On the Plasticity Collapse in Continual Machine Unlearning

Yingdan Shi[1], Xiang Xu[2], Kaize Ding[3], Alfred O. Hero[4], and Ren Wang[1⋆]

[1] Illinois Institute of Technology, Chicago IL 60616, USA
[2] Amazon, New York NY 10001, USA
[3] Northwestern University, Evanston IL 60208, USA
[4] University of Michigan, Ann Arbor MI 48109, USA

Abstract. Machine unlearning enables deep neural networks to selectively remove the influence of specific data in response to privacy and regulatory requirements. While prior work largely studies single-shot unlearning, real-world systems must accommodate continual unlearning, where multiple unlearning requests occur sequentially over time. In this work, we identify a fundamental limitation of this setting: plasticity collapse, a progressive breakdown in a model's ability to effectively forget. Through theoretical analysis of continual unlearning dynamics, we show that continual unlearning operations accumulate geometric constraints in parameter space, leading to saturated subspaces that restrict future updates. This structural effect induces two distinct failure modes: (1) Forward failure – diminishing forgetting quality for subsequent tasks, and (2) Backward failure – spontaneous re-memorization of previously forgotten information. Extensive experiments across multiple architectures, datasets, and methods in image classification confirm that plasticity collapse is not an artifact of specific implementations, but a pervasive phenomenon inherent to continual unlearning. Our findings reveal a critical barrier to the long-term reliability of machine unlearning systems and motivate the development of plasticity-preserving unlearning algorithms. Our code is available at `https://github.com/TIML-Group/Continual-Machine-Unlearning-Plasticity-Collapse`.

Keywords: Machine Unlearning · Plasticity · Privacy

1 Introduction

The rapid proliferation of machine learning across sensitive domains has intensified scrutiny over data governance and user privacy. Recent legislative frameworks, including the European Union's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), establish the "right to be forgotten", a legal mandate requiring organizations to delete user data upon request. Beyond regulatory compliance, there is an escalating need to purge harmful or biased information from models to mitigate potential misuses, such as

⋆ Corresponding author: rwang74@illinoistech.edu

the generation of toxic content or the propagation of social biases. For machine learning models, compliance extends beyond removing records from training sets. It requires eliminating information that has been encoded into model parameters during training, a process known as machine unlearning.

Significant progress has been made in developing efficient unlearning methods for single unlearning requests. However, practical deployments face a more complex reality: models must accommodate multiple sequential unlearning[5] requests over their operational lifetime. A content moderation system, for instance, may need to progressively forget different categories of prohibited content as policies evolve. A recommendation system must continuously honor unlearning requests without complete retraining. These scenarios demand that unlearning methods remain effective not just once, but repeatedly and reliably over time.

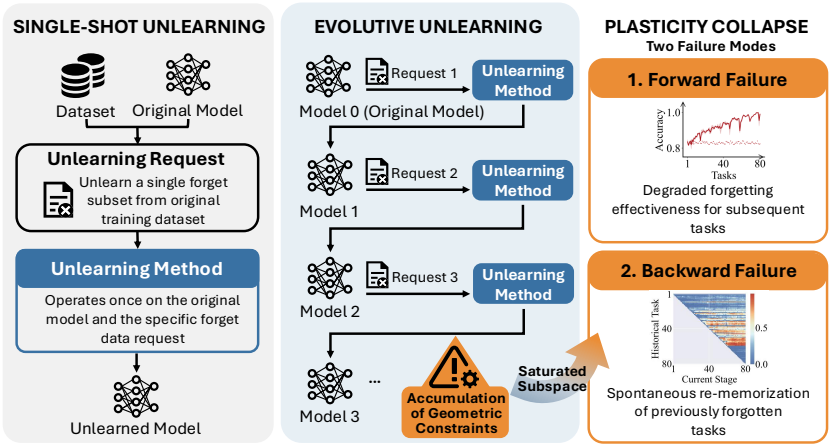


Fig. 1: Comparison between single-shot unlearning and continual unlearning. Unlike the single-shot setting, continual unlearning suffers from the accumulation of geometric constraints, leading to a saturated subspace that induces plasticity collapse. This collapse is characterized by two distinct phenomena: forward failure, which represents a progressive decline in forgetting quality, and backward failure, referring to the spontaneous re-memorization of previously forgotten tasks.

The overview of this work is shown in Fig. 1. In this work, we identify a fundamental limitation of existing unlearning methods in continual settings: models progressively lose their ability to effectively forget, a phenomenon we term ***plasticity collapse***. Through extensive experiments, we identify two distinct failure modes induced by plasticity collapse: (1) **Forward failure**, in which forgetting quality deteriorates progressively for subsequent tasks, and (2) **Backward failure**, in which previously forgotten tasks spontaneously re-emerge. Through theoretical analysis of continual unlearning dynamics, we show that successive unlearning operations accumulate geometric constraints in parameter space, leading to saturated subspaces that restrict future parameter updates. This structural degradation manifests in both failure modes. These findings raise serious

[5] In this work, the terms *continual unlearning* and *sequential unlearning* are used interchangeably.

concerns about the long-term reliability of machine unlearning in the continual unlearning setting. If forgetting quality degrades over sequential requests or earlier forgetting is silently reversed, the privacy guarantees that unlearning methods promise become difficult to sustain. Our contributions are as follows.

- We are the first to systematically characterize plasticity collapse in the continual unlearning. Through extensive experiments spanning diverse architectures, datasets, and unlearning methods in image classification, we demonstrate that this is not an artifact of specific implementations, but a pervasive phenomenon inherent to the continual unlearning setting.
- We provide a theoretical analysis of the underlying mechanism. Our analysis shows that the accumulation of parameter constraints from successive unlearning operations creates increasingly restrictive subspaces, leading to degraded optimization and persistent residual information.
- We design targeted experiments to isolate and validate the mechanisms underlying plasticity collapse, bridging our theoretical predictions with observed empirical behavior.

While our analysis focuses on image classification tasks, the identified mechanisms of subspace saturation and geometric constraints stem from the fundamental optimization dynamics of neural networks. Consequently, this work lays a critical theoretical and diagnostic foundation for understanding plasticity collapse in broader contexts, such as generative modeling and large language models (LLMs), where sequential unlearning is equally pivotal yet more complex.

2 Related Work

2.1 Machine Unlearning

Machine unlearning seeks to remove the influence of specific training data from a learned model without requiring complete retraining. Early approaches framed unlearning as the inverse of learning, applying gradient ascent on the data to be forgotten [4, 10]. However, naive gradient ascent often causes catastrophic degradation of model utility on retained data.

