Revisiting Gradient Ascent: Machine Unlearning from a Geometric Perspective for Source-Free Scenarios

Yufeng Cao[1], Xuyang Teng[2,*], Naen Xu[3], Tianyu Du[3], and Qiang Zhao[2]

[1]School of Automation (School of Artificial Intelligence), Hangzhou Dianzi University, Hangzhou, China,
[2]School of Communication Engineering, Hangzhou Dianzi University, Hangzhou, China,
[3]School of Software Technology, Zhejiang University, Hangzhou, China,

Abstract

Machine unlearning is becoming increasingly indispensable for meeting ever-stringent data compliance requirements, such as the "right to be forgotten". While Gradient Ascent (GA) based approximate methods have drawn attention for their conceptual simplicity, their practical deployment is severely constrained by three fundamental issues: first, the lack of feasibility analysis leads to divergent update directions; second, gradient conflict inevitably degrades the performance maintenance on retained data; and third, heavy reliance on the retain set restricts their applicability in real-world settings. To address these challenges, we conduct an in-depth investigation from a geometric perspective into the underlying root causes of instability in conventional GA methods. Our analysis reveals that the feasibility of unlearning intrinsically depends on specific geometric constraint relationships within the parameter space. Building upon this theoretical insight, we formally establish the methods governing unlearning feasibility. The first is the direction feasibility, which constrains the feasible update trajectory and keeps the model within reasonable parameter regions during unlearning. The second is the step-size feasibility, which further restricts the update magnitude along the feasible direction to prevent model collapse caused by large steps. Notably, we find that the direction feasibility does not rely on any information from the retain set. In light of this, we further explore viable approaches for effectively estimating the step-size feasibility under strict source-free scenarios.

1 Introduction

With the increasing data compliance requirements (e.g., the "right to be forgotten") faced by machine learning models in practical deployment, machine unlearning has emerged as a research hotspot in the field of machine learning [6, 31, 33, 16]. Its core objective is to enable deployed models to unlearn the influence of specific training samples, thereby eliminating their contribution to the models behavior and meeting the demand for selective forgetting. However, how to ensure unlearning efficacy while maintaining the models performance on the retain set (i.e., the data not to be unlearned) remains a highly challenging problem [45, 12, 34].

Existing machine unlearning methods can be broadly categorized into exact unlearning and approximate unlearning [46]. Exact unlearning methods typically require retraining the model at a relatively high computational cost, while approximate unlearning methods aim for more efficient update mechanisms [48, 36]. Approximate unlearning methods are typically Fine-Tuning (FT)-based, Hessian-based (i.e., influence functions), and Gradient Ascent (GA)-based. FT-based methods suffer

*Corresponding author: tengxuyang@hdu.edu.cn

from persistent parameter residuals and remain trapped near the original optimum [20], and Hessianbased methods suffer from poor scalability and reliance on convexity assumptions that rarely hold in practice [5, 28]. Conversely, GA-based methods have gained attention due to their simplicity and theoretical clarity. However, their deployment in practical unlearning tasks remains constrained by three fundamental limitations:

First, GA-based methods lack a prior assessment of the feasibility of unlearning, such as parameter update direction. By merely updating parameters along the GA direction of the forget set (i.e., the data to be unlearned), they overlook the geometric relationships between the gradients of samples in forget set, leading to divergent update directions and a sharp decline in model performance [37]. *Second*, the gradient conflict between the unlearning efficacy and the performance maintenance (i.e., forgetting vs retention) is difficult to reconcile. Due to the inherent statistical dependencies between forget set and retain set, GA-based methods often compromise model performance on the retain set, easily resulting in excessive unlearning [40]. *Third*, GAbased methods typically require access to retain set in order to finetune the model, thereby preserving performance on retain set. However, due to limited data storage and strict privacy regulations, the retain set is typically inaccessible in realworld unlearning settings (i.e., sourcefree). Consequently, these methods are rendered inapplicable.

To address these limitations, we revisit the GA-based unlearning approach and propose a theoretically guaranteed GA framework whose unlearning feasibility is established through local linearization. By analyzing the geometric structure of the gradient space, we find that the *trade-off between unlearning efficacy and performance maintenance is not inherently opposing, but depends on the geometric relationship between the parameter update direction and the sample gradients*. This geometric perspective reveals the root cause of instability in the GA method, and leads to the derivation of the geometric sufficient conditions for the feasibility of unlearning. Furthermore, by adding perturbations to the forget set, we eliminate the need for an explicit retain set. Consequently, our framework provides a principled definition of unlearning feasibility, and extends it to sourcefree scenarios.

In summary, our main contributions are summarized as follows:

- We geometrically diagnose the root cause of instability in GA methods (gradient direction conflict and step-size mismatch problem) and derive the sufficient conditions for unlearning feasibility.
- We estimate the gradient statistics and threshold step size of the retain set by adding Gaussian perturbations to the model parameters, thereby extending the unlearning feasibility to sourcefree scenarios.
- We experimentally validate that this theory effectively guides unlearning algorithms to achieve efficient, thorough, and interpretable data removal while ensuring performance maintenance.

2 Related Work

2.1 Machine Unlearning

Existing machine unlearning methods can be divided into exact unlearning and approximate unlearning [38]. Exact unlearning methods typically remove the influence of specific samples completely through retraining or sharded storage (e.g., SISA [7]), but incur high computational costs [32, 18]. Approximate unlearning methods pursue more efficient update mechanisms, including FT-based methods, Hessian-based methods, bi-level optimization-based methods, and GA-based methods. FT-based methods overwrite the influence of the forget set by further fine-tuning the model on the retain set [20, 21]. Hessian-based methods leverage second-order information to approximate unlearning updates [51, 43, 35]. Bi-level optimization-based methods formulate unlearning as a hierarchical optimization problem, where the lower-level objective minimizes the forgetting loss and the upper-level objective preserves model utility, thereby mitigating the forgettingretention tradeoff [3, 42, 44, 15, 47]. However, FT-based struggles to truly eliminate data influence as parameter residuals persist and the model remains trapped near the original optimum, Hessian-based methods are computationally expensive and unstable, while GA attracts attention due to its simplicity.

Early GA methods directly update parameters along the GA direction of the forget set, eliminating their influence by maximizing the models loss on the forget set [23, 50, 19]. Latest GA methods attempt to introduce retain set constraints to mitigate the conflict between forgetting and retention objectives, mainly including: Normalized Gradient Difference (NGDiff), which simultaneously minimizes retain loss and maximizes forget loss [26]; Forget the Token and Pixel (FTTP), which

applies GA only to target tokens/pixels and Gradient Descent (GD) elsewhere to preserve performance [30]; Negative Preference Optimization (NPO) avoids GA divergence through preference optimization [14]; label smoothing-based methods improve GA by introducing a smoothing mechanism [11, 39]. However, these methods commonly suffer from divergent update directions and poor retention.

2.2 Source-Free Unlearning

In practice, privacy regulations (e.g., GDPR) or storage limitations may render the retain set inaccessible or too large to repeatedly access. This leads to a more challenging sourcefree unlearning setting: only the forget set is available, while the retain set is completely unavailable. In source-free scenarios, existing unlearning methods mainly include: Hessian-based methods [1], which indirectly estimate retain set Hessian information by adding random perturbations to model parameters and observing loss changes, thereby approximating Newton updates without accessing the original retain data; KL divergencebased methods [4, 29], which only need to calibrate noise using the statistical distance between a surrogate dataset and the source data, and the calculation of KL divergence does not depend on actual retained samples, thus achieving certified unlearning without source data; score forgetting distillation method [9], which only requires aligning the conditional scores of unsafe categories with those of safe categories based on the models own outputs, and the entire process does not involve the original training data, making it naturally applicable to sourcefree scenarios.

