gradual shrinkage toward the boundary. At time step t, the regularization term for the current network θ_t is formally defined as:

$$\mathcal{L}_{\text{MarginalL2}} = \sum_d \mathbb{I}(|\theta_{t,d}| > c) \cdot \theta_{t,d}^2 \tag{3}$$

where $\theta_{t,d}$ denotes the d-th dimension of the parameter vector θ_t and $\mathbb{I}(|\theta_{t,d}| > c)$ is an indicator function that activates the L2 penalty only for parameters exceeding the threshold c. This approach mitigates the over-constraining of parameter learning by setting a free margin, while preventing uncontrolled weight divergence by imposing the L2 regularization outside the margin threshold.

To further mitigate forgetting, we incorporate a rehearsal-based component that maintains a buffer of historical observations and policy distributions from previous tasks. During training on a new task, a subset of past observations is sampled, and the policy is regularized via behavior cloning to minimize the discrepancy between the current and stored action distributions, which is formulated as a Rehearsal loss. Let $\mathcal{B}_i$ denote the rehearsal buffer at task i, containing samples of past observations and policies. The Rehearsal loss regarding task i for the current parameter θ_t at time step t is computed as:

$$\mathcal{L}_{\text{Rehearsal}} = \mathbb{E}_{s \sim \mathcal{B}_{i-1}} \left[\text{CE} \left(\pi_{i-1}(a|s), \pi_{\theta_t}(a|s) \right) \right] \tag{4}$$

where CE is the cross-entropy loss between the action distributions $\pi_{i-1}(a|s)$ and $\pi_{\theta_t}(a|s)$ due to the discrete action space in MinAtar, and s represents a state sample from the rehearsal buffer $\mathcal{B}_{i-1}$. Unlike periodic regenerative loss, which constrains weights toward previous values, this approach directly preserves decision-making behavior by aligning policy outputs across tasks. To ensure efficient memory utilization, a fixed-size rehearsal buffer is maintained, with samples selected uniformly across all encountered tasks. By incorporating historical decision knowledge, this component enables the agent to retain prior skills while continuously adapting to new environments.

Finally, REMARL2 extends the conventional RL training loss functions by optimizing the loss functions for the two components concurrently. The final loss function is computed as:

$$\mathcal{L} = \mathcal{L}_\pi + \lambda_1 \mathcal{L}_Q + \lambda_2 \mathcal{L}_{\text{MarginalL2}} + \lambda_3 \mathcal{L}_{\text{Rehearsal}} \tag{5}$$

where $\mathcal{L}_\pi$ is the PPO policy objective in our context, $\mathcal{L}_Q$ is the value function loss, and $\lambda_1, \lambda_2, \lambda_3$ are weighting coefficients that balance the contribution of each loss term. The update procedure is as follows:

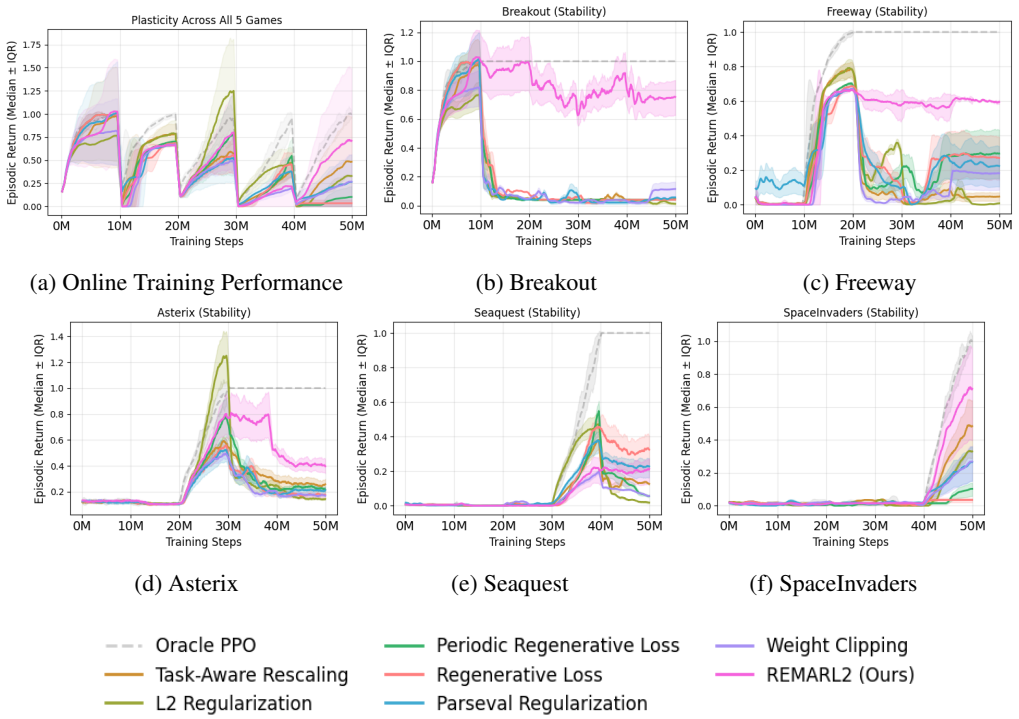
- At the end of each task i, save a batch of observations $\{s_j\}$ and corresponding action distributions $\pi_i(a|s_j)$ in $\mathcal{B}_i$.
- For each new task i, sample a subset of past observations and action distributions from $\mathcal{B}_{i-1}$.
- The agent trains using behavior cloning to align current and past policies (Rehearsal loss) while applying Marginal L2 Regularization to constrain excessive weight updates and maintain adaptability (Marginal L2 penalty).