Exact unlearning methods offer certified guarantees by partitioning training data and maintaining sub-models that can be efficiently retrained upon deletion [3, 13, 21]. SISA training [3], for example, divides the dataset into shards and trains separate models on each, allowing selective retraining when data is removed. While these approaches provide strong guarantees, they incur substantial storage and computational overhead, limiting their scalability, particularly in continual unlearning settings.

Approximate unlearning methods [8, 12, 17, 22, 23, 26, 27] trade formal guarantees for efficiency, offering practical alternatives through diverse techniques including fine-tuning, parameter perturbation, and selective gradient manipulation. These methods are generally more applicable to real-world deployment but have been studied almost exclusively in single-shot unlearning settings.

2.2 Continual Unlearning

A line of work addresses scenarios where a model alternates between continual learning and selective forgetting [1,5,14,25]. In this mixed setting, the model continues to receive new training data alongside unlearning requests, which helps preserve model utility throughout the process.

More closely related to our work, [9] studied incremental unlearning, where the goal is to achieve thorough forgetting across sequential unlearning tasks on a fixed model, without interleaving new learning. That is, the model must completely remove targeted knowledge while retaining the rest, purely through successive forgetting operations. This purely sequential unlearning setting is the focus of our work. We go further by identifying a fundamental limitation of this setting, namely plasticity collapse, and providing both theoretical and empirical characterization of this failure mode.

2.3 Neural Network Plasticity

Loss of plasticity has been identified as a fundamental challenge in deep learning, referring to the progressive decline in a network's capacity to learn new information over time [6,16,19]. [19] showed that neural networks in deep reinforcement learning lose learning capacity even when the underlying task remains stationary. Proposed mechanisms include dormant neurons [24], implicit rank collapse [15], and gradient interference [11]. [16] further demonstrated that continual learning agents suffer similar rigidity, attributing the effect to feature reuse constraints and conflicting gradient signals. Remedies include weight resetting [6], optimization adaptations [20], and the shrink-and-perturb strategy [2,7], which combines parameter pruning with controlled perturbation.

Despite this progress, plasticity in the context of machine unlearning remains unexplored. The mechanisms governing learning plasticity do not transfer directly to unlearning, since the optimization objectives and parameter-space constraints differ fundamentally. Our work is the first to characterize plasticity collapse specifically within the unlearning context, establishing it as a distinct and pervasive phenomenon with its own theoretical grounding.

3 Plasticity Collapse in Continual Unlearning

3.1 Preliminaries and Notation

Continual unlearning setup. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ denote the full training dataset used to train an initial model with parameters $\boldsymbol{\theta}_0 \in \mathbb{R}^d$. We consider a sequence of T unlearning requests arriving one at a time. At each stage $t \in \{1, \ldots, T\}$, a subset is designated as the forget set $F_t \subset \mathcal{D}$, containing samples whose influence must be removed from the model. An optional retain set $R_t \subseteq \mathcal{D} \setminus F_t$ is provided to preserve model utility on non-forgotten data. After processing request t, the model parameters are updated from $\boldsymbol{\theta}_{t-1}$ to $\boldsymbol{\theta}_t$.

Unlearning objective. For any dataset $S = \{(x_i, y_i)\}_{i=1}^{n_S}$, let $\ell(\boldsymbol{\theta}; x, y)$ denote the per-sample loss (e.g., cross-entropy). We define the empirical loss over S as

$$\mathcal{L}_S(\boldsymbol{\theta}) \;:=\; \frac{1}{n_S} \sum_{i=1}^{n_S} \ell(\boldsymbol{\theta};\, x_i,\, y_i). \tag{1}$$

Most unlearning methods share a common two-component structure: a **forget loss** $\mathcal{L}_{F_t}(\boldsymbol{\theta})$ that drives the model away from predictions on F_t, and a **retain loss** $\mathcal{L}_{R_t}(\boldsymbol{\theta})$ that anchors model behavior on R_t. Methods differ primarily in how each loss function is designed. For example, NegGrad+ [17] additionally minimizes the retain loss, RandomLabeling [12] replaces ground-truth labels in F_t with random ones, and SalUn [8] further restricts updates to salient parameters. Despite these differences, in the local geometry of parameter space, the parameter update induced by any such method has a positive projection onto the gradient ascent direction of $\mathcal{L}_{F_t}$, reflecting a shared inductive bias toward reducing the model's fit on F_t. Thus, all such methods fit within the unified objective

$$\mathcal{J}_t(\boldsymbol{\theta}) \;:=\; \lambda\, \mathcal{L}_{R_t}(\boldsymbol{\theta}) \;-\; \mathcal{L}_{F_t}(\boldsymbol{\theta}), \qquad \lambda \geq 0, \tag{2}$$

where λ controls the balance between forgetting and utility preservation. One step of gradient descent on $\mathcal{J}_t$ gives

$$\boldsymbol{\theta}_t \;=\; \boldsymbol{\theta}_{t-1} - \eta_t \nabla_{\boldsymbol{\theta}} \mathcal{J}_t(\boldsymbol{\theta}_{t-1}) \;=\; \boldsymbol{\theta}_{t-1} - \eta_t \Big(\lambda\, \nabla \mathcal{L}_{R_t}(\boldsymbol{\theta}_{t-1}) - \nabla \mathcal{L}_{F_t}(\boldsymbol{\theta}_{t-1}) \Big), \tag{3}$$

where $\eta_t > 0$ is the step size at stage t. For tractability, we analyze the single-step case in Eq. 3, but extending to multiple steps per task does not affect our theoretical conclusions. Our theoretical analysis is developed within this general framework and does not presuppose any particular design of $\mathcal{L}_{F_t}$ or $\mathcal{L}_{R_t}$. The plasticity collapse we characterize is therefore a structural consequence of sequentially applying any training-based unlearning method.

3.2 Definition of Plasticity Collapse

Through extensive experiments across diverse datasets and architectures, we identify two distinct failure modes inherent in the continual unlearning setting. Prior to our theoretical analysis, we formally define these modes, which we attribute to the plasticity collapse that occurs during sequential unlearning.

