Controlling Path Dependence in Gradient Ascent Unlearning through Forget-Set Ordering

Varun Sampath Kumar[1] **Esmaeil S Nadimi**[1] **Vinay Chakravarthi Gogineni**[1]

[1]Applied AI and Data Science Unit, The Maersk McKinney Moller Institute, University of Southern Denmark, Denmark

Abstract

Machine unlearning aims to selectively remove the influence of designated training data from a trained model. Gradient ascent-based unlearning methods are widely used in machine unlearning, but are often observed to exhibit instability and sensitivity to optimization hyperparameters. In this work, we show that the ordering of forget-set samples during ascent materially influences the resulting optimization trajectory, yielding markedly different forgetting–retention trade-offs under identical optimization budgets. We provide a local second-order analysis illustrating how non-commutativity of ascent updates induces path dependence, and we propose a simple ordering strategy that processes forget samples from low to high predictive uncertainty. By deferring high-uncertainty samples, this strategy attenuates early parameter drift and leads to more stable trajectories. Across image classification benchmarks with convolutional and vision transformer architectures, as well as text classification tasks, we observe that this ordering consistently achieves effective forgetting while better preserving retain-set utility compared to random ordering. In the retain-data-free regime, this trajectory-aware modification substantially strengthens vanilla gradient ascent, improving its position on the forgetting–retention Pareto frontier relative to naive ascent and several retain-data-free unlearning baselines. In general, our results identify sample ordering as a practically meaningful and previously underexplored degree of freedom in optimization-based unlearning, highlighting the importance of trajectory-aware design and evaluation.

1 INTRODUCTION

Modern machine learning models are routinely trained on large-scale datasets containing sensitive or user-generated information. After training, the influence of individual data points becomes implicitly encoded in the model parameters, rendering their targeted removal non-trivial without full retraining. The European Union's General Data Protection Regulation (GDPR) Mantelero [2013] has elevated the "right to be forgotten" to a legal requirement, compelling the effective erasure of user data from learned models. Unlike traditional databases, where records can be deleted explicitly, negating the contribution of specific training examples in neural networks while preserving performance on the remaining data is substantially more challenging. Machine unlearning has thus emerged as a vital research direction, seeking efficient methods to remove the influence of designated data from trained models without the prohibitive expense of retraining from scratch Xu et al. [2023], Bourtoule et al. [2021].

Most of the existing unlearning methods Golatkar et al. [2020], Kurmanji et al. [2023] rely on a post-unlearning fine-tuning stage to recover the model utility lost during the unlearning process. These approaches typically perform gradient ascent on the forget set over a small number of steps to diminish its influence, optionally followed by gradient descent on the retain set to mitigate any collateral degradation in utility. However, empirical evaluations reveal that such ascent-based techniques are highly sensitive to hyper-parameters including learning rate, batch size, optimizer choice, and number of steps, resulting in substantial variance in unlearning quality and retained performance. Recent analyses further demonstrate that these methods can behave unstably, particularly when the representations of forget samples are statistically entangled with those of the retain set Mavrothalassitis et al. [2025]. These findings motivate a more fundamental question: *to what extent is gradient ascent-based unlearning intrinsically well-defined?*

The order of the training samples significantly influences the

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:3224–3236.

optimization dynamics during learning. This effect has been well studied in the context of stochastic gradient descent through curriculum learning Bengio et al. [2009], where presenting examples from easy to hard typically accelerates convergence and improves generalization. In contrast, machine unlearning seeks an optimization trajectory that drives the model toward a parameter region where it retains high utility on the retain set while exhibiting poor performance (i.e., effective forgetting) on the forget set. Critically, this trajectory is shaped by intra-batch gradients during ascent updates, which contributes to the observed instability and high variance in gradient ascent-based unlearning methods. This fundamental distinction between the objectives and dynamics of learning (minimization) versus unlearning (maximization on forget data) motivates the exploration of mechanisms such as careful sample ordering during ascent that can guide the optimization toward more stable, controllable paths, thereby achieving better trade-offs between retained utility and effective forgetting.