The formalized algorithm is described in Appendix C.

5.1 Evaluation & Comparison

Our findings demonstrate that REMARL2 achieves a superior balance between plasticity and stability, as evidenced by its strong performance in both Online Evaluation and Till-Now AUC.

REMARL2 Strikes an Effective Balance Between Plasticity and Stability As shown in Figure 1 and Table 4, REMARL2 achieves the highest Till-Now AUC along with a strong Online score, highlighting its exceptional ability to retain past knowledge while adapting to new tasks. This balance



(a) Online Training Performance (b) Breakout (c) Freeway

(d) Asterix (e) Seaquest (f) SpaceInvaders

Figure 3: Performance of REMARL2 as normalized episodic return (median) across six seeds, plotted with a running window size of 10, with shaded regions showing interquartile range (IQR).

between stability and plasticity is further evidenced by its superior Normalized Total Reward, which outperforms other methods, underscoring its competitive plasticity across tasks. While CRELU excels in a single game, Asterix, with a high Online Evaluation score, its overall performance remains limited. As a result, we exclude CRELU from further comparisons and plotting due to its skewed behavior. On the other hand, REMARL2 demonstrates consistent and strong performance across multiple games, as evidenced by its highest Till-Now AUC, reflecting its ability to sustain plasticity while maintaining long-term retention. Additionally, rescaling-based strategies such as TASK-AWARE Rescaling exhibit high plasticity but are hindered by instability, as indicated by their lower AUC scores. REMARL2 mitigates this issue by combining selective regularization with efficient rehearsal, ensuring robust stability without sacrificing plasticity, particularly in the first three games (subfigures 3b to 3d), where it demonstrates strong performance with minimal forgetting.

Table 4: Comparisons with Existing Methods: Online evaluation as median ± IQR; Till-Now (AUC) evaluation as mean ± std after normalization.

Method	Metrics	Breakout	Freeway	Asterix	Seaquest	SpaceInvaders	Raw Total	Normalized Total
Baseline for Comparison								
NAÏVE PPO	ONLINE	12.71±0.61	32.32±0.96	2.07±0.10	2.33±0.79	17.93±3.35	67.36±5.62	2.60±0.19
	AUC	5.50±2.12	6.87±3.08	0.70±0.28	0.17±0.25	1.69±1.46	14.93±11.04	1.95±0.79
Selected Existing Methods for Comparison								
TASK-AWARE RESCALING	ONLINE	12.23±2.93	**36.04±2.93**	2.45±0.22	4.81±1.38	49.39±24.70	104.93±32.44	3.25±0.71
	AUC	4.98±3.87	7.97±5.49	0.77±0.38	0.35±0.29	2.66±1.99	16.73±11.59	0.87±0.70
L2 REGULARIZATION	ONLINE	9.85±1.73	34.73±2.70	5.45±0.84	6.18±0.40	34.40±3.46	90.61±9.14	3.71±0.47
	AUC	4.09±3.17	8.68±6.25	0.93±0.78	**0.44±0.38**	1.69±1.04	15.83±10.98	2.34±1.23
PERIODIC REGENERATIVE LOSS	ONLINE	12.68±0.66	30.49±1.15	3.03±0.66	**7.83±2.38**	10.32±5.16	64.34±10.03	3.15±0.48
	AUC	5.35±4.34	8.29±6.81	0.82±0.48	0.34±0.29	1.38±1.02	16.18±9.90	2.17±0.92
Our Method								
REMARL2	ONLINE	**13.05±3.13**	29.16±0.85	3.13±0.63	2.47±1.23	**73.30±36.65**	**121.11±42.93**	**3.37±0.88**
	AUC	**10.33±2.14**	**12.83±9.86**	**1.01±0.64**	0.27±0.20	**2.83±2.06**	**27.27±10.93**	**3.28±1.36**