Definition 1 (Plasticity collapse) *Let $\{\boldsymbol{\theta}_t\}_{t=1}^{T}$ be the sequence of model parameters produced by an unlearning method applied to a sequence of forget sets $\{F_t\}_{t=1}^{T}$. Assuming the model maintains a stable level of predictive utility during the entirety of the unlearning stages, we say the model exhibits plasticity collapse if, as the number of unlearning tasks t increases, one or both of the following failure modes manifest:*

❶ ***Forward Failure.*** *The forgetting quality on task t progressively deteriorates as the number of sequential tasks increases.*

❷ ***Backward Failure.*** *Information from an earlier forget data F_s ($s < t$) re-emerges in the model at a later stage $t > s$ without any explicit re-exposure to the data in F_s, which is a spontaneous re-memorization phenomenon.*

Both failure modes are empirically demonstrated in Section 4. The following theoretical analysis reveals that both failure modes stem from a single underlying cause: the accumulation of geometric constraints in parameter space across sequential unlearning tasks.

3.3 Theoretical Analysis

In this section, we theoretically analyze why unlearning plasticity collapse exists in the continual unlearning dynamics. The key insight is that sequential unlearning tasks often induce updates that share directions in the parameter space $\mathbb{R}^d$. When the corresponding update operators are composed across tasks, repeated action along these shared directions can lead to geometric amplification under the conditions of the proposed theorems below. To prevent unbounded growth of the parameter iterates, the optimization dynamics must neutralize this expansion. This can occur either by suppressing future updates along the shared subspace, which limits the ability to modify predictions for new tasks (forward failure), or by introducing cancelling updates along the same directions, which partially undo earlier forgetting steps (backward failure).

Linear model. To analyze the continual unlearning dynamics in a tractable setting, we first instantiate the objective Eq. 2 with squared loss on a linear model. For any dataset $D = (\mathbf{X}_D, \mathbf{y}_D)$ with feature matrix $\mathbf{X}_D \in \mathbb{R}^{n_D \times d}$ and labels $\mathbf{y}_D \in \mathbb{R}^{n_D}$, the squared loss is

$$\mathcal{L}_D(\boldsymbol{\theta}) \;:=\; \frac{1}{2n_D}\|\mathbf{X}_D\boldsymbol{\theta} - \mathbf{y}_D\|_2^2, \tag{4}$$

with gradient $\nabla\mathcal{L}_D(\boldsymbol{\theta}) = \frac{1}{n_D}\mathbf{X}_D^\top(\mathbf{X}_D\boldsymbol{\theta} - \mathbf{y}_D)$. Substituting into Eq. 3 yields an affine recursion in $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}_t \;=\; \underbrace{\left(\mathbf{I} + \mathbf{M}_t^F - \lambda\mathbf{M}_t^R\right)}_{=:\,\mathbf{A}_t}\boldsymbol{\theta}_{t-1} \;-\; \underbrace{\left(\mathbf{b}_t^F - \lambda\mathbf{b}_t^R\right)}_{=:\,\mathbf{c}_t}, \tag{5}$$

where the matrices $\mathbf{M}_t^F, \mathbf{M}_t^R \in \mathbb{R}^{d\times d}$ and vectors $\mathbf{b}_t^F, \mathbf{b}_t^R \in \mathbb{R}^d$ are defined as

$$\mathbf{M}_t^F := \frac{\eta_t\mathbf{X}_{F_t}^\top\mathbf{X}_{F_t}}{n_{F_t}}, \quad \mathbf{M}_t^R := \frac{\eta_t\mathbf{X}_{R_t}^\top\mathbf{X}_{R_t}}{n_{R_t}}, \quad \mathbf{b}_t^F := \frac{\eta_t\mathbf{X}_{F_t}^\top\mathbf{y}_{F_t}}{n_{F_t}}, \quad \mathbf{b}_t^R := \frac{\eta_t\mathbf{X}_{R_t}^\top\mathbf{y}_{R_t}}{n_{R_t}}. \tag{6}$$

Note that $\mathbf{M}_t^F \succeq \mathbf{0}$ and $\mathbf{M}_t^R \succeq \mathbf{0}$ are positive semidefinite by construction. To capture the directional alignment across sequential tracking tasks, let $W \subset \mathbb{R}^d$ denote the shared subspace spanned by the continuous model-drift trajectories. The operator $A_t = I + M_t^F - \lambda M_t^R$ governs the evolution of parameter updates. In subspaces where the forget curvature dominates the scaled retain curvature,

A_t becomes expansive. The remainder of our analysis focuses on these expanding subspaces, which give rise to the operator-product instability underlying plasticity collapse. For clarity, we assume a single gradient step per task in Eq. 5. Performing K steps simply replaces $\mathbf{A}_t$ with $\mathbf{A}_t^K$, which preserves the expansive structure on the subspace W and leaves our theoretical conclusions below unchanged.

Special case: $\lambda = 0$. When no retain set is used, Eq. 5 simplifies to pure gradient ascent on the forget loss. Writing $\mathbf{X}_t \in \mathbb{R}^{k\times d}$ and $\mathbf{y}_t \in \mathbb{R}^k$ for the forget data at stage t, the update becomes

$$\boldsymbol{\theta}_{t+1} = (\mathbf{I} + \mathbf{M}_t)\,\boldsymbol{\theta}_t - \mathbf{b}_t, \qquad \mathbf{M}_t := \frac{\eta_t}{k}\mathbf{X}_t^\top \mathbf{X}_t \succeq \mathbf{0}, \qquad \mathbf{b}_t := \frac{\eta_t}{k}\mathbf{X}_t^\top \mathbf{y}_t. \tag{7}$$

We define the update increment $\mathbf{u}_t := \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}$ and the forget span

$$U_t \;:=\; \mathrm{Im}(\mathbf{X}_t^\top) \;=\; \mathrm{Im}(\mathbf{M}_t) \;\subseteq\; \mathbb{R}^d, \tag{8}$$

which is the subspace spanned by the input features of F_t. Since $\nabla\mathcal{L}_t(\boldsymbol{\theta}) = \frac{1}{k}\mathbf{X}_t^\top(\mathbf{X}_t\boldsymbol{\theta} - \mathbf{y}_t) \in U_t$ for all $\boldsymbol{\theta}$, every gradient step satisfies $\mathbf{u}_t \in U_t$. For a subspace $V \subseteq \mathbb{R}^d$, we write $\mathbf{P}_V \in \mathbb{R}^{d\times d}$ for the orthogonal projector onto V.

Definition 2 (Time-ordered product) *For indices $s \le t$, define the time-ordered product of the linear update operators as*

$$\mathbf{A}_{t:s} \;:=\; \mathbf{A}_t\mathbf{A}_{t-1}\cdots\mathbf{A}_s,$$

with the convention $\mathbf{A}_{t:t+1} := \mathbf{I}$. In the special case $\lambda = 0$, $\mathbf{A}_j = \mathbf{I} + \mathbf{M}_j$ and $\mathbf{A}_{t:s} = \prod_{j=s}^{t}(\mathbf{I} + \mathbf{M}_j)$.