In this work, we identify a critical and often underexplored degree of freedom in gradient ascent-based unlearning methods: the ordering of forget-set samples during ascent updates. Concretely, we show that while stochastic gradient descent is known to be path-dependent, in ascent-based unlearning trajectory effects play a dominant empirical role in the forgetting–retention trade-off. Under identical optimization budgets, the ordering of forget-set samples alone induces substantially different outcomes. More broadly, our findings advocate for a shift toward trajectory-aware design and evaluation of unlearning methods, moving beyond purely algorithm-centric perspectives to explicitly account for optimization dynamics and path sensitivity.

To address this path dependence, we propose a theoretically motivated sample-ordering strategy during gradient ascent that systematically alters the forgetting–retention trade-off under fixed optimization budgets. Specifically, ordering forget samples from low to high predictive uncertainty defers large-gradient updates, yielding more stable trajectories that better preserve retain-set utility, albeit requiring more ascent steps. In contrast, the reverse ordering induces aggressive forgetting with substantial collateral degradation. Random ordering exhibits higher variance and less predictable trade-offs. Across benchmarks and model scales, these results indicate that sample ordering serves as an interpretable and effective mechanism for controlling ascent-based unlearning dynamics.

Our main contributions are as follows.

1. We show that gradient ascent-based unlearning exhibits evident *path dependence*: under identical optimization budgets, the ordering of forget-set samples alone governs the forgetting vs retention trade-offs.
2. We provide a local second-order analysis and empirical evidence showing that deferring large-gradient samples leads to more stable trajectories and improved forgetting–retention trade-offs.
3. Our proposed ordering achieves effective forgetting with reduced retain-set utility degradation, while its reverse attains strong forgetting rapidly but with substantially larger utility loss.
4. Finally, via mode connectivity analysis between the pre-unlearning model and post-unlearning models obtained via different orderings, we demonstrate path characteristics induced by sample ordering.

2 RELATED WORK

Machine unlearning. Unlearning methods are broadly categorized into *exact* and *approximate* approaches. Exact techniques, such as SISA Bourtoule et al. [2021] and checkpoint-based retraining schemes, produce models statistically indistinguishable from those retrained from scratch on the retain set, offering strong removal guarantees at the expense of high computational cost or specialized training protocols. Consequently, research has shifted toward approximate unlearning Golatkar et al. [2020], Kurmanji et al. [2023], Fan et al. [2024], Gogineni and Nadimi [2025], which emulates retrain equivalence via efficient post-training updates (e.g., gradient ascent on the forget set or parameter-efficient fine-tuning) while trading perfect equivalence for scalability and practicality. These methods often provide weaker formal guarantees on complete erasure of the forget set's influence.

Path dependence and optimization dynamics in unlearning. Path dependence complicates unlearning in multi-stage training pipelines, as demonstrated by Yu et al. [2025], where local gradient ascent outcomes depend on the order of training stages, rendering retrain equivalence impossible for path-oblivious methods.

This resonates with our findings on ascent-based unlearning but focuses on inter-stage ordering rather than intra-forget-set sample ordering during ascent. In non-convex settings, Jin et al. [2025] frame unlearning as multi-task optimization with normalized gradient differences and adaptive rates to balance forgetting and retention. Our work is complementary: while they enhance gradient-level control, we introduce forget-sample ordering as an orthogonal, trajectory-shaping mechanism that influences path dependence and the achievable Pareto frontier.