REMARL2 Mitigates Rigid Constraints of Existing Regularization Methods Traditional regularization techniques, such as L2 REGULARIZATION and WEIGHT CLIPPING, are primarily designed to ensure stability by constraining weight updates, but this often limits adaptability in dynamic environments. While L2 REGULARIZATION maintains strong retention, its rigid constraints hinder the ability to adapt in environments requiring strategic shifts, as observed in games like Space Invaders. Similarly, WEIGHT CLIPPING enforces hard parameter bounds, reducing flexibility and impeding the ability to learn effectively in more complex scenarios. In contrast, REMARL2 addresses these challenges by selectively applying regularization, thereby preventing parameter divergence without sacrificing the adaptability needed for efficient learning in dynamic environments.

5.2 Ablation Study

To understand the individual contributions of the components in our proposed method, we conduct an ablation study comparing REHEARSAL REGULARIZATION, MARGINAL L2, and the combined REMARL2. The results, summarized in Table 5, reveal key insights into the effectiveness of each component in balancing plasticity and stability.

Table 5: Ablation results of our methods: Online evaluation as median ± IQR; Till-Now (AUC) evaluation as mean ± std after normalization.

Method	Metrics	Breakout	Freeway	Asterix	Seaquest	SpaceInvaders	Raw Total	Normalized Total
REHEARSAL REGULARIZATION	ONLINE	11.96±2.84	32.30±1.88	2.58±0.61	3.05±1.53	40.50±5.19	90.39±12.24	2.95±0.59
	AUC	10.04±1.87	**14.18±8.46**	0.85±0.44	0.25±0.17	2.16±1.20	**27.48±10.91**	3.01±0.95
MARGINAL L2	ONLINE	10.17±1.70	**32.95±2.15**	2.09±0.33	**6.15±1.95**	42.56±17.88	93.91±24.20	2.97±0.59
	AUC	4.47±3.50	8.40±6.62	0.71±0.31	**0.37±0.23**	2.14±1.21	16.09±10.54	2.25±1.05
REMARL2	ONLINE	**13.05±3.13**	29.16±0.85	**3.13±0.63**	2.47±1.23	**73.30±36.65**	**121.11±42.93**	**3.37±0.88**
	AUC	**10.33±2.14**	12.83±9.86	**1.01±0.64**	0.27±0.20	**2.83±2.06**	27.27±10.93	**3.28±1.36**

The ablation study reveals that REHEARSAL REGULARIZATION excels at retaining knowledge, particularly evident in its strong AUC performance in Breakout and Freeway, demonstrating its ability to mitigate catastrophic forgetting. However, its performance in more plasticity-demanding tasks, such as Seaquest and Space Invaders, is less robust, suggesting that while rehearsal methods are effective for stability, they struggle in environments requiring substantial plasticity. On the other hand, MARGINAL L2 enhances plasticity by selectively applying L2 regularization, which prevents uncontrolled parameter divergence while promoting adaptability. While it shows strong performance in tasks like Freeway and Space Invaders, it fails to fully address catastrophic forgetting, as seen in its suboptimal performance in Breakout.

The ablation study demonstrates the complementary strengths of REHEARSAL REGULARIZATION and MARGINAL L2, which, when combined in REMARL2, deliver superior performance across tasks, particularly excelling in Space Invaders. MARGINAL L2 enhances plasticity by preventing excessive weight growth, while REHEARSAL REGULARIZATION preserves stability through behavior cloning. By integrating selective weight regularization with rehearsal-based learning, REMARL2 successfully balances these two critical aspects, overcoming the individual limitations of each component. This synergistic approach ensures robust knowledge retention while enabling effective adaptation to new tasks, offering a more comprehensive solution for continual reinforcement learning than either method in isolation. The findings underscore the importance of integrating multiple strategies to address the complex challenges of continual learning.