Lemma 1 (Closed-form solution) *Let $\{\boldsymbol{\theta}_t\}$ follow the affine recursion Eq. 5. Then for any $1 \le s \le t$,*

$$\boldsymbol{\theta}_t \;=\; \mathbf{A}_{t:s}\,\boldsymbol{\theta}_{s-1} \;-\; \sum_{j=s}^{t} \mathbf{A}_{t:j+1}\,\mathbf{c}_j. \tag{9}$$

Proof. Substitute $\boldsymbol{\theta}_j = \mathbf{A}_j\boldsymbol{\theta}_{j-1} - \mathbf{c}_j$ recursively for $j = s, s{+}1, \ldots, t$. Collecting terms gives Eq. 9, using the convention $\mathbf{A}_{t:t+1} = \mathbf{I}$.

Lemma 1 exposes the key structural tension: the term $\mathbf{A}_{t:s}\,\boldsymbol{\theta}_{s-1}$ grows with t if $\mathbf{A}_{t:s}$ is expansive, and the forcing term $\sum_j \mathbf{A}_{t:j+1}\mathbf{c}_j$ must counterbalance this growth to keep $\boldsymbol{\theta}_t$ stable. The following theorems characterize when and how fast this growth occurs.

Theorem 1 (Exponential growth on a shared subspace, $\lambda = 0$) *Consider the $\lambda = 0$ dynamics Eq. 7. Fix $s \le t$ and let $\mathcal{T} \subseteq \{s, \ldots, t\}$ be any subset of task indices. Suppose there exists a nonzero subspace $W \subseteq \mathbb{R}^d$ satisfying:*

B1 *Shared subspace (invariance). For every $j \in \mathcal{T}$, $\mathbf{M}_j W \subseteq W$.*

B2 *Uniform relevance. There exists $\rho > 0$ such that for every $j \in \mathcal{T}$ and every $\mathbf{v} \in W$,*

$$\mathbf{v}^\top \mathbf{M}_j \mathbf{v} \;\geq\; \rho \|\mathbf{v}\|^2.$$

Then for all $\mathbf{v} \in W$,

$$\|\mathbf{A}_{t:s}\, \mathbf{v}\| \;\geq\; (1+\rho)^{|\mathcal{T}|}\, \|\mathbf{v}\|. \tag{10}$$

Consequently, for the trajectory θ_t to remain bounded, the dynamics must therefore suppress or counteract growth within W, which generically forces forward or backward failure.

The proof can be found in Appendix B.

Remark 1 (Geometric interpretation of (B1) and (B2)) *(**B1**) requires W to be a subspace shared across unlearning tasks. In practice, W may be approximated by the intersection of forget spans $\bigcap_t U_t$, or estimated as the dominant principal subspace of the stacked update matrix. (**B2**) requires that forget features project nontrivially onto W: since $\mathbf{v}^\top \mathbf{M}_j \mathbf{v} = \frac{\eta_j}{k} \|\mathbf{X}_j \mathbf{v}\|^2$, (**B2**) is equivalent to $\|\mathbf{X}_j \mathbf{v}\|^2 \geq \frac{k\rho}{\eta_j} \|\mathbf{v}\|^2$, which holds when different forget sets repeatedly activate the same feature directions. In practice, ρ is small due to moderate curvature imbalance between forget and retain directions, combined with learning rate scaling.*

Theorem 1 establishes the growth result for $\lambda = 0$. We now extend it to the general case with retain regularization.

Theorem 2 (Exponential growth under forget-dominance, general λ) *Fix $s \leq t$ and suppose there exists a nonzero subspace $W \subseteq \mathbb{R}^d$ such that for all $j \in \{s, \ldots, t\}$:*

G1 *Invariance. $\mathbf{A}_j W \subseteq W$, where $\mathbf{A}_j = \mathbf{I} + \mathbf{M}_j^F - \lambda \mathbf{M}_j^R$.*
G2 *Forget-dominance on W. There exists $\rho > 0$ such that for all $\mathbf{v} \in W$,*

$$\mathbf{v}^\top (\mathbf{M}_j^F - \lambda \mathbf{M}_j^R) \mathbf{v} \;\geq\; \rho \|\mathbf{v}\|^2.$$

Then for all $\mathbf{w} \in W$,

$$\|\mathbf{A}_{t:s}\, \mathbf{w}\| \;\geq\; (1+\rho)^{t-s+1}\, \|\mathbf{w}\|. \tag{11}$$

Proof. Under (**G2**), $\mathbf{A}_j|_W = (\mathbf{I} + \mathbf{M}_j^F - \lambda \mathbf{M}_j^R)|_W \succeq (1+\rho)\mathbf{I}_W$, so $\|\mathbf{A}_j \mathbf{w}\| \geq (1+\rho)\|\mathbf{w}\|$ for all $\mathbf{w} \in W$. Iterating via (**G1**) gives Eq. 11.

Remark 2 (The role of λ in (G2)) *(**G2**) holds when forget curvature exceeds the scaled retain curvature on W. When $\lambda = 0$ this reduces to (**B2**). For $\lambda > 0$, if the retain data overlaps strongly with W, then $\mathbf{M}_j^R|_W$ may dominate and $\mathbf{A}_j|_W$ becomes contractive, precluding exponential growth. However, when forget and retain sets are drawn from the same data distribution, their feature subspaces substantially overlap, and (**G2**) tends to hold for moderate λ.*