Recently, CUFG Miao et al. [2025] proposes a fine-tuning-based unlearning method that uses easy-to-hard scheduling of forget samples, driven by a difficulty score derived from model inference (confidence on forgetting targets) together with a forgetting-gradient corrector for stability. Although CUFG empirically demonstrates the benefits of progressive forgetting, it provides only high-level descriptions of the difficulty score and its interplay with unlearning dynamics without a precise formal definition of the metric and of-

fers no analysis of gradient-ascent instability arising from ordering mechanisms. In contrast, our work diagnoses the instability of gradient-ascent updates in machine unlearning from a local second-order perspective. We show that non-commutativity and gradient–curvature interactions induce path dependence, providing a mechanistic explanation for curriculum-style designs. This perspective motivates a simple low-to-high ordering strategy using gradient-related proxies that improves the forgetting–retention trade-off under fixed optimization budgets, without modifying the optimizer or introducing additional correction terms. More broadly, our analysis connects classical observations on the non-commutativity of stochastic updates in non-convex optimization to ascent-based unlearning objectives.

3 PRELIMINARIES

We consider the standard supervised setting, where a model f_θ is trained on dataset $\mathcal{D} = \mathcal{F} \cup \mathcal{R}$, with forget set $\mathcal{F}$ and retain set $\mathcal{R}$. Machine unlearning seeks to remove the influence of $\mathcal{F}$ while preserving utility on $\mathcal{R}$.

Ascent-based unlearning erases information associated with $\mathcal{F}$ via gradient ascent on the forget-set loss. Starting from pre-trained parameters θ_0, updates proceed as

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta \ell(\theta_t; z_{\pi(t)}), \quad z_{\pi(t)} \in \mathcal{F}, \tag{1}$$

where η is the learning rate and $\pi(t)$ indexes the forget sample selected at step t under ordering π. Prior works typically treat π as a nuisance, defaulting to uniform random shuffling. In contrast, we decouple the ascent rule from sample selection and treat forget-sample ordering as an explicit design choice that shapes the unlearning trajectory and resulting trade-offs.

4 MECHANISTIC ANALYSIS OF SAMPLE ORDERING

A natural question arises: why does the ordering of forget samples alone induce qualitatively different forgetting–retention behaviors under identical optimization budgets?

To address this, we provide a local second-order mechanistic explanation that highlights an important source of path dependence in gradient ascent updates. Our analysis is deliberately local (in the neighborhood of the initial parameters θ_0) and does not claim global guarantees; instead, it qualitatively accounts for the empirical behaviors observed in our experiments. While later-stage dynamics may deviate from this linearized regime, the analysis captures the dominant early-stage interactions that shape trajectory divergence under fixed budgets.

4.1 SETUP AND LOCAL SECOND-ORDER APPROXIMATION

Let $\theta_0 \in \mathbb{R}^d$ denote the parameters of the trained model. For a sample $z = (x, y)$, let $\ell(\theta; z)$ be a twice continuously differentiable loss function. Ascent-based unlearning applies updates of the form

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta \ell(\theta_t; z_{\pi(t)}), \tag{2}$$

where $\eta > 0$ is the step size and π is the sample ordering.

To expose order dependence, consider a first-order Taylor expansion of the gradient around θ_0:

$$\nabla \ell(\theta; z_i) \approx g_i + H_i(\theta - \theta_0), \tag{3}$$

with $g_i = \nabla \ell(\theta_0; z_i)$ and $H_i = \nabla^2 \ell(\theta_0; z_i)$. This local approximation captures the leading second-order interactions over short unlearning trajectories; higher-order terms introduce additional non-commutativity beyond this regime.