3 Preliminaries

In this paper, we primarily focus on discriminative models. Given the rapidly increasing adoption of generative models, our theory can also be extended to the generative setting. Please refer to Appendix **??** for a more detailed discussion. Let $\theta_0 \in \Theta$ be the initial model parameters optimized on the dataset $\mathcal{D}$, $\mathcal{F}$ be the forget set, and $\mathcal{R}$ be the retain set ($\mathcal{R} = \mathcal{D} \setminus \mathcal{F}$). In the formulation of machine unlearning, it is typically necessary to simultaneously prevent significant degradation of model performance on the retain set while enforcing a marked degradation on the forget set [17]. To quantify the allowed deviations in these two opposite directions, we introduce the following definitions [53, 13]. For each sample i, the loss function $\mathcal{L}_i : \Theta \to \mathbb{R}$ satisfies $\mathcal{L}_i(\theta_0) \leq \epsilon_i$, where $\epsilon_i \geq 0$ is a retention threshold that ensures retention objective. For forgotten samples $j \in \mathcal{F}$, we require their performance to degrade (forgetting objective), i.e., there exists a forgetting threshold $\tau_j > \mathcal{L}_j(\theta_0)$, and we expect $\mathcal{L}_j(\theta) \geq \tau_j$. For the convenience of later notation, we denote

$$\Delta_i = \epsilon_i - \mathcal{L}_i(\theta_0) \geq 0, \quad \Delta_j = \tau_j - \mathcal{L}_j(\theta_0) > 0. \tag{1}$$

After defining the above retention and forgetting threshold, we introduce the following assumptions to analyze the existence of local approximate unlearning, which restricts parameter changes within a range. To ensure that the error of the linear approximation is bounded by a quadratic term, thereby providing a basis for local analysis, we introduce the smoothness assumption.

Assumption 1 (Smoothness)**.** *There exists a constant $L > 0$ such that for any small perturbation δ,*

$$|\mathcal{L}_i(\theta_0 + \delta) - \mathcal{L}_i(\theta_0) - \nabla \mathcal{L}_i(\theta_0)^\top \delta| \leq \frac{L}{2}\|\delta\|^2. \tag{2}$$

Corollary 1 (Exact Quadratic Constraints)**.** *Under Assumption 1, for $\delta = tu$ with $t > 0$, $\|u\| = 1$, the unlearning and retention constraints are equivalent to*

$$t\nabla \mathcal{L}_j(\theta_0)^\top u \geq \Delta_j - \frac{L}{2}t^2 \quad (j \in \mathcal{F}), \qquad t\nabla \mathcal{L}_i(\theta_0)^\top u \leq \Delta_i + \frac{L}{2}t^2 \quad (i \in \mathcal{R}). \tag{3}$$

These form the basis for analyzing the feasibility of local approximate unlearning.

Under source-free scenarios, we cannot directly compute $\nabla \mathcal{L}_i(\theta_0)$ and can only infer relevant statistics through the forget set. Thus, we introduce the following two assumptions. First, the local correlation assumption captures the intuition that the local behavioral correlation between the forget set and the retain set is key to inferring properties of the retain set from the forget set alone. Second, the gradient direction consistency assumption ensures that the projection of retainset gradients onto any direction is not excessively small, which in turn guarantees a conservative lower bound on the step size.

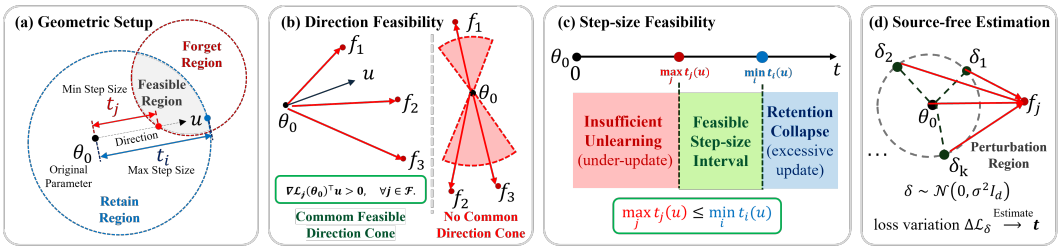


Figure 1: Geometric illustration of source-free unlearning feasibility. Subfigure (a) shows the geometric framework setup, including intuitive illustrations of direction and stepsize conditions. Subfigure (b) illustrates the direction feasibility scenario, i.e., the forgotten samples share a common direction. Subfigure (c) illustrates the stepsize feasibility scenario, i.e., the adopted step size lies between the minimum step size threshold by the forgetting objective and the maximum step size threshold by the retention objective. Subfigure (d) demonstrates how to extend the theory to the sourcefree setting by adding perturbation to the model parameters.

Assumption 2 (Local Correlation)**.** *There exists $C \geq 1$ such that for any small perturbation δ,*

$$\frac{1}{|\mathcal{R}|}\sum_{i\in\mathcal{R}}|\mathcal{L}_i(\theta_0+\delta)-\mathcal{L}_i(\theta_0)| \leq C\cdot\frac{1}{|\mathcal{F}|}\sum_{j\in\mathcal{F}}|\mathcal{L}_j(\theta_0+\delta)-\mathcal{L}_j(\theta_0)|. \tag{4}$$

Assumption 3 (Gradient Direction Consistency)**.** *There exists a constant $\alpha \in (0,1)$ such that for any unit direction u,*

$$\min_{i:\nabla\mathcal{L}_i(\theta_0)^\top u>0}\nabla\mathcal{L}_i(\theta_0)^\top u \leq \alpha\cdot\sqrt{\frac{1}{|\mathcal{R}|}\sum_{i\in\mathcal{R}}\|\nabla\mathcal{L}_i(\theta_0)\|^2}. \tag{5}$$

4 Methodology

In this section, we first conduct a detailed analysis of the limitations of existing GA-based methods. Next, we propose unlearning feasibility conditions based on the changes in loss values on the retain set and the forget set. We then extend our theory to sourcefree scenarios.

4.1 Limitations of Existing Methods

Existing GA-based unlearning methods typically adopt the following update rule:

$$\theta_{t+1} = \theta_t + \eta\cdot\nabla_\theta\mathcal{L}_f(\theta_t;\mathcal{F}) - \lambda\cdot\nabla_\theta\mathcal{L}_r(\theta_t;\mathcal{R}), \tag{6}$$

where $\mathcal{L}_f$ and $\mathcal{L}_r$ are the loss functions on the forget set and the retain set, respectively, and $\eta, \lambda > 0$ are step-size coefficients. This update attempts to increase the loss on the forgotten samples (GA) while decreasing the loss on the retained samples (GD). Although this update rule is intuitive, it lacks rigorous theoretical support. Thus, this approach suffers from two fundamental flaws: direction conflict and step-size mismatch.

Direction conflict: Since $\mathcal{F}$ and $\mathcal{R}$ are user-specified, they are not necessarily independent and may in fact be correlated. Consequently, the average gradients on the forget set and the retain set are often non-orthogonal and may even form an acute angle. In such cases, the two gradient directions come into conflict: moving along one direction also increases the retention loss, while moving along the other undermines the forgetting performance. More severely, when there is a conflict in $\mathcal{F}$, such as when samples with completely different features need to be forgotten, no single-direction update can simultaneously increase the loss on all forgotten samples. This leads to the direction feasibility condition: there must exist a common direction u such that the projection of the gradient of every forgotten sample onto u is positive. If no such u exists, then regardless of the step size, no movement can achieve unlearning.

Step size mismatch: In traditional GA-based unlearning, the effective update magnitude $t = \|\theta_{t+1} - \theta_t\|$, implicitly needs to lie between the minimum tolerable for forgetting objective and the maximum tolerable for retention objective. Intuitively, adaptive learning rates can dynamically adjust

parameter step sizes and seem to automatically fall into this interval. However, recent studies [52, 8] point out that unlearning algorithms are highly sensitive to the learning rate, and even the introduction of adaptive schedulers cannot fully eliminate this dependency, which can also be observed in our experiments Section 5. Therefore, the step-size mismatch problem persists.

4.2 Feasibility Condition for Unlearning

To unify the analysis of the two aforementioned flaws, we integrate the direction condition and the step-size condition into a unified local update condition. Theorem 1 provides a sufficient condition for the existence of a small perturbation $\delta = tu$ that simultaneously satisfies the forgetting and the retention threshold. This theorem transforms the problems of direction conflict and step-size mismatch into numerical inequalities concerning gradient projections and retention thresholds.