6 Related Work

Various recent studies have examined the balance between stability and plasticity in CRL. For instance, Lyle et al. (2023) identify a "primacy bias" in deep reinforcement learning, where agents overfit to early experiences, hindering subsequent learning. They propose a resetting mechanism to address this bias; however, their analysis primarily focuses on the impact of loss landscape curvature on plasticity, potentially overlooking other contributing factors. Expanding on stability-plasticity

trade-offs, van de Ven et al. (2024) review strategies to mitigate catastrophic forgetting in neural networks, while Kong et al. (2022) introduce Advanced Null Space (AdNS) to prevent task interference through gradient projection, and Liu et al. (2024) propose LRFR, leveraging low-rank feature representation for improved stability and plasticity. Although these approaches provide valuable insights, they primarily focus on supervised learning and overlook reinforcement learning's unique challenges, such as non-stationary rewards and evolving state distributions. Further highlighting the complexities of plasticity in RL, Abbas et al. (2023) investigate how prolonged training can erode learning capabilities in deep RL agents. However, their work primarily focuses on architectural modifications and does not explore strategies that explicitly balance stability and plasticity during learning, leaving open questions about how to effectively mitigate plasticity loss while retaining past knowledge.

While these studies provide important contributions, they often overlook the stability-plasticity trade-off in structured RL environments like MinAtar, which strikes a balance between complexity and computational efficiency (Ceron & Castro, 2021; Gogianu et al., 2021). Games such as Freeway and Seaquest within MinAtar serve as key benchmarks for evaluating exploration capabilities and knowledge retention (Ceron & Castro, 2021). This work systematically assesses the stability-plasticity trade-off in continual RL, addressing gaps in prior evaluations by exploring underexamined aspects and leveraging MinAtar's structured yet computationally efficient environment to provide insights applicable to more complex RL settings.

7 Conclusion

In this study, we systematically evaluated the stability-plasticity trade-off in CRL through a sequence of MinAtar games. Our findings reveal that no method consistently performs the best across all scenarios. Approaches prioritizing plasticity, such as reward rescaling, excel at adapting to new tasks but perform moderately with long-term stability. In contrast, stability-oriented methods, like L2 REGULARIZATION, aimed to mitigate forgetting but hinder plasticity. REMARL2 addresses these challenges by combining selective weight regularization with rehearsal-based learning, achieving a strong balance between plasticity and stability. Empirical results demonstrate that REMARL2 effectively balances Till-Now AUC and Online evaluation scores, preserving knowledge while acquiring new skills. By ensuring continual adaptation without excessive forgetting, REMARL2 establishes itself as a robust solution for CRL.

Limitations and Future Work This study has several limitations that offer opportunities for future research. First, our experiments were conducted exclusively in the MinAtar environment, limiting generalizability; extending this work to larger-scale environments like Meta-World (Yu et al., 2019) would provide a more robust assessment. Second, our evaluation focused solely on PPO, leaving open how REMARL2 performs with other algorithms. Expanding benchmarking to include diverse approaches would deepen insights into the stability-plasticity trade-off. Finally, the fixed task order in our experiments may impact performance; future work could explore task sequencing effects in environments with varying difficulty and reward scales.

Building on these limitations, future research could extend this approach to larger-scale environments, where balancing stability and plasticity becomes more challenging. Optimizing memory efficiency in rehearsal buffers is also critical for scaling to complex tasks, as storage requirements can grow significantly. Additionally, investigating hybrid models that combine REMARL2 with other advanced reinforcement learning techniques could provide deeper insights into handling a wider range of environments.

References

Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C Machado. Loss of plasticity in continual deep reinforcement learning. In *Conference on lifelong learning agents*, pp. 620–636. PMLR, 2023.

Matthew Aitchison, Penny Sweetser, and Marcus Hutter. Atari-5: Distilling the arcade learning environment down to five games. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 421–438. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/aitchison23a.html.

Glen Berseth, Cheng Xie, Paul Cernek, and Michiel Van de Panne. Progressive reinforcement learning with distillation for multi-skilled motion control. In *International Conference on Learning Representations*, 2018.