4.2 NON-COMMUTATIVITY OF ASCENT UPDATES

For any two forget samples z_i and z_j, applying ascent in order i then j yields

$$\begin{aligned} \theta_{ij} &\approx \theta_0 + \eta g_i + \eta\big(g_j + H_j(\eta g_i)\big) \\ &= \theta_0 + \eta(g_i + g_j) + \eta^2 H_j g_i. \end{aligned} \tag{4}$$

Reversing the order gives

$$\begin{aligned} \theta_{ji} &\approx \theta_0 + \eta g_j + \eta\big(g_i + H_i(\eta g_j)\big) \\ &= \theta_0 + \eta(g_j + g_i) + \eta^2 H_i g_j. \end{aligned} \tag{5}$$

The resulting parameter difference is

$$\theta_{ij} - \theta_{ji} \approx \eta^2 (H_j g_i - H_i g_j). \tag{6}$$

This difference is generally nonzero, demonstrating that ascent updates are inherently *non-commutative*. Order sensitivity emerges from second-order interactions between gradients and local curvature; it vanishes only in degenerate cases (e.g., linear models or flat curvature). Taking the vector norm of the parameter difference in (6) and applying the triangle inequality together with submultiplicativity of the induced operator norm gives:

$$\|\theta_{ij} - \theta_{ji}\| \lesssim \eta^2 \big(\|H_j\| \, \|g_i\| + \|H_i\| \, \|g_j\|\big). \tag{7}$$

Drift is thus controlled by products of gradient magnitudes and curvature strength. Samples inducing large gradients (high $\|g_i\|$) or lying in regions of strong curvature amplify path-dependent effects when encountered early. While stochastic gradient descent also exhibits non-commutativity, gradient ascent can amplifies drift by pushing parameters *away* from the local basin around θ_0. Deferring large-gradient samples can attenuate early drift and better preserve retain-set utility under fixed optimization budgets.

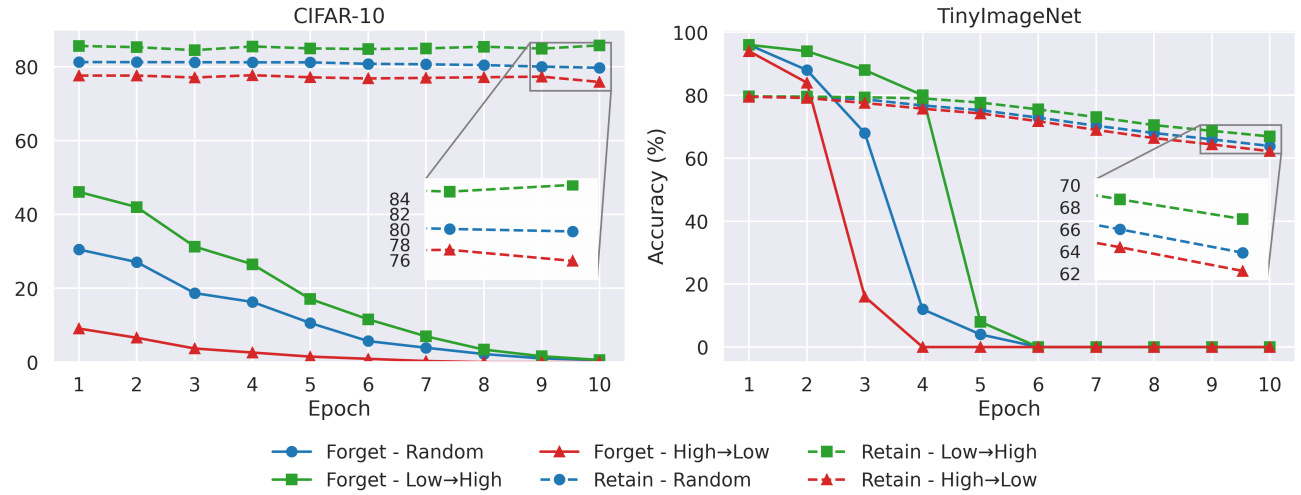

Figure 1: Unlearning trajectories over 10 epochs on CIFAR-10 (left) and TinyImageNet (right).

4.3 PREDICTIVE UNCERTAINTY AS A PROXY FOR UPDATE MAGNITUDE

In standard classification settings with cross-entropy loss and softmax output, the ascent gradient magnitude $\|g_i\|$ at θ_0 correlates strongly with predictive uncertainty on sample z_i: confidently correct predictions produce small gradients, while highly uncertain or misclassified samples yield larger gradients.