Remark 3 (Connecting to the two failure modes) *Theorems 1 and 2 characterize an operator-product instability: if there exists a (possibly approximate) shared subspace W such that the linear operators $\mathbf{A}_t$ satisfy invariance on W and expand on W (i.e., $(\mathbf{M}_t^F - \lambda \mathbf{M}_t^R)|_W \succeq \rho \mathbf{I}_W$), then any component along W is amplified exponentially under repeated composition, $\|\mathbf{A}_{t:s}\mathbf{w}\| \geq (1+\rho)^{t-s+1}\|\mathbf{w}\|$ for all $\mathbf{w} \in W$. Meanwhile, Lemma 1 shows that the iterate admits the decomposition $\boldsymbol{\theta}_t = \mathbf{A}_{t:s}\boldsymbol{\theta}_{s-1} - \sum_{j=s}^{t} \mathbf{A}_{t:j+1}\mathbf{c}_j$. Consequently, unless controlled, $\|\mathbf{P}_W \boldsymbol{\theta}_t\|$ grows geometrically, where $\mathbf{P}_W$ denotes the orthogonal projector onto W. For the trajectory to remain bounded, the dynamics must neutralize the expansion along W. One stabilization mechanism is to progressively reduce motion along W, thereby limiting further excitation of the expanding directions. However, if successive forget tasks share components in W, this restriction directly reduces the model's ability to modify predictions for new forget requests, resulting in progressive deterioration of forgetting quality (**Forward Failure**). Alternatively, the second mechanism is cancellation within W, whereby later optimization steps generate updates that oppose the accumulated parameter displacement along the shared subspace. Specifically, Specifically, the propagated forcing terms $\sum_{j=s}^{t} \mathbf{A}_{t:j+1}\mathbf{c}_j$ collectively encode the historical forgetting trajectories produced by successive unlearning tasks within W. To maintain a bounded parameter trajectory despite the exponential expansion of the homogeneous component $\mathbf{A}_{t:s}\boldsymbol{\theta}_{s-1}$, the optimization dynamics must introduce opposing components inside the same shared subspace. Consequently, the net parameter displacement associated with earlier forgetting operations is partially reduced. Since previous forget tasks repeatedly modify the model through this common subspace, reducing the accumulated displacement along W weakens the functional effect of earlier forgetting updates, allowing the model outputs on previously forgotten data to move back toward their pre-unlearning behavior. This manifests as spontaneous re-memorization of earlier forget tasks (**Backward Failure**). Overall, both failure modes thus emerge as distinct stabilization responses to the same operator-product expansion phenomenon characterized in Theorems 1 and 2.*

Now consider a neural network $f(\mathbf{x};\boldsymbol{\theta}) : \mathbb{R}^p \to \mathbb{R}$ with parameters $\boldsymbol{\theta} \in \mathbb{R}^d$. Let $\boldsymbol{\theta}_0$ denote the initialization, and assume training remains in the Neural Tangent Kernel (NTK) regime, so that the first-order linearization

$$f(\mathbf{x};\boldsymbol{\theta}) \approx f(\mathbf{x};\boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}} f(\mathbf{x};\boldsymbol{\theta}_0)^\top (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

is valid throughout optimization. For each task j, define the NTK feature matrices $\boldsymbol{\Phi}_{F_j} \in \mathbb{R}^{n_{F_j} \times d}$ and $\boldsymbol{\Phi}_{R_j} \in \mathbb{R}^{n_{R_j} \times d}$, whose rows are $\nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i;\boldsymbol{\theta}_0)^\top$ for samples in the forget and retain sets, respectively. Under squared loss with step size η_j, the parameter deviation $\boldsymbol{\delta}_j := \boldsymbol{\theta}_j - \boldsymbol{\theta}_0$ satisfies the affine recursion

$$\boldsymbol{\delta}_j = \mathbf{A}_j \boldsymbol{\delta}_{j-1} - (\mathbf{b}_j^F - \lambda \mathbf{b}_j^R),$$

where $\mathbf{A}_j = \mathbf{I} + \mathbf{M}_j^F - \lambda \mathbf{M}_j^R$, $\quad \mathbf{M}_j^F = \frac{\eta_j}{n_{F_j}} \boldsymbol{\Phi}_{F_j}^\top \boldsymbol{\Phi}_{F_j}$, $\quad \mathbf{M}_j^R = \frac{\eta_j}{n_{R_j}} \boldsymbol{\Phi}_{R_j}^\top \boldsymbol{\Phi}_{R_j}$.

Theorem 3 (NTK Extension) *Fix $s \leq t$. Suppose there exists a nonzero subspace $W \subset \mathbb{R}^d$ such that for all $j \in \{s, \ldots, t\}$:*

N1 *Invariance.* $\mathbf{A}_j W \subseteq W$.
N2 *Forget-dominance on* W. $(\mathbf{M}_j^F - \lambda \mathbf{M}_j^R)|_W \succeq \rho \mathbf{I}_W$ *for some* $\rho > 0$.

Then for all $\mathbf{w} \in W$,

$$\|\mathbf{A}_{t:s}\,\mathbf{w}\| \;\geq\; (1+\rho)^{t-s+1}\|\mathbf{w}\|. \tag{12}$$

Moreover, the parameter deviation satisfies

$$\boldsymbol{\delta}_t \;=\; \mathbf{A}_{t:s}\,\boldsymbol{\delta}_{s-1} \;-\; \sum_{j=s}^{t} \mathbf{A}_{t:j+1}\big(\mathbf{b}_j^F - \lambda \mathbf{b}_j^R\big), \tag{13}$$

so boundedness of $\boldsymbol{\delta}_t$ *along* W *requires cancellation of the exponentially growing homogeneous component* $\mathbf{A}_{t:s}\,\boldsymbol{\delta}_{s-1}$.

See Appendix C for proof.

3.4 Diagnostic Quantities

Theorems 1 and 2 establish that when unlearning tasks share a common expanding subspace W, the iterates must remain bounded through one of two compensating mechanisms, corresponding precisely to failure modes and in Definition 1. To detect these mechanisms empirically, we introduce two quantities that directly probe the geometry of parameter updates relative to W.

In practice, W is estimated from the update increments $\{\mathbf{u}_1, \ldots, \mathbf{u}_T\}$ by computing the top-r right singular vectors of the stacked matrix $\mathbf{U} = [\mathbf{u}_1 \mid \cdots \mid \mathbf{u}_T] \in \mathbb{R}^{d\times T}$. Let $\mathbf{Q} \in \mathbb{R}^{d\times r}$ denote the resulting orthonormal basis, so that $\mathbf{P}_W = \mathbf{Q}\mathbf{Q}^\top$ is the orthogonal projector onto the estimated W.

Definition 3 (Energy Ratio and Projection Coefficient) *For each stage* t, *we define Energy Ratio (ER) and Projection Coefficient (CO) as follows:*

$$\mathrm{ER}_t \;:=\; \frac{\|\mathbf{P}_W \mathbf{u}_t\|^2}{\|\mathbf{u}_t\|^2} \;=\; \frac{\|\mathbf{Q}\mathbf{Q}^\top \mathbf{u}_t\|^2}{\|\mathbf{u}_t\|^2}\,, \tag{14}$$

$$\mathrm{CO}_t \;:=\; \mathbf{Q}^\top \mathbf{u}_t\,. \tag{15}$$

ER_t measures what proportion of the update $\mathbf{u}_t$ lies within W: a value close to 1 indicates strong alignment with the shared subspace. CO_t records the signed projection of $\mathbf{u}_t$ onto each principal direction in W, capturing the direction and magnitude of engagement with W.