Theorem 1 (Unlearning Feasibility)**.** *Under Assumption 1, there exists a small perturbation $\delta = tu$ that satisfies the forgetting and retention threshold if the following two conditions hold:*

1. ***Direction condition:*** *For all $j \in \mathcal{F}$, there exists a unit direction u such that*

$$\nabla \mathcal{L}_j(\theta_0)^\top u > 0. \tag{7}$$

2. ***Step-size condition:*** *Given that the direction condition holds, define the minimum step size required for forgetting*

$$t_j(u) = \frac{-\nabla \mathcal{L}_j(\theta_0)^\top u + \sqrt{(\nabla \mathcal{L}_j(\theta_0)^\top u)^2 + 2L\Delta_j}}{L}, \tag{8}$$

and the maximum step-size tolerable for retention objective (when $\nabla \mathcal{L}_i^\top u > 0$)

$$t_i(u) = \frac{\nabla \mathcal{L}_i(\theta_0)^\top u - \sqrt{(\nabla \mathcal{L}_i(\theta_0)^\top u)^2 - 2L\Delta_i}}{L}. \tag{9}$$

Then there exists $t > 0$ if and only if

$$\max_{j \in \mathcal{F}} t_j(u) \leq \min_{i: \nabla \mathcal{L}_i^\top u > 0} t_i(u). \tag{10}$$

If no retain sample satisfies $\nabla \mathcal{L}_i^\top u > 0$, the right-hand side is treated as $+\infty$.

Please see Appendix A for the detailed derivation.

The direction condition guarantees the existence of a common direction u such that the loss of every forgotten sample can increase along this direction. The step-size condition further requires that the minimum step size needed for unlearning does not exceed the maximum step size tolerable for the retention objective. Specifically, $t_j(u)$ is given by the smallest positive solution to the quadratic inequality $\mathcal{L}_j(\theta_0 + tu) - \mathcal{L}_j(\theta_0) \geq \Delta_j$, and $t_i(u)$ is given by the largest solution to $\mathcal{L}_i(\theta_0 + tu) - \mathcal{L}_i(\theta_0) \leq \Delta_i$. Our derived Theorem 1 provides a complete criterion for local unlearning feasibility.

4.3 Step-size Feasibility Estimation in the Source-Free Scenario

Note that *the direction condition involves only the gradients of forgotten samples and is completely independent of the retain set, so it can be directly adapted to in source-free scenarios.*

Traditional unlearning methods assume that the retain set $\mathcal{R}$ is fully accessible, allowing direct computation of retain gradients or verification of constraints. However, in source-free unlearning scenarios, we cannot directly obtain $\nabla \mathcal{L}_i(\theta_0)$ to test the step-size condition. Inspired by [1], we estimate the gradient statistics of the forgotten samples by adding small random perturbations to the model and observing the loss changes on the forget set. On this basis, we infer lower bounds on the retain gradient projections from the perturbation responses of the forgotten samples, thereby conservatively testing the step-size condition.

Here, we first estimate an upper bound on the mean squared norm to keep the projection within a certain range; after obtaining this upper bound, we can derive a lower bound for each projection and complete the verification of the step-size condition. See the Appendix B for detailed derivation.

4.3.1 Precondition: Upper Bound on the Retain Gradient Norm

The step-size condition involves the gradients of the retained samples, which are inaccessible in the source-free setting. However, under Assumption 2, the average absolute loss change of the retain set is bounded by that of the forget set up to a constant factor C. This allows us to estimate the retain gradient statistics using only the forget sets responses to random parameter perturbations. Specifically, by adding small i.i.d. Gaussian perturbations $\delta \sim \mathcal{N}(0, \sigma^2 I_d)$ to the model parameters and measuring the resulting loss changes on the forget set, we can recover an upper bound on $\frac{1}{|\mathcal{R}|}\sum_i \|\nabla \mathcal{L}_i\|^2$. Theorem 2 formalizes this bound.

Theorem 2 (Upper Bound on the Retain Gradient Norm)**.** *Under Assumption 1 and Assumption 2, we have*

$$\frac{1}{|\mathcal{R}|}\sum_{i\in\mathcal{R}} \|\nabla \mathcal{L}_i\|^2 \le 2C^2\big(\|G_f\|^2 + V_f^2\big) + \frac{(C+1)^2 L^2}{2}\sigma^2(d^2+2d). \tag{11}$$

where $V_f^2 = \frac{1}{|\mathcal{F}|}\sum_j \|\nabla \mathcal{L}_j - G_f\|^2$ *is the variance of the forgotten sample gradients.*

The quantity V_f^2 in Theorem 2 can also be estimated via perturbation observations. Define the second moment of loss changes for each forgotten sample $M_j^{(2)} = \frac{1}{K}\sum_k Y_{jk}^2$, and construct

$$\widehat{V_f^2} = \frac{1}{\sigma^2}\Big(\frac{1}{|\mathcal{F}|}\sum_j M_j^{(2)} - M_f^{(2)}\Big). \tag{12}$$

Because $\mathbb{E}[\widehat{V_f^2}] = V_f^2 + O(\sigma^2)$, so $\widehat{V_f^2}$ is an approximately unbiased estimator of V_f^2. Substituting $\widehat{V_f^2}$ and $\widehat{\|G_f\|^2}$ into Theorem 2 yields a computable upper bound on the mean squared norm of the retain gradients.

4.3.2 Step-size Condition in Source-free Scenarios

With an upper bound on the mean squared norm, we can derive a lower bound on their projections onto a candidate direction u, i.e., a key quantity required in the expression for $t_i(u)$. To achieve this, we rely on Assumption 3, which ensures that the minimal positive projection among retained samples is at least a fraction α of the root-mean-square gradient norm. Combining it with the upper bound from Theorem 2, we obtain a conservative estimate for $\nabla \mathcal{L}_i^\top u$ and subsequently a lower bound on the retention threshold step size $\min_i t_i(u)$.

Specifically, combining Theorem 2, we obtain a lower bound estimate for $\nabla \mathcal{L}_i^\top u$:

$$\nabla \mathcal{L}_i^\top u \ge \alpha\big(G_{\text{est}} + \sqrt{\epsilon_0}\big), \tag{13}$$

where

$$G_{\text{est}} = \frac{\sqrt{2}C}{\sigma}\sqrt{M_f^{(2)}}, \quad \epsilon_0 = 2C^2\widehat{V_f^2} + \frac{(C+1)^2L^2}{2}\sigma^2(d^2+2d). \tag{14}$$

Furthermore, the step-size condition also requires the threshold margins Δ_i of the retained samples. In the source-free scenario, we assume a known conservative lower bound $\Delta_{\min} = \min_{i\in\mathcal{R}} \Delta_i$ (which can be obtained from domain knowledge or a small validation set). Since $t_i(u)$ is monotonically increasing in $\nabla \mathcal{L}_i^\top u$ and also monotonically increasing in Δ_i, substituting the above lower bound into the expression for $t_i(u)$ gives a lower bound estimate for the retention threshold step size.

Theorem 3 (Lower Bound on the Retention Threshold Step Size)**.** *Under Assumption 1, Assumption 2, and Assumption 3, let* $\hat{A} = \alpha(G_{est} + \sqrt{\epsilon_0})$*. If* $\hat{A}^2 \ge 2L\Delta_{\min}$*, then*

$$\min_{i:\nabla \mathcal{L}_i^\top u>0} t_i(u) \ge \hat{t}_{\min}(u) := \frac{\hat{A} - \sqrt{\hat{A}^2 - 2L\Delta_{\min}}}{L}. \tag{15}$$

The required step size for the forgetting objective, $\max_j t_j(u)$, can be directly computed using the estimated $\widehat{\nabla \mathcal{L}_j^\top u}$ (e.g., via the first moment $\frac{1}{K}\sum_k Y_{jk}(u^\top \delta_k)/\sigma^2$). Finally, the step-size condition in source-free scenarios is:

$$\max_{j\in\mathcal{F}} \widehat{t_j(u)} \le \hat{t}_{\min}(u). \tag{16}$$

Our theory provides a concise two-step diagnosis for assessing the feasibility of unlearning under the source-free scenarios:

- **Direction Condition Check**: Construct a common direction u (e.g., the average gradient direction) such that the projections of all forget gradients are positive (Eq. 7); if no such u exists, one may sacrifice part of the retention and forgetting objective.
- **Step-size Condition Estimation**: Observe loss changes on the forget set via random perturbations, estimate a lower bound on the retain gradient projections, and then compute the lower bound $\hat{t}_{\min}(u)$ (Eq. 15) of the threshold step size for retention as well as the threshold step size $\max_j \widehat{t_j(u)}$ (Eq. 8) for forgetting. If $\max_j \widehat{t_j(u)} \leq \hat{t}_{\min}(u)$, choose the maximum step size and perform the update $\theta_{\text{new}} = \theta_0 + tu$ to improve unlearning efficiency; otherwise, we need to relax the retention or forgetting objective.