Johan Samir Obando Ceron and Pablo Samuel Castro. Revisiting rainbow: Promoting more insightful and inclusive deep reinforcement learning research. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 1373–1383. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/ceron21a.html.

Wesley Chung, Lynn Cherif, Doina Precup, and David Meger. Parseval regularization for continual reinforcement learning. *Advances in Neural Information Processing Systems*, 37:127937–127967, 2025.

Michael Dann and John Thangarajah. Adapting to reward progressivity via spectral reinforcement learning. *arXiv preprint*, arXiv:2104.14138, 2021. URL https://arxiv.org/abs/2104.14138.

Mohamed Elsayed, Qingfeng Lan, Clare Lyle, and A. Rupam Mahmood. Weight clipping for deep continual and reinforcement learning. *arXiv preprint*, arXiv:2407.01704, 2024. URL https://api.semanticscholar.org/CorpusID:270878729.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep policy gradients: A case study on PPO and TRPO. *arXiv preprint*, arXiv:2005.12729, 2020. URL https://arxiv.org/abs/2005.12729.

Florin Gogianu, Tudor Berariu, Mihaela C Rosca, Claudia Clopath, Lucian Busoniu, and Razvan Pascanu. Spectral normalisation for deep reinforcement learning: An optimisation perspective. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3734–3744. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/gogianu21a.html.

Muhammad Burhan Hafez and Kerim Erekmen. Continual deep reinforcement learning with task-agnostic policy distillation. *Scientific Reports*, 14(1), December 2024. ISSN 2045-2322. DOI: 10.1038/s41598-024-80774-8. URL http://dx.doi.org/10.1038/s41598-024-80774-8.

Tyler L Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. *arXiv preprint*, September 2017.

Maximilian Igl, Gregory Farquhar, Jelena Luketina, Wendelin Boehmer, and Shimon Whiteson. The impact of non-stationarity on generalisation in deep reinforcement learning. *arXiv preprint*, arXiv:2006.05826, 2020. URL https://arxiv.org/abs/2006.05826.

Arthur Juliani and Jordan T. Ash. A study of plasticity loss in on-policy deep reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=MsUf8kpKTF.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2020. URL https://api.semanticscholar.org/CorpusID:229679944.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476, 2022. URL https://doi.org/10.1613/jair.1.13673.

James Kirkpatrick, Razvan Pascanu, Neil C. 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. *arXiv preprint*, arXiv:1612.00796, 2016. URL http://arxiv.org/abs/1612.00796.

Yajing Kong, Liu Liu, Zhen Wang, and Dacheng Tao. Balancing stability and plasticity through advanced null space in continual learning. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 219–236, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19809-0.

Saurabh Kumar, Henrik Marklund, and Benjamin Van Roy. Maintaining plasticity in continual learning via regenerative regularization, 2024. URL https://arxiv.org/abs/2308.11958.

Zhenrong Liu, Yang Li, Yi Gong, and Yik-Chung Wu. Learning a low-rank feature representation: Achieving better trade-off between stability and plasticity in continual learning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5885–5889, 2024. DOI: 10.1109/ICASSP48485.2024.10446458.

Clare Lyle, Zeyu Zheng, Evgenii Nikishin, Bernardo Avila Pires, Razvan Pascanu, and Will Dabney. Understanding plasticity in neural networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23190–23211. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/lyle23b.html.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. *arXiv preprint*, arXiv:1711.05769, 2017. URL http://arxiv.org/abs/1711.05769.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint*, arXiv:1312.5602, 2013. URL http://arxiv.org/abs/1312.5602.

Anusha Nagabandi, Chelsea Finn, and Sergey Levine. Deep online learning via meta-learning: Continual adaptation for model-based RL. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyxAfnA5tm.

Sindhu Padakandla, Prabuchandran K. J., and Shalabh Bhatnagar. Reinforcement learning in non-stationary environments. *arXiv preprint*, arXiv:1905.03970, 2019. URL http://arxiv.org/abs/1905.03970.

Mark B Ring. Child: A first step towards continual learning. *Machine Learning*, 28(1):77–104, 1997.

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. URL http://arxiv.org/abs/1606.04671.