These two quantities connect directly to the two failure modes:

- **Energy Ratio for diminishing forgetting quality.** If $\mathbf{A}_{t:s}$ grows exponentially along W and the iterates remain stable, the algorithm must reduce the component of $\mathbf{u}_t$ along W, thereby causing ER_t toward 0. Since future forget sets have features in $U_t \subseteq W$ by (**B1**), an update nearly orthogonal to W induces diminishing forgetting quality for subsequent tasks.

(a) FT **(b)** NegGrad+ **(c)** RL **(d)** SalUn **(e)** MUNBa

Fig. 2: Forgetting accuracy of various unlearning methods on Tiny-ImageNet with VGG-16-BN under the random data forgetting scenario. The solid line denotes the forgetting accuracy on each task, while the scatter points represent the forgetting accuracy for single-task unlearning. Lower accuracy indicates better forgetting quality. For all methods, the forgetting accuracy increases as the number of sequential unlearning tasks grows, suggesting a progressive degradation in forgetting quality.

– **Coefficient for spontaneous re-memorization.** Alternatively, the algorithm may inject updates with opposing signs along W. Because the same shared directions also carry contributions from earlier unlearning steps (captured in the forcing terms c_j), such cancellation can partially undo prior forgetting-induced parameter changes, thereby restoring performance on previously forgotten data.

4 Experiments

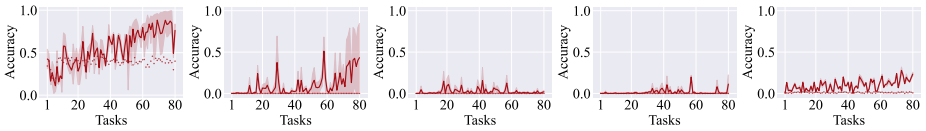
4.1 Experimental Setup

We evaluate plasticity collapse on image classification tasks. Experiments are conducted on **Tiny-ImageNet** [18] and **CIFAR-100**, using **VGG-16-BN** and **PreResNet-110** as backbone architectures. The results on CIFAR-100 can be found in Appendix D. We study five representative unlearning methods: **Fine-tuning (FT)**, **NegGrad+** [17], **RandomLabeling (RL)** [12], **SalUn** [8], and **MUNBa** [28]. We exclude FG-OrIU [9] from our baselines due to its inherent limitation to ViT architectures. Specifically, FG-OrIU relies on parameter-efficient fine-tuning via LoRA, which restricts its applicability to ViT models and prevents straightforward extension to architectures such as VGG and ResNet.

We consider two forgetting scenarios. In **random data forgetting**, each task F_t consists of 1% of training data selected uniformly at random, producing forget spans U_t with relatively unstructured overlap. In **class-wise forgetting**, each F_t consists of all samples from a designated class, producing more structured and concentrated forget spans.

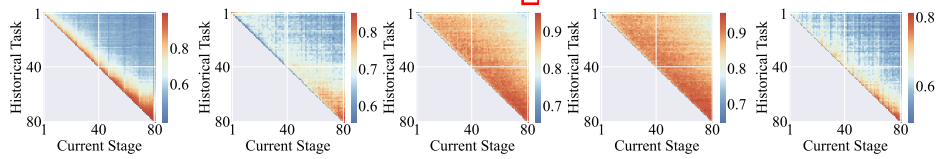
4.2 Results

Empirical Evidence for Two Failure Modes. The accuracy of the forget data is set as the primary metric for unlearning quality, as it is the most commonly used measure in prior work [8, 14, 28, 29]. Lower forgetting accuracy indicates higher forgetting quality. Specifically, at each stage t, we measure model accuracy on



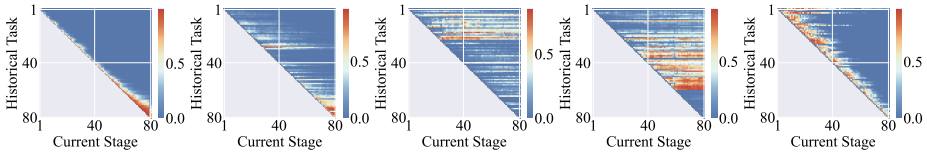
(a) FT **(b)** NegGrad+ **(c)** RL **(d)** SalUn **(e)** MUNBa

Fig. 3: Forgetting accuracy on Tiny-ImageNet with VGG-16-BN under the class-wise forgetting scenario. FT, NegGrad+ and MUNBa also show diminishing forgetting quality for subsequent tasks. RL and SalUn do not exhibit the clear forward failure in this setting, but we show its backward failure in Fig 5.



(a) FT **(b)** NegGrad+ **(c)** RL **(d)** SalUn **(e)** MUNBa

Fig. 4: Forgetting accuracy heatmaps on Tiny-ImageNet with VGG-16-BN for historical tasks on the random data forgetting scenario. Entry (i, j) records forgetting accuracy of task F_i on the model $\boldsymbol{\theta}_j$ where $j \geq i$.



(a) FT **(b)** NegGrad+ **(c)** RL **(d)** SalUn **(e)** MUNBa

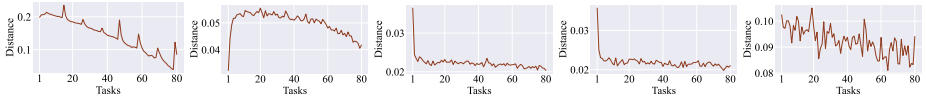
Fig. 5: Forgetting accuracy heatmaps on Tiny-ImageNet with VGG-16-BN for historical tasks on the class-wise forgetting scenario. RL and SalUn, which do not exhibit forward failure, show obvious backward failure in this setting.

F_t after the unlearning update. Results on retain and test accuracy are reported in Appendix D. The results are average values from 3 independent trials.

Figs. 2 and 3 plot forgetting accuracy across sequential tasks for all five methods under both scenarios. As a reference baseline, we also report the forgetting accuracy achieved by each method when applied to a single task in isolation (scatter points in Figs. 2 and 3), allowing a direct comparison between single-task unlearning and continual unlearning to highlight the degradation induced by the sequential setting.

Under **random data forgetting** as shown in Fig. 2, all five methods exhibit increasing forgetting accuracy as tasks accumulate. This trend empirically validates our forward failure mode of plasticity collapse. On average across methods, forgetting accuracy of the last task in the continual setting is approximately 33.2% higher than in the single-task baseline, with NegGrad+ showing the most severe degradation. Notably, even MUNBa, one of the most recent unlearning methods, is also not immune to this effect.