5 Experiments

In this section, we conduct comprehensive experiments to evaluate the effectiveness of our proposed theoretically guided unlearning framework. We focus on three key aspects: (i) unlearning efficacy measured by forget accuracy and privacy leakage; (ii) performance maintenance on the retain set; (iii) the practicality under source-free scenarios. We compare our method against several baselines on three standard image classification datasets. All experiments were conducted on 15 vCPU Intel(R) Xeon(R) Platinum 8474C, NVIDIA RTX 4090D, and 80GB RAM within 30 minutes.

5.1 Experiment Setting

Dataset. We use three standard image classification datasets: CIFAR-10 [27], CIFAR-100 [27], and [10]. For each dataset, the forget set is constructed in three ways: (i) class-wise forgetting, i.e., selecting all samples from a specific class whose dot product with the class-average gradient direction is positive; (ii) common direction forgetting screening samples with positive gradient projection via random direction search, where only the direction condition is satisfied while the step-size condition typically does not hold; (iii) random forgetting randomly sampling 10% of the training samples without enforcing the direction condition. The retain set consists of the remaining training samples after removing the forget set.

Models. Two network architectures are adopted for the datasets. For CIFAR-10 and MNIST, a CNN is used, which consists of three convolutional layers, an adaptive pooling layer, and two fully connected layers. We use an adapted version of ResNet18 [22] on CIFAR-100, with the first convolutional layer modified to accommodate 32Œ32 inputs and the initial max-pooling layer removed. See parameters in Appendix D.

Baseline. We compare several unlearning methods: (i) Retrain as the exact unlearning method; (ii) NGDiff, which actively uses direction with a fixed learning rate; (iii) NGDiff+SF, which adopts theoretically guided direction and step-size conditions in a source-free scenario; (iv) Selective Fine-tuning with Targeted Confusion (SFTC) [41] based on a teacherstudent framework, where the retain set is finetuned and the forget set is confused via KL divergence using a biased random generator; (v) SourceFree Hessian Estimation (SFU) [2], which uses only the forget set and the original model to estimate the diagonal of the Hessian of the retain set and performs a Newton update. Basaran et al. [4] rely on a surrogate dataset, while Chen et al. [9] address diffusion models. The problem settings and technical approaches of both works are inconsistent with our strictly sourcefree, discriminative model setting, so they are not included as baselines for comparison.

Evaluation metrics. We evaluate the unlearning effect from multiple dimensions: (i) Forget Accuracy (FA) measures the models performance on the forget set (lower is better); (ii) Retain Accuracy (RA) measures the models performance on the retain set (higher is better). For privacy protection, we employ Membership Inference Attack (MIA) and report AUC and accuracy, where lower values indicate better privacy protection. For representation similarity, we compute Centered Kernel Alignment (CKA) [49], where higher values indicate more similar representations. We use the Information Difference Index (IDI) [24], which quantifies the residual information about the forget set in intermediate features, where a value closer to 0 indicates more similarity to the retrained model.

5.2 Experiment Results

5.2.1 Overall Results

Table 1 reports the performance of all methods across the three datasets under the class-wise forgetting setting where both geometric conditions are satisfied. From the results, we make the following observations:

Theoretical guidance significantly improves unlearning efficacy. Under the joint guidance of the direction and step-size conditions, most metrics improve after extending GA and NGDiff to source- free scenarios, achieving substantially lower FA than their unguided counterparts (see Figure 2). In particular, the improvement brought by theoretical guidance is most pronounced for GA-based methods, with FA reduced by up to 44.86% on CIFAR-10.

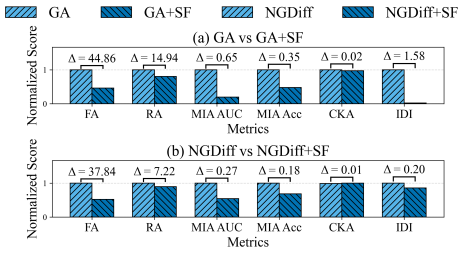


Figure 2: The metrics after extending GA and NGDiff to source-free scenarios.

Residual information is effectively removed, enhancing privacy. Metrics such as IDI and MIA show that after successful unlearning, our theoretically guided algorithms significantly reduce the residual information of forgotten samples within the model. For instance, on CIFAR-100, GA+SF achieves an IDI of -0.04 (close to 0) and an MIA AUC of 0.16, indicating near-complete removal of forget-set information.

Table 1: Experimental results on the forget set that satisfies two conditions.

Metric	Retrain	GA	GA+SF	NGDiff	NGDiff+SF	SFTC	SFU
CIFAR-10							
FA(%)	78.40	50.91	18.55	24.98	20.86	95.96	0.00
RA (%)	89.57	75.14	68.27	76.72	69.20	66.74	56.34
MIA AUC	–	0.11	0.14	0.15	0.12	0.15	0.21
MIA Acc (%)	–	24.79	27.54	27.37	26.34	25.43	29.35
CKA Avg	–	0.83	0.82	0.83	0.83	0.84	0.83
IDI	–	0.93	0.62	1.07	0.80	0.85	1.39
CIFAR-100							
FA (%)	86.49	83.24	38.38	79.46	41.62	91.35	0.00
RA (%)	99.89	75.09	60.15	70.75	63.53	11.92	1.00
MIA AUC	–	0.81	0.16	0.59	0.32	0.25	0.61
MIA Acc (%)	–	67.30	31.89	57.30	38.65	32.16	58.92
CKA Avg	–	0.77	0.75	0.75	0.76	0.65	0.78
IDI	–	1.62	-0.04	1.43	1.23	-1.79	1.41
MNIST							
FA (%)	39.62	27.16	14.04	0.00	17.43	99.63	80.94
RA (%)	99.19	94.77	95.54	64.22	95.03	98.70	98.43
MIA AUC	–	0.05	0.05	0.14	0.05	0.05	0.05
MIA Acc (%)	–	7.42	6.55	24.20	7.18	6.45	6.98
CKA Avg	–	0.97	0.97	0.90	0.97	0.97	0.97
IDI	–	0.99	0.98	0.99	0.99	1.00	0.96

5.2.2 Ablation Experiment

We conduct an ablation experiment to isolate the contribution of the direction condition and the step-size condition. Figures 3 and 4 show the following: (i) Under both conditions satisfied, the unlearning algorithm can effectively reduce FA, which decreases 88.5%. At the same time, although RA declines, the drop remains within a controllable range, and the model does not collapse. (ii) Only the directional condition is satisfied, the update direction is correct, but the excessive step size leads to drastic parameter updates: FA drops to 0, and RA also deteriorates severely (decrease

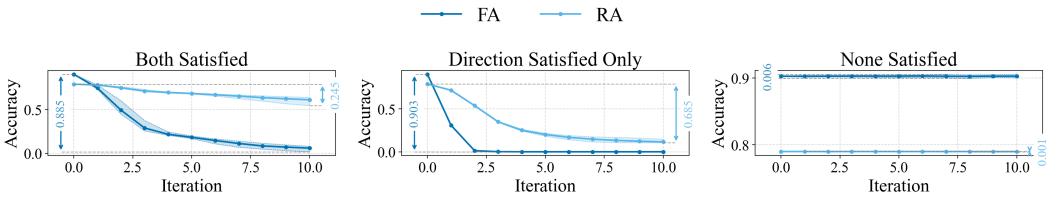


Figure 3: Unlearning process under different conditions.

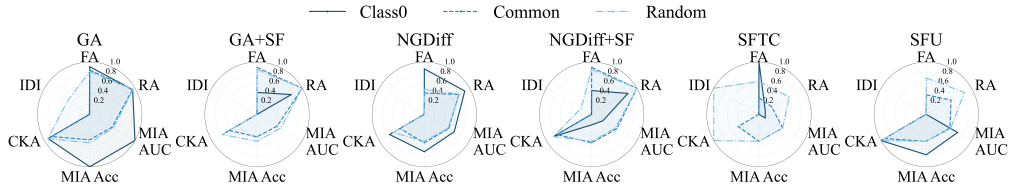


Figure 4: The metrics across different conditions in the CIFAR-100 ablation experiment.