Tom Schaul, Georg Ostrovski, Iurii Kemaev, and Diana Borsa. Return-based scaling: Yet another normalisation trick for deep RL. *arXiv preprint*, arXiv:2105.05347, 2021. URL https://arxiv.org/abs/2105.05347.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint*, arXiv:1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4528–4537. PMLR, 2018a.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress amp; compress: A scalable framework for continual learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4528–4537. PMLR, 10–15 Jul 2018b. URL https://proceedings.mlr.press/v80/schwarz18a.html.

Freek Stulp. Adaptive exploration for continual reinforcement learning. In *IROS*, pp. 1631–1636. IEEE, 2012. URL https://doi.org/10.1109/IROS.2012.6385818.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL http://incompleteideas.net/book/the-book-2nd.html.

Sebastian Thrun. A lifelong learning perspective for mobile robot control. In *IROS*, pp. 23–30, 1994. URL https://doi.org/10.1109/IROS.1994.407413.

Gido M. van de Ven, Nicholas Soures, and Dhireesha Kudithipudi. Continual learning and catastrophic forgetting. *arXiv preprint*, arXiv:2403.05175, 2024. URL https://doi.org/10.48550/arXiv.2403.05175.

Hado van Hasselt, Arthur Guez, Matteo Hessel, and David Silver. Learning functions across many orders of magnitudes. *arXiv preprint*, arXiv:1602.07714, 2016. URL http://arxiv.org/abs/1602.07714.

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*, 46(8):5362–5383, August 2024.

Jiayi Weng, Huayu Chen, Dong Yan, Kaichao You, Alexis Duburcq, Minghao Zhang, Yi Su, Hang Su, and Jun Zhu. Tianshou: A highly modularized deep reinforcement learning library. *Journal of Machine Learning Research*, 23(267):1–6, 2022. URL http://jmlr.org/papers/v23/21-1127.html.

Annie Xie, James Harrison, and Chelsea Finn. Deep reinforcement learning amidst continual structured non-stationarity. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11393–11403. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/xie21c.html.

K. Young and Tian Tian. Minatar: An atari-inspired testbed for thorough and reproducible reinforcement learning experiments. *arXiv preprint*, 2019. URL https://api.semanticscholar.org/CorpusID:174801095.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. *arXiv preprint*, arXiv:1910.10897, 2019. URL http://arxiv.org/abs/1910.10897.

Tiantian Zhang, Zichuan Lin, Yuxing Wang, Deheng Ye, Qiang Fu, Wei Yang, Xueqian Wang, Bin Liang, Bo Yuan, and Xiu Li. Dynamics-adaptive continual reinforcement learning via progressive contextualization. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):14588–14602, 2024.

A Preliminaries: Proximal Policy Optimization (PPO).

For all of our experiments in this study, we adopt Proximal Policy Optimization (Schulman et al., 2017) as our base algorithm due to its stability and empirical success in on-policy settings. PPO maximizes a clipped surrogate objective:

$$J^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)} \hat{A}_t, \text{clip} \left(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}, 1-\epsilon, 1+\epsilon \right) \hat{A}_t \right) \right], \tag{6}$$

where $\hat{A}_t$ is the advantage estimate and ϵ being the clipping threshold. This formulation ensures policy updates remain bounded, preventing destructive parameter shifts. However, in CRL, the absence of explicit mechanisms to address non-stationarity leads to plasticity loss and catastrophic forgetting. As the agent transitions between tasks, previously learned knowledge is overwritten, resulting in performance degradation. These challenges are further exacerbated by the inherent complexities of sequential decision-making, where variations in task properties and shifting data distributions introduce additional sources of instability.

B Experiment Details

For all the games in MinAtar, we used their v0 versions, where the action space consists of six available actions (some of which may be equivalent to a no-op depending on the game), ensuring a consistent setup for training.

A PPO agent is sequentially trained on each task for 10 million iterations, with evaluations conducted every 10,000 iterations. Training and evaluation are performed using 128 parallel environments, and the mini-batch size is set to 4,096. Further details on hyperparameters and implementation specifics can be found in the appendix. To ensure an unbiased evaluation, test environments are initialized with random seeds different from those used in training.

Performance is assessed throughout training using a standard approach that interleaves training and evaluation blocks (van de Ven et al., 2024). During each evaluation phase, the policy is tested on all encountered tasks, including the current one, for up to 10,000 steps. This ensures proper termination conditions and prevents scenarios where the agent fails to conclude its actions.