Under **class-wise forgetting** as shown in Fig. 3, the situation is more nuanced. Since class-wise forgetting is an inherently easier task, most existing methods achieve near-0% forgetting accuracy in the single-task setting. In the contin-



(a) FT **(b)** NegGrad+ **(c)** RL **(d)** SalUn **(e)** MUNBa

Fig. 6: L2 norm of parameter differences between consecutive updates on Tiny-ImageNet with VGG-16-BN under the random data forgetting scenario. It shows an obvious decrease in the magnitude of parameter updates as the number of tasks increases, indicating that the model becomes increasingly frozen in the parameter space.

(a) FT **(b)** NegGrad+ **(c)** RL **(d)** SalUn **(e)** MUNBa

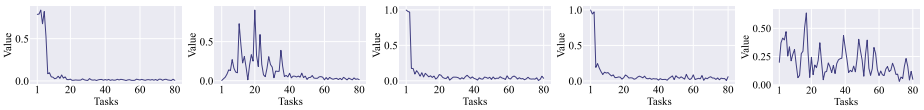
Fig. 7: L2 norm of parameter differences between consecutive updates on Tiny-ImageNet with VGG-16-BN under the class-wise forgetting scenario. For RL and SalUn, which exhibit backward failure, the tasks where re-memorization occurs show lower L2 norm values compared to other tasks.

ual setting, however, this strong performance no longer holds. FT, NegGrad+, and MUNBa all show clearly increasing forgetting accuracy as tasks accumulate, confirming forward failure mode. RL and SalUn exhibit more moderate increases and do not display a pronounced forward failure under this scenario. However, as we show next, both methods suffer from a backward failure mode, i.e., spontaneous re-memorization.

Figs. 4 and 5 track historical task forgetting accuracy over time, where entry (i, j) records forgetting accuracy of task F_i on the model $\boldsymbol{\theta}_j$ where $j \geq i$. The heatmap results are shown for a single random seed to facilitate the CO analysis presented later. The diagonal entries, representing current-task forgetting quality, show a consistent blue-to-red gradient as t increases across Figs. 4a–4e and Figs. 5a, 5b, 5e, confirming progressive forgetting quality degradation, i.e., **forward failure**. The off-diagonal entries above the diagonal reveal the **backward failure** mode. Tasks that were initially well-forgotten (blue on diagonal entries) later transition to red for RL (Fig. 5c) and SalUn (Fig. 5d) under class-wise forgetting. Taking SalUn as a concrete example (Fig. 5d), most of the tasks in the index range 35-60 exhibit pronounced re-memorization, even though these tasks were well forgotten at their current stages.

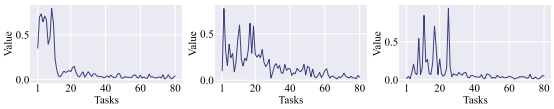
Importantly, the two failure modes are not mutually exclusive. MUNBa, for instance, displays both forward (Fig. 3e) and backward (Fig. 5e) failure under class-wise forgetting scenario, demonstrating that a single method can suffer from both forms of plasticity collapse. **Taken together, these results show that plasticity collapse (both forward and backward failure modes) is a universal phenomenon, manifesting across different unlearning methods and forgetting scenarios.**

Update magnitude diminishes over time. Theorems 1 and 2 predict that as W saturates, the suppression mechanism increasingly drives parameter updates away from W. This phenomenon is expected to manifest as a systematic reduc-



(a) FT **(b)** NegGrad+ **(c)** RL **(d)** SalUn **(e)** MUNBa

Fig. 8: Energy Ratio for forward failure diagnosis on Tiny-ImageNet with VGG-16-BN under the random data forgetting scenario.



(a) FT **(b)** NegGrad+ **(c)** MUNBa

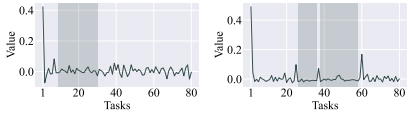
Fig. 9: Energy Ratio for forward failure diagnosis on Tiny-ImageNet with VGG-16-BN under the class-wise forgetting scenario.

tion in the update magnitude, defined as $\mathbf{u}_t := \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}$. Figs. 6 and 7 empirically corroborate this. $\|\mathbf{u}_t\|$ declines consistently in scenarios exhibiting forward failure, with the most pronounced decay aligning precisely with the stages where forgetting quality degrades most severely. Furthermore, for cases characterized by backward failure (Figs. 7c-7d), tasks exhibiting the re-memorization phenomenon (shaded region) also display significantly lower L_2 norms compared to others. **These observations provide direct evidence that the model becomes increasingly frozen in the parameter space, progressively losing its capacity to generate the meaningful updates required to fulfill new unlearning requests.**

Diagnostic Quantities for Plasticity Collapse. We now apply ER_t and CO_t in Section 3.4 to directly validate the theoretical predictions. The subspace W is estimated from $\{\mathbf{u}_1, \ldots, \mathbf{u}_T\}$ via PCA of the stacked update matrix $\mathbf{U}$, with $\mathbf{Q}$ taken as the leading $r = 1$ right singular vector for visualization clarity. Patterns are consistent for $r > 1$.

Figs. 8 and 9 plot ER_t across sequential tasks. In the cases with **forward failure**, ER_t decays progressively toward zero, confirming the suppression mechanism: as $\mathbf{A}_{t:s}$ accumulates exponential growth along W, it is forced to route updates away from W to maintain stability. The decay of ER_t closely tracks the degradation of forgetting accuracy, providing quantitative evidence that geometric saturation of W is the proximate cause of forward failure.

Fig. 10 plots CO_t for methods exhibiting re-memorization. Tasks undergoing re-memorization show $\mathrm{CO}_t \approx 0$. Their updates fail to engage the principal forgetting direction, indicating the model cannot make progress along W. Tasks surrounding re-memorization events display pronounced sign oscillations in CO_t, alternating between large positive and negative values. These oscillations directly



(a) RL **(b)** SalUn

Fig. 10: Coefficient for backward failure diagnosis on Tiny-ImageNet with VGG-16-BN under the class-wise forgetting scenario.

reflect antagonistic interference: forgetting a new task requires injecting updates along W that partially undo prior forgetting operations, triggering re-memorization of earlier tasks.

Together, the decay of ER_t and the oscillatory behavior of CO_t provide consistent, theory-grounded evidence for both failure modes of plasticity collapse, closing the loop between the theoretical predictions of Theorems 1–2 and the empirical observations.