68.5%), rendering the model nearly unusable. (iii) In random unlearning, the algorithm completely fails to drive the unlearning process: FA shows almost no change, and RA also remains essentially unchanged. In summary, both the directional condition and the step-size condition are indispensable; only when both are satisfied can stable and effective machine unlearning be achieved.

5.2.3 Parameter Sensitivity Analysis

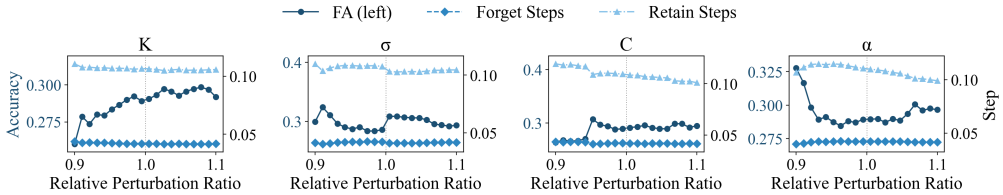


Figure 5: Sensitivity of parameter.

We examine the sensitivity of our method to four key parameters: the number of perturbations K, the perturbation scale σ, the local correlation constant C (Assumption 2), and the gradient consistency factor α (Assumption 3). Figure 5 summarizes the results after smoothing: (i) The proposed method exhibits high robustness in terms of the retention step size. It remains stable around the baseline level regardless of which parameter (K, σ, C, or α) is adjusted. (ii) The forget step size is relatively insensitive to changes in K and σ; however, as C and α increase, the forget step size exhibits a gradually decreasing trend. (iii) FA is more sensitive to parameter variations, and the degrees of sensitivity differ significantly across parameters. Specifically, FA increases with K; as σ increases, FA exhibits slight fluctuations; as α increases, FA exhibits a V-shaped pattern and attains smaller values near the baseline; while as C increases slowly, FA shows a gradual upward trend.

5.3 Extension to Generative Models

Although the theory in Section 4 is presented for discriminative models, the direction and step-size conditions rely only on the smoothness of the loss function (Assumption 1) and do not require class-probability outputs. Hence the framework can be directly extended to generative models by defining suitable differentiable forgetting and retention losses, e.g., reconstruction error or negative log-likelihood. We validate this on VAE-MNIST. The forget set is digit "5" and the retain set contains the other nine digits. We set $\sigma = 1.5 \times 10^{-4}$, $K = 200$, $C = 1.1$, $\alpha = 0.1$, and use multi-step iteration up to 20 steps.

Table 2 reports the results. A single-step update already reduces the recognition rate for the forgotten class from 90.7% to 67.05%, and multi-step iteration further lowers it to 22.12%, while the

Table 2: Unlearning, retain performance, and generation quality on VAE-MNIST.

Metric	Before	Single step	Multi-step
Unlearning effect			
Reconstruction loss for "5"	115.40	2493.49	2661.85
Recognition rate for "5"	90.7%	67.05%	22.12%
Retain performance and generation quality			
Avg. reconstruction loss (non-forgotten)	101.66	–	1483.28
Forgotten class generation rate	10%	–	8%
Latent FD	0.19	–	0.27

reconstruction loss for the forgotten class increases substantially, indicating that the forgetting effect accumulates with steps. Unlike discriminative models, the evaluation here also involves the generated distribution. Nevertheless, the generated class distribution remains uniform and the Latent FD stays stable at 0.27, showing that the generative capacity has not collapsed. These results confirm that the proposed geometric framework is also effective for generative models.

6 Conclusion

In this paper, we revisit GA-based unlearning methods and address two key limitations: lack of feasibility assessment and reliance on the retain set. By analyzing the geometric structure of gradient space, we derive sufficient conditions for unlearning feasibility (including both direction and step-size conditions) and extend our framework to sourcefree scenarios via Gaussian perturbations. Our geometric framework can serve as a plugin component for existing unlearning algorithms to achieve more complete forgetting. Experiments demonstrate that our approach enables effective unlearning while preserving model performance on the retain set. Our geometric perspective provides a principled theoretical foundation for deploying machine unlearning in practice, covering both traditional and source-free scenarios.

7 Broader Impacts and Limitations

Our work helps meet compliance requirements such as the right to be forgotten and reduces reliance on the retain set. However, unlearning techniques may be misused to evade audits, and insufficient verification may also harm model fairness and robustness. Theoretically, our paper guarantees only the feasibility of a single-step, local, fixed-direction update and does not guarantee the effectiveness of multi-step iteration; the source-free extension relies on Gaussian perturbations to estimate retain gradients, whose estimation variance and computational overhead are significant in high-dimensional large models, and it requires a known lower bound on the retention tolerance, so it remains limited in scenarios such as large-scale and online unlearning.

Acknowledgments and Disclosure of Funding

This work was partly supported by the NSFC under No. 62402418, the "Pioneer and Leading Goose" R&D Program of Zhejiang under No. 2026C02A1233 and 2025C02034, and the Ningbo Yongjiang Talent Project.

References

[1] Sk Miraj Ahmed, Umit Yigit Basaran, Dripta S. Raychaudhuri, Arindam Dutta, Rohit Kundu, Fahim Faisal Niloy, Basak Guler, and Amit K. Roy-Chowdhury. Towards source-free machine unlearning, 2025.

[2] Sk Miraj Ahmed, Umit Yigit Basaran, Dripta S. Raychaudhuri, Arindam Dutta, Rohit Kundu, Fahim Faisal Niloy, Basak Guler, and Amit K. Roy-Chowdhury. Towards Source-Free Machine Unlearning, August 2025. arXiv:2508.15127 [cs].

[3] Sadia Asif and Mohammad Mohammadi Amiri. Ofmu: Optimization-driven framework for machine unlearning, 2026.

[4] Umit Yigit Basaran, Sk. Miraj Ahmed, Amit K. Roy-Chowdhury, and Basak Guler. A certified unlearning approach without access to source data. *ArXiv*, abs/2506.06486, 2025.

[5] Samyadeep Basu, Phil Pope, and Soheil Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations*, 2021.

[6] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159, 2019.

[7] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159, 2021.

[8] Chaochao Chen, Yizhao Zhang, Yuyuan Li, Jun Wang, Lianyong Qi, Xiaolong Xu, Xiaolin Zheng, and Jianwei Yin. Post-training attribute unlearning in recommender systems. *ACM Transactions on Information Systems*, 43(1):1–28, 2024.

[9] Tianqi Chen, Shujian Zhang, and Mingyuan Zhou. Score forgetting distillation: A swift, data-free method for machine unlearning in diffusion models, 2025.

[10] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[11] Zonglin Di, Zhaowei Zhu, Jinghan Jia, Jiancheng Liu, Zafar Takhirov, Bo Jiang, Yuanshun Yao, Sijia Liu, and Yang Liu. Label smoothing improves machine unlearning, 2024.

[12] Meng Ding, Rohan Sharma, Changyou Chen, Jinhui Xu, and Kaiyi Ji. Understanding fine-tuning in approximate unlearning: A theoretical perspective, 2025.

[13] Taha Entesari, Arman Hatami, Rinat Khaziev, Anil Ramakrishna, and Mahyar Fazlyab. Constrained entropic unlearning: A primal-dual framework for large language models, 2025.

[14] Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning, 2025.

[15] Xiaohua Feng, Yuyuan Li, Huwei Ji, Jiaming Zhang, Li Zhang, Tianyu Du, and Chaochao Chen. Leveraging machine unlearning for cost-efficient preference alignment, 2026.

[16] Xiaohua Feng, Jiaming Zhang, Fengyuan Yu, Chengye Wang, Li Zhang, Kaixiang Li, Yuyuan Li, Chaochao Chen, and Jianwei Yin. A survey on generative model unlearning: Fundamentals, taxonomy, evaluation, and future direction. *arXiv preprint arXiv:2507.19894*, 2025.

[17] Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast Machine Unlearning without Retraining through Selective Synaptic Dampening. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12043–12051, March 2024.

[18] Jiawei Ge, Jiuxin Cao, Xiangmei Chen, Xuelin Zhu, Weijia Liu, Chang Liu, Kun Wang, and Bo Liu. Beyond visual cues: Synchronously exploring target-centric semantics for vision-language tracking. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(5):1–21, 2025.