C REMARL2 Algorithm

The REMARL2 algorithm extends the original PPO framework by incorporating two key components: **MARGINAL L2** and **REHEARSAL REGULARIZATION**. Below is the detailed algorithm.

Algorithm 1 REMARL2 Algorithm

Initialize policy parameters θ_0, value function parameters ϕ_0
Set MARGINAL L2 hyperparameter threshold c
Set number of tasks K andinitialize the rehearsal buffers $\mathcal{B}_i = \emptyset$ for $i = 1, 2, \ldots, K$
for $i = 1, 2, \ldots, K$ **do** ▷ Outer loop over tasks
  for $t = 0, 1, 2, \ldots$ **do** ▷ Inner loop for training on task i
    // Interact with the environment and collect samples
    Collect trajectories $\mathcal{D}_t$ using policy π_{θ_t}
    Compute rewards-to-go $\hat{R}_t$ and advantages $\hat{A}_t$
    MARGINAL L2:
    Compute regularization term:

$$\mathcal{L}_{\text{MarginalL2}} = \sum_d \mathbb{I}(|\theta_{t,d}| > c) \cdot \theta_{t,d}^2$$

where $\theta_{t,d}$ denotes the d-th dimension of the parameter vector θ_t and $\mathbb{I}(|\theta_{t,d}| > c)$ is an indicator function that activates the L2 penalty only for parameters exceeding the threshold c.

    REHEARSAL REGULARIZATION:
    Sample a batch of $\{s_j, \pi_{i-1}(a|s_j)\}$ from $\mathcal{B}_{i-1}$
    Compute Rehearsal loss:

$$\mathcal{L}_{\text{Rehearsal}} = \mathbb{E}_{s \sim \mathcal{B}_{i-1}} \left[\text{CE} \left(\pi_{i-1}(a|s), \pi_{\theta_t}(a|s) \right) \right]$$

    Update ϕ and θ using:

$$\mathcal{L} = \mathcal{L}_\pi + \lambda_1 \mathcal{L}_Q + \lambda_2 \mathcal{L}_{\text{MarginalL2}} + \lambda_3 \mathcal{L}_{\text{Rehearsal}}$$

where $\mathcal{L}_\pi$ is the PPO-Clip objective, $\mathcal{L}_Q$ is the value function loss, and λ_1, λ_2, λ_3 are coefficients controlling the weight of each loss term.

  end for
  Update Rehearsal Buffer:
  Save a batch of $\{s_j, \pi_i(a|s_j)\}$ in $\mathcal{B}_i$
  If $|\mathcal{B}_i| >$ capacity, remove oldest samples
end for

D Hyperparameter Search Space

In this section, we present the hyperparameter search space used to benchmark both existing and proposed methods, as shown in Table 6, along with the best configurations identified for each. For the remaining settings, we follow the recommendations from the original papers of each method.

Table 6: Hyperparameter Search Space and Optimal Configurations

Method	Parameter	Search Space	Best Value
L2 REGULARIZATION	penalty coefficient λ	{0.1, 0.01, 0.001}	$\lambda = 0.001$
REGENERATIVE LOSS	penalty coefficient λ	{0.1, 0.01, 0.001}	$\lambda = 0.01$
WEIGHT CLIPPING	clipping param κ	{1, 2, 3, 4}	$\kappa = 2.0$
PARSEVAL REGULARIZATION	penalty coefficient λ, Adding diagonal layers, Input scaling	{1, 0.1, 0.01, 0.001}, (True/False), (True/False)	$\lambda = 0.01$, Adding diagonal layers=True, Input scaling=True
PERIODIC REGENERATIVE LOSS	penalty coefficient λ	{0.1, 0.01, 0.001}	$\lambda = 0.01$
EWC	penalty coefficient λ	{1, 0.1, 0.01, 0.001}	$\lambda = 0.1$
MARGINAL L2	penalty coefficient λ, clipping param κ	{1, 0.1, 0.01, 0.001}, {1, 2, 3, 4}	$\lambda = 0.001$, $\kappa = 2.0$
REHEARSAL REGULARIZATION	Rehearsal Buffer sample batch size $B_{\text{rehearsal}}$	{4, 8, 16, 32, 64, 128}	$B_{\text{rehearsal}} = 16$