5 Conclusion

We identify loss of unlearning plasticity as a fundamental challenge in continual machine unlearning. Through theoretical analysis and empirical validation, we demonstrate that models undergoing sequential unlearning operations exhibit deteriorating forgetting quality and spontaneous re-memorization. These findings reveal that current unlearning methods, designed for single-shot scenarios, face severe limitations in realistic continual settings, necessitating fundamentally new plasticity-preserving approaches for practical deployment.

Our theoretical analysis provides actionable guidance for plasticity-preserving unlearning methods in the future. Specifically, the Energy Ratio and Projection Coefficient introduced in Section 3.4 can serve as lightweight online monitors for detecting impending collapse, and can be naturally incorporated as regularization signals into existing gradient-based unlearning objectives. We hope this work serves as both a cautionary analysis and a constructive foundation for the next generation of continual unlearning algorithms.

Acknowledgments

This work was supported in part by the National Science Foundation under grants IIS-2246157, FMitF-2319243, and the Department of Energy under grant DE-CR0000042. The project was also supported by computational resources provided by the NSF ACCESS and Argonne National Lab.

References

1. Adhikari, S., Kumaravelu, V., Srijith, P.: An unlearning framework for continual learning. arXiv preprint arXiv:2509.17530 (2025)
2. Ash, J., Adams, R.P.: On warm-starting neural network training. Advances in neural information processing systems **33**, 3884–3894 (2020)
3. Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C.A., Jia, H., Travers, A., Zhang, B., Lie, D., Papernot, N.: Machine unlearning. In: 2021 IEEE symposium on security and privacy (SP). pp. 141–159. IEEE (2021)
4. Cao, Y., Yang, J.: Towards making systems forget with machine unlearning. In: 2015 IEEE symposium on security and privacy. pp. 463–480. IEEE (2015)
5. Chatterjee, R., Chundawat, V., Tarun, A., Mali, A., Mandal, M.: A unified framework for continual learning and unlearning. arXiv preprint arXiv:2408.11374 (2024)

6. Dohare, S., Hernandez-Garcia, J.F., Lan, Q., Rahman, P., Mahmood, A.R., Sutton, R.S.: Loss of plasticity in deep continual learning. Nature **632**(8026), 768–774 (2024)
7. Evron, I., Moroshko, E., Ward, R., Srebro, N., Soudry, D.: How catastrophic can catastrophic forgetting be in linear regression? In: Conference on Learning Theory. pp. 4028–4079. PMLR (2022)
8. Fan, C., Liu, J., Zhang, Y., Wei, D., Wong, E., Liu, S.: Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In: International Conference on Learning Representations (2024)
9. Feng, Q., Tu, J., Kang, M., Zhao, H., Zhang, C., Qian, H.: Fg-oriu: Towards better forgetting via feature-gradient orthogonality for incremental unlearning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1957–1967 (2025)
10. Ginart, A., Guan, M., Valiant, G., Zou, J.Y.: Making ai forget you: Data deletion in machine learning. Advances in neural information processing systems **32** (2019)
11. Goyal, A., Bengio, Y.: Inductive biases for deep learning of higher-level cognition. Proceedings of the Royal Society A **478**(2266), 20210068 (2022)
12. Graves, L., Nagisetty, V., Ganesh, V.: Amnesiac machine learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 11516–11524 (2021)
13. Guo, C., Goldstein, T., Hannun, A., Van Der Maaten, L.: Certified data removal from machine learning models. arXiv preprint arXiv:1911.03030 (2019)
14. Huang, Z., Cheng, X., Zhang, J., Zheng, J., Wang, H., He, Z., Li, T., Huang, X.: A unified gradient-based framework for task-agnostic continual learning-unlearning. arXiv preprint arXiv:2505.15178 (2025)
15. Kumar, A., Agarwal, R., Ghosh, D., Levine, S.: Implicit under-parameterization inhibits data-efficient deep reinforcement learning. arXiv preprint arXiv:2010.14498 (2020)
16. Kumar, S., Marklund, H., Van Roy, B.: Maintaining plasticity in continual learning via regenerative regularization. arXiv preprint arXiv:2308.11958 (2023)
17. Kurmanji, M., Triantafillou, P., Hayes, J., Triantafillou, E.: Towards unbounded machine unlearning. Advances in neural information processing systems **36**, 1957–1987 (2023)
18. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015)
19. Lyle, C., Rowland, M., Dabney, W.: Understanding and preventing capacity loss in reinforcement learning. arXiv preprint arXiv:2204.09560 (2022)
20. Mirzadeh, S.I., Chaudhry, A., Yin, D., Hu, H., Pascanu, R., Gorur, D., Farajtabar, M.: Wide neural networks forget less catastrophically. In: International conference on machine learning. pp. 15699–15717. PMLR (2022)
21. Neel, S., Roth, A., Sharifi-Malvajerdi, S.: Descent-to-delete: Gradient-based methods for machine unlearning. In: Algorithmic Learning Theory. pp. 931–962. PMLR (2021)
22. Shi, Y., Liu, S., Ding, K., Wang, R.: Tackling fake forgetting through uncertainty quantification. In: Forty-third International Conference on Machine Learning (2026)
23. Shi, Y., Wang, R.: Exploring nonlinear pathway in parameter space for machine unlearning. In: Forty-third International Conference on Machine Learning (2026)
24. Sokar, G., Agarwal, R., Castro, P.S., Evci, U.: The dormant neuron phenomenon in deep reinforcement learning. In: International Conference on Machine Learning. pp. 32145–32168. PMLR (2023)

25. Tang, J., Zhuang, H., Fang, D., Li, J., Han, F., Huang, Y., Fan, K., Wang, L., Zhu, Z., Zhang, S., et al.: Acu: Analytic continual unlearning for efficient and exact forgetting with privacy preservation. arXiv preprint arXiv:2505.12239 (2025)
26. Thudi, A., Deza, G., Chandrasekaran, V., Papernot, N.: Unrolling sgd: Understanding factors influencing machine unlearning. In: 2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P). pp. 303–319. IEEE (2022)
27. Warnecke, A., Pirch, L., Wressnegger, C., Rieck, K.: Machine unlearning of features and labels. arXiv preprint arXiv:2108.11577 (2021)
28. Wu, J., Harandi, M.: Munba: Machine unlearning via nash bargaining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4754–4765 (2025)
29. Zhao, K., Kurmanji, M., Bărbulescu, G.O., Triantafillou, E., Triantafillou, P.: What makes unlearning hard and what to do about it. Advances in Neural Information Processing Systems **37**, 12293–12333 (2024)