[19] Jiawei Ge, Xinyu Zhang, Jiuxin Cao, Xuelin Zhu, Weijia Liu, Qingqing Gao, Biwei Cao, Kun Wang, Chang Liu, Bo Liu, et al. Gen4track: A tuning-free data augmentation framework via self-correcting diffusion model for vision-language tracking. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3037–3046, 2025.

[20] Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. *CoRR*, abs/2010.10981, 2020.

[21] Preethi Gurumurthy and P.K. Srijith. Pseudo-inverse prefix tuning for effective unlearning in llms. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25, page 47434747, New York, NY, USA, 2025. Association for Computing Machinery.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[23] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models, 2023.

[24] Dongjae Jeon, Wonje Jeung, Taeheon Kim, Albert No, and Jonghyun Choi. An Information Theoretic Evaluation Metric For Strong Unlearning, November 2025. arXiv:2405.17878 [cs].

[25] Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning, 2024.

[26] Xiaomeng Jin, Zhiqi Bu, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, and Mingyi Hong. Unlearning as multi-task optimization: A normalized gradient difference approach with an adaptive learning rate. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Albuquerque, New Mexico, April 2025.

[27] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[28] Ranjit Kumar, Ravi Prakash, Narendra Kumar, Suchismita Chinara, and Mazin Abed Mohammed. Machine unlearning for trustworthy AI: A systematic review of techniques, challenges, and applications. *Arch. Comput. Methods Eng.*, October 2025.

[29] {Varun Sampath} Kumar, Esmaeil Nadimi, and {Vinay Chakravarthi} Gogineni. Rklu: Redistributive kl distillation for efficient retain-free machine unlearning. In *ESANN 2026 Proceedings - 34th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 41–46. i6doc.com, 2026.

[30] Jiaqi Li, Chuanyi Zhang, Miaozeng Du, Hui Zhang, Yongrui Chen, Qianshan Wei, Junfeng Fang, Ruipeng Wang, Sheng Bi, and Guilin Qi. Forget the token and pixel: Rethinking gradient ascent for concept unlearning in multimodal generative models. In *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria, July 2025. Association for Computational Linguistics.

[31] Na Li, Chunyi Zhou, Yansong Gao, Hui Chen, Anmin Fu, Zhi Zhang, and Yu Shui. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects, 2024.

[32] Yuyuan Li, Chaochao Chen, Yizhao Zhang, Weiming Liu, Lingjuan Lyu, Xiaolin Zheng, Dan Meng, and Jun Wang. Ultrare: Enhancing receraser for recommendation unlearning via error decomposition. *Advances in Neural Information Processing Systems*, 36:12611–12625, 2023.

[33] Yuyuan Li, Xiaohua Feng, Chaochao Chen, and Qiang Yang. A survey on recommendation unlearning: Fundamentals, taxonomy, evaluation, and open questions. *IEEE Transactions on Knowledge and Data Engineering*, 38(2):781–799, 2025.

[34] Yuyuan Li, Yizhao Zhang, Weiming Liu, Xiaohua Feng, Zhongxuan Han, Chaochao Chen, and Chenggang Yan. Multi-objective unlearning in recommender systems via preference guided pareto exploration. *IEEE Transactions on Services Computing*, 2025.

[35] Junxu Liu, Mingsheng Xue, Jian Lou, Xiaoyu Zhang, Li Xiong, and Zhan Qin. Muter: Machine unlearning on adversarially trained models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4869–4879, 2023.

[36] Shuo Lu, Haohan Wang, Wei Feng, Weizhen Wang, Shen Zhang, Yaoyu Li, Ao Ma, Zheng Zhang, Jingjing Lv, Junjie Shen, et al. One size, many fits: Aligning diverse group-wise click preferences in large-scale advertising image generation. In *Proc. CIKM*, 2026.

[37] Ioannis Mavrothalassitis, Pol Puigdemont, Noam Itzhak Levi, and Volkan Cevher. Ascent fails to forget. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[38] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Trans. Intell. Syst. Technol.*, 16(5), September 2025.

[39] Zirui Pang, Hao Zheng, Zhijie Deng, Ling Li, Zixin Zhong, and Jiaheng Wei. Label smoothing improves gradient ascent in llm unlearning, 2025.

[40] Zirui Pang, Hao Zheng, Zhijie Deng, Ling Li, Zixin Zhong, and Jiaheng Wei. Label smoothing improves gradient ascent in LLM unlearning, 2026.

[41] Vasileios Perifanis, Efstathios Karypidis, Nikos Komodakis, and Pavlos Efraimidis. SFTC: Machine Unlearning via Selective Fine-tuning and Targeted Confusion. In *European Interdisciplinary Cybersecurity Conference*, pages 29–36, Xanthi Greece, June 2024. ACM.

[42] Hadi Reisizadeh, Jinghan Jia, Zhiqi Bu, Bhanukiran Vinzamuri, Anil Ramakrishna, Kai-Wei Chang, Volkan Cevher, Sijia Liu, and Mingyi Hong. Blur: A bi-level optimization approach for llm unlearning, 2025.

[43] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *CoRR*, abs/2103.03279, 2021.

[44] Md Toufikuzzaman, Ahmad Mousavi, and Dongwon Lee. Blade: Bilevel low-rank augmented-lagrangian erasure for llm unlearning, 2026.

[45] Martin Van Waerebeke, Marco Lorenzi, Giovanni Neglia, and Kevin Scaman. When to forget? complexity trade-offs in machine unlearning. In *Forty-second International Conference on Machine Learning*, 2025.

[46] Weiqi Wang, Zhiyi Tian, Chenhan Zhang, and Shui Yu. Machine unlearning: A comprehensive survey, 2024.

[47] Zhihao Wang, Bingqian Liu, Jian Wang, and Bing Li. Enhancing intent understanding and preference learning for sequential recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 38(5):2762–2778, 2026.

[48] Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(3):2150–2168, 2024.

[49] Xiaoyu Xu, Xiang Yue, Yang Liu, Qingqing Ye, Huadi Zheng, Peizhao Hu, Minxin Du, and Haibo Hu. Unlearning isn't deletion: Investigating reversibility of machine unlearning in llms, 2025.

[50] David Zagardo. A more practical approach to machine unlearning. *ArXiv*, abs/2406.09391, 2024.

[51] Binchi Zhang, Yushun Dong, Tianhao Wang, and Jundong Li. Towards Certified Unlearning for Deep Neural Networks, May 2025.

[52] Xuyang Zhong, Haochen Luo, and Chen Liu. Dualoptim: Enhancing efficacy and stability in machine unlearning with dual optimizers, 2025.

[53] Shiji Zhou, Tianbai Yu, Zhi Zhang, Heng Chang, Xiao Zhou, Dong-Dong Wu, and Han Zhao. Efficient utility-preserving machine unlearning with implicit gradient surgery. *ArXiv*, abs/2510.22124, 2025.

A Conditions for Local Approximate Unlearning

A.1 Definitions

Model parameters. Let θ_0 be the model parameters before unlearning, and θ the desired model parameters after unlearning.

Performance function. Define a function $\mathcal{L}_i : \Theta \to \mathbb{R}$ for every sample $i \in \mathcal{D}$, such that $\mathcal{L}_i(\theta_0) \leqslant \epsilon_i$ where $\epsilon_i \geqslant 0$ is a preset tolerance representing the training error.

Forget set and retain set. Let the forget set be $\mathcal{F} \subseteq \mathcal{D}$ and the retain set be $\mathcal{R} \subseteq \mathcal{D} \backslash \mathcal{F}$.

Unlearning threshold. For a forgotten sample $j \in \mathcal{F}$, we require its performance on the model to degrade, i.e., $\mathcal{L}_j(\theta)$ becomes larger. Hence we set a unlearning threshold τ_j and require $\mathcal{L}_j(\theta) > \tau_j$ with $\tau_j > \mathcal{L}_j(\theta_0)$.

Retaining requirement. For the retain set, we require the model performance to remain unchanged, i.e., $\mathcal{L}_i(\theta) \leqslant \epsilon_i$.

A.2 Local Linearized Directional Conditions

By the Assumption 1, for any small δ we have the Taylor expansion with remainder

$$\mathcal{L}_i(\theta_0 + \delta) = \mathcal{L}_i(\theta_0) + \nabla \mathcal{L}_i(\theta_0)^\top \delta + R_i(\delta), \quad |R_i(\delta)| \leq \frac{L}{2}\|\delta\|^2. \tag{17}$$

Retain sample constraints. For a retain sample i, we require $\mathcal{L}_i(\theta_0 + \delta) \leq \epsilon_i$, which gives

$$\nabla \mathcal{L}_i(\theta_0)^\top \delta \leq \Delta_i + \frac{L}{2}\|\delta\|^2 \tag{18}$$

where $\Delta_i = \epsilon_i - \mathcal{L}_i(\theta_0) \geq 0$.

Forgotten sample constraints. For a forgotten sample $j \in \mathcal{F}$, we require $\mathcal{L}_j(\theta_0 + \delta) \geq \tau_j$, i.e.,

$$\nabla \mathcal{L}_j(\theta_0)^\top \delta \geq \Delta_j - \frac{L}{2}\|\delta\|^2, \tag{19}$$

where $\Delta_j = \tau_j - \mathcal{L}_j(\theta_0) > 0$.

Reduction to directional conditions. Let $\delta = tu$ with $t = \|\delta\| > 0$ and u a unit vector. Substituting into (R) and (F) yields

$$t\nabla \mathcal{L}_i(\theta_0)^\top u \leq \Delta_i + \frac{L}{2}t^2, \quad t\nabla \mathcal{L}_j(\theta_0)^\top u \geq \Delta_j - \frac{L}{2}t^2. \tag{20}$$

The sufficient conditions for the existence of $t > 0$ and a unit vector u satisfying $(*)$ are:

1. **Directional condition**: $\nabla \mathcal{L}_j(\theta_0)^\top u > 0$ for all $j \in \mathcal{F}$.
2. **Step-size condition**: $\max_{j \in \mathcal{F}} t_j(u) \leq \min_{i:\nabla \mathcal{L}_i(\theta_0)^\top u > 0} t_i(u)$, where

$$t_j(u) = \frac{-\nabla \mathcal{L}_j(\theta_0)^\top u + \sqrt{(\nabla \mathcal{L}_j(\theta_0)^\top u)^2 + 2L\Delta_j}}{L}, \tag{21}$$

$$t_i(u) = \frac{\nabla \mathcal{L}_i(\theta_0)^\top u - \sqrt{(\nabla \mathcal{L}_i(\theta_0)^\top u)^2 - 2L\Delta_i}}{L}. \tag{22}$$

A.3 Existence Condition for Local Approximate Unlearning

Under the Assumption 1, if either of the following holds, then there exists no small perturbation $\delta = tu$ that simultaneously satisfies all unlearning and retaining constraints:

1. $\nabla \mathcal{L}_j(\theta_0)^\top u > 0, \quad \forall j \in \mathcal{F}$.(directional impossibility);
2. There exists a u satisfying the directional condition such that $\max_j t_j(u) > \min_i t_i(u)$ (step-size conflict).

B Step-Size Computation in the Source-Free Scenario

B.1 Problem Setup and Recap of Known Results

We adopt the notation from Appendix A. Define $\Delta_j = \tau_j - \mathcal{L}_j(\theta_0) > 0$ and $\Delta_i = \epsilon_i - \mathcal{L}_i(\theta_0) \geq 0$. The minimum step needed for unlearning and the upper bound on the tolerable step for retention are respectively:

$$t_j(u) = \frac{-\nabla\mathcal{L}_j(\theta_0)^\top u + \sqrt{(\nabla\mathcal{L}_j(\theta_0)^\top u)^2 + 2L\Delta_j}}{L}, \quad j \in \mathcal{F}, \tag{23}$$

$$t_i(u) = \frac{\nabla\mathcal{L}_i(\theta_0)^\top u - \sqrt{(\nabla\mathcal{L}_i(\theta_0)^\top u)^2 - 2L\Delta_i}}{L}, \quad i \in \mathcal{R},\ \nabla\mathcal{L}_i(\theta_0)^\top u > 0. \tag{24}$$

The deterministic existence theorem (Appendix A) states that a small perturbation $\delta = tu$ satisfying both unlearning and retaining constraints exists iff:

1. There exists a unit direction u such that $\nabla\mathcal{L}_j(\theta_0)^\top u > 0$ for all $j \in \mathcal{F}$;
2. $\max_{j\in\mathcal{F}} t_j(u) \leq \min_{i:\nabla\mathcal{L}_i(\theta_0)^\top u>0} t_i(u)$.

In the source-free scenario, $\nabla\mathcal{L}_i(\theta_0)$ and $t_i(u)$ are unknown and can only be estimated via perturbation observations.

B.2 Estimation of the Unlearning Average Gradient Norm

Perturbation observations. Add K i.i.d. small perturbations $\delta_k \sim \mathcal{N}(0, \sigma^2 I_d)$ to the model parameters, with $\sigma > 0$ small. Record the loss change for each forgotten sample:

$$Y_{jk} := \mathcal{L}_j(\theta_0 + \delta_k) - \mathcal{L}_j(\theta_0), \quad j \in \mathcal{F},\ k = 1, \ldots, K. \tag{25}$$

The average loss change over forgotten samples is

$$Y_k := \frac{1}{|\mathcal{F}|}\sum_{j\in\mathcal{F}} Y_{jk}. \tag{26}$$

By the Assumption 1,

$$Y_k = G_f^\top \delta_k + R_k, \quad G_f := \frac{1}{|\mathcal{F}|}\sum_{j\in\mathcal{F}} \nabla\mathcal{L}_j(\theta_0), \quad |R_k| \leq \frac{L}{2}\|\delta_k\|^2. \tag{27}$$

Estimation of the unlearning average gradient norm. Define the sample second moment $M_f^{(2)} := \frac{1}{K}\sum_{k=1}^K Y_k^2$. Its expectation is

$$\mathbb{E}[M_f^{(2)}] = \sigma^2\|G_f\|^2 + \mathbb{E}[R_k^2] \leq \sigma^2\|G_f\|^2 + \frac{L^2}{4}\sigma^4 d(d+2). \tag{28}$$

Choosing σ sufficiently small (e.g., $\sigma = O(K^{-1/4})$) gives an approximately unbiased estimate:

$$\|G_f\|^2 \approx \frac{M_f^{(2)}}{\sigma^2}. \tag{29}$$

B.3 Upper Bound for the Retaining Gradient Second Moment

Deriving the upper bound. Under Assumption 2, using the triangle inequality and the boundedness of the remainder, we obtain

$$\frac{1}{|\mathcal{R}|}\sum_i |\nabla\mathcal{L}_i^\top \delta| \leq C \cdot \frac{1}{|\mathcal{F}|}\sum_j |\nabla\mathcal{L}_j^\top \delta| + (C+1)\frac{L}{2}\|\delta\|^2. \tag{30}$$

Squaring both sides and taking expectation over $\delta \sim \mathcal{N}(0, \sigma^2 I)$, then applying the CauchySchwarz inequality, yields

$$\frac{1}{|\mathcal{R}|}\sum_i \|\nabla\mathcal{L}_i\|^2 \leq 2C^2 \cdot \frac{1}{|\mathcal{F}|}\sum_j \|\nabla\mathcal{L}_j\|^2 + \frac{(C+1)^2 L^2}{2}\sigma^2(d^2+2d). \tag{31}$$

Define $V_f^2 := \frac{1}{|\mathcal{F}|}\sum_j \|\nabla \mathcal{L}_j\|^2 - \|G_f\|^2$. Then

$$\frac{1}{|\mathcal{R}|}\sum_i \|\nabla \mathcal{L}_i\|^2 \le 2C^2\|G_f\|^2 + \epsilon_0, \tag{32}$$

where $\epsilon_0 = 2C^2V_f^2 + \frac{(C+1)^2L^2}{2}\sigma^2(d^2+2d)$. Substituting the estimate of $\|G_f\|^2$ gives

$$\frac{1}{|\mathcal{R}|}\sum_i \|\nabla \mathcal{L}_i\|^2 \le \frac{2C^2M_f^{(2)}}{\sigma^2} + \epsilon_0. \tag{33}$$

B.4 Lower Bound for the Retaining Tolerable Step

Set $G_{\text{est}} := \frac{\sqrt{2}C}{\sigma}\sqrt{M_f^{(2)}}$ and $\epsilon_1 = \sqrt{\epsilon_0}$. Then

$$\sqrt{\frac{1}{|\mathcal{R}|}\sum_i \|\nabla \mathcal{L}_i\|^2} \le G_{\text{est}} + \epsilon_1. \tag{34}$$

Combining with Assumption 3 gives

$$\nabla \mathcal{L}_i^\top u \ge \alpha G_{\text{est}} + \alpha\epsilon_1, \quad \forall i \text{ with } \nabla \mathcal{L}_i^\top u > 0. \tag{35}$$

Let $\Delta_{\min} = \min_{i\in\mathcal{R}} \Delta_i$ be a known conservative lower bound on the retaining tolerances. Substituting into the expression for $t_i(u)$ yields the lower bound estimate:

$$\hat{t}_{\min}(u) = \frac{(\alpha G_{\text{est}} + \alpha\epsilon_1) - \sqrt{(\alpha G_{\text{est}} + \alpha\epsilon_1)^2 - 2L\Delta_{\min}}}{L}. \tag{36}$$

B.5 Estimation of the Unlearning Gradient Variance

The variance of the forgotten sample gradients is

$$V_f^2 := \frac{1}{|\mathcal{F}|}\sum_{j\in\mathcal{F}} \|\nabla \mathcal{L}_j\|^2 - \|G_f\|^2. \tag{37}$$

For each forgotten sample j, define the sample second moment $M_j^{(2)} := \frac{1}{K}\sum_{k=1}^K Y_{jk}^2$. Then

$$\mathbb{E}[M_j^{(2)}] = \sigma^2\|\nabla \mathcal{L}_j\|^2 + \mathbb{E}[R_{jk}^2] \le \sigma^2\|\nabla \mathcal{L}_j\|^2 + \frac{L^2}{4}\sigma^4 d(d+2). \tag{38}$$

Construct the estimator

$$\hat{V}_f^2 := \frac{1}{\sigma^2}\left(\frac{1}{|\mathcal{F}|}\sum_j M_j^{(2)} - M_f^{(2)}\right), \tag{39}$$

whose expectation is $V_f^2 + O(\sigma^2)$, i.e., approximately unbiased. There exists a constant c_2 such that with high probability,

$$\left|\hat{V}_f^2 - V_f^2\right| \le c_2\left(\frac{\sqrt{\log(1/\delta)}}{\sqrt{K}} + \sigma^2L^2d\right). \tag{40}$$

C Discussion and Limitations

C.1 Discussion on the Dependence on Problem Constants

In this section, we analyze the dependence of the lower bound $\hat{t}_{\min}$ in Theorem 3 on the problem constants d, C, and α.

Table 3: Influence of problem constants on the lower bound $\hat{t}_{\min}$.

Constant	Dependency	Intuitive meaning
d (model dimension)	$\epsilon_0 \propto \sigma^2 d^2$	The higher the dimension, the stronger the cumulative effect of parameter perturbations, the looser the estimated upper bound on the retain gradient, and thus the more conservative (smaller) $\hat{t}_{\min}$. A "safe step size" becomes harder to guarantee in high-dimensional spaces. However, if the model exhibits structural sparsity, we can replace d with the effective parameter dimension d_{eff} (e.g., the dimension of a sparse subspace), which significantly tightens the upper bound.
C (local correlation)	$\epsilon_0 \propto C^2$	C measures the degree of coupling between the local behaviours of the forget set and the retain set. A larger C implies lower confidence in inferring retain-set properties from the forget set (amplified uncertainty), and thus a more conservative step-size lower bound. When $C = 1$, the local behaviours of the two sets are perfectly synchronised, and the estimation is most accurate.
α (gradient consistency)	$\hat{A} \propto \alpha$	α measures the ratio of the minimum positive projection of the retain gradient onto direction u to its RMS norm. A larger α means a larger component of the retain gradient along u, and $\hat{t}_{\min}$ increases accordingly – the model can better "tolerate" updates along u. When $\alpha = 0$, no retain performance can be guaranteed.

C.2 Discussion on Conservativeness and Sparsity Potential

The retain gradient norm upper bound derived in Theorem 2 is inherently conservative in high-dimensional settings. The conservativeness of the bound can be substantially alleviated by exploiting model sparsity. If the model exhibits structural sparsity (i.e., only a small number of parameters significantly influence the unlearning/retaining objective), the parameters can be partitioned into an effective parameter space d_{eff} (nonzero parameters or dimensions that contribute to the objective), while the rest are fixed. By applying perturbations and updates within this effective subspace, all d-related terms in the theorem are replaced by d_{eff}. For high-dimensional sparse models where $d_{\text{eff}} \ll d$, the bound becomes significantly tighter. This approach introduces model structural priors to reduce estimation variance, aligning with the principles of "sparse unlearning" [25]. The theoretical extension and empirical validation of this sparse unlearning framework are left for future work.

D Parameter Setting

Table 4: Hyperparameter settings for each experiment.

Dataset	Forget Set Type	σ	K	C	α
CIFAR-10	common	1×10^{-3}	200	1.0	0.1
CIFAR-10	class0	1×10^{-3}	500	1.5	0.5
CIFAR-10	random	1×10^{-3}	200	1.0	0.1
CIFAR-100	common	1×10^{-4}	500	1.0	0.1
CIFAR-100	class0	1×10^{-4}	500	1.5	0.5
CIFAR-100	random	1×10^{-4}	500	1.0	0.1
MNIST	common	5×10^{-4}	100	1.0	0.1
MNIST	class0	5×10^{-4}	100	1.0	0.1
MNIST	random	5×10^{-4}	100	1.0	0.1

Table 5: Source of step size for different algorithms.

Algorithm	Source of step size
GA	Follows the setting in the original paper
NGDiff	Adopts the adaptive learning rate mechanism (GeN) from the original paper; no fixed step size needs to be preset
SFTC	Follows the setting in the original paper
SFU	Newton update, no step size hyperparameter

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contributions of this paper: diagnosing the instability of the GA method from a geometric perspective (direction conflict and step-size mismatch), deriving sufficient conditions for forgetting feasibility, and extending to source-free scenarios via Gaussian perturbation. These claims are supported by the methodological exposition in Sections 4.1–4.3, the theoretical analysis in Appendix A, and the experimental results in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix C.2 for limitations on multi-step updates and source-free estimation in large models.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The complete assumptions (Assumptions 1–3 in Section 3) and the proofs of all theorems (Theorems 1–3) are provided in Appendices A–B.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all necessary information to reproduce the main experimental results. Section 5.1 describes the datasets, model architectures, baseline methods, and evaluation metrics. The hyperparameters for each experiment (including σ, K, C, α) are provided in Appendix D. With this information, readers can reproduce the results even without access to the exact code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The core contribution of this work lies in the derivation of theoretical feasibility conditions and the proposal of an algorithmic method, which is methodological in nature rather than tied to a specific code implementation. The algorithm steps have been clearly described in Sections 4 and the Appendix A–B. Readers can independently implement the algorithm based on this description.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5.1 specifies the dataset splits, model architectures, baselines, and evaluation metrics. The hyperparameters for each experiment (including σ, K, C, α) are listed in Appendix D. These details are sufficient to understand the results.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The upper and lower bounds of accuracy during the unlearning process are shown in Figure 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 5 for experimental setup. All experiments were conducted on 15 vCPU Intel(R) Xeon(R) Platinum 8474C, NVIDIA RTX 4090D, and 80GB RAM within 30 minutes.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: This research aims to improve the security and privacy protection capabilities of machine unlearning, aligning with NeurIPS's ethical guidelines for promoting trustworthy AI. The experiments in Section 5 use publicly available image classification datasets (CIFAR-10, CIFAR-100, MNIST) and standard baseline methods, and do not involve human subjects, crowdsourced data, or high-risk privacy data release. The paper is submitted anonymously and contains no identifiable institutional or personal information. All methods are theory-driven model parameter updates and do not generate synthetic content or involve automated decision-making. Therefore, there is no violation of the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: See the Broader Impacts discussion in Appendix D, where both positive and negative societal impacts are explicitly analyzed.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets (CIFAR, MNIST, ResNet18, baseline methods) are properly cited with original papers. The datasets and models are used under permissive licenses (MIT, CC BY-SA 3.0).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.