Predictive uncertainty thus serves as a cheap, easily computed proxy for expected update magnitude and potential instability. Ordering forget samples by *increasing* uncertainty (low to high) defers large-gradient samples, attenuating early second-order interactions and yielding more stable trajectories under a fixed budget. Conversely, ordering by *decreasing* uncertainty (high to low) prioritizes large updates early, amplifying cumulative drift and leading to aggressive but unstable forgetting.

For a full forget set, the cumulative second-order drift can be approximated as

$$\sum_{t<s} \eta^2 H_{\pi(s)} g_{\pi(t)}, \tag{8}$$

where uncertainty-guided ordering (increasing uncertainty) tends to reduces the magnitude of these terms relative to random or reverse orderings.

In practice, we compute uncertainty via predictive entropy on the forget set at the initial model θ_0, then sort deterministically. This yields our proposed low-to-high proxy ordering without requiring expensive per-sample Hessian computations.

4.4 IMPLICATIONS FOR RETAIN PERFORMANCE

Let $\mathcal{R}$ be the retain set with aggregate loss $\mathcal{L}_\mathcal{R}(\theta)$. Near θ_0, a second-order Taylor expansion yields

$$\mathcal{L}_\mathcal{R}(\theta) \approx \mathcal{L}_\mathcal{R}(\theta_0) + \frac{1}{2}(\theta - \theta_0)^\top H_\mathcal{R}(\theta - \theta_0), \tag{9}$$

where $H_\mathcal{R} = \nabla^2 \mathcal{L}_\mathcal{R}(\theta_0)$ is the retain-set Hessian (assumed positive semi-definite near a local minimum). Excessive parameter drift during ascent thus incurs quadratic degradation in retain performance.

By attenuating early second-order drift through careful sample ordering (as analyzed in the preceding subsections), uncertainty-guided ascent better preserves retain accuracy while still achieving effective forgetting without any modification to the optimizer, learning rate, or other hyperparameters.

This perspective yields a principled notion of *hardness* in unlearning: samples that induce large ascent updates are *hard to unlearn stably*, in that applying them early amplifies path-dependent drift and accelerates retain-set degradation. Conversely, deferring such hard samples reduces early drift and produces more controlled unlearning trajectories.

Within this framework, any per-sample signal correlated with ascent update magnitude such as per-sample loss, gradient norm, or predictive uncertainty can serve to construct orderings that either amplify or suppress cumulative path drift. Predictive uncertainty is therefore a convenient and empirically effective illustrative proxy rather than a necessary element of the analysis. This theoretical view directly motivates our empirical results: sample ordering during gradient ascent can be understood as a trajectory-aware mechanism, for controlling the forgetting vs retention trade-off under fixed optimization budgets.

Table 1: Effect of sample ordering under batch size 32 for 1 epoch on CIFAR-10 (ResNet-18, lr=1e-5) and TinyImageNet (ViT-B/16, lr=5e-5).

Dataset	Method	Forget Train Acc ↓	Retain Train Acc ↑	Forget Test Acc ↓	Retain Test Acc ↑	MIA ↓
CIFAR-10	Original	87.38	80.66	96.33	89.87	1.00
	Random	$15.18_{\pm 0.33}$	$67.11_{\pm 0.35}$	$33.27_{\pm 0.62}$	$81.03_{\pm 0.28}$	0.77
	Low → High	$26.08_{\pm 0.67}$	$\mathbf{71.53}_{\pm 0.18}$	$47.93_{\pm 0.68}$	$\mathbf{84.69}_{\pm 0.15}$	0.81
	High → Low	$\mathbf{3.32}_{\pm 0.07}$	$63.48_{\pm 0.15}$	$\mathbf{8.23}_{\pm 0.05}$	$77.25_{\pm 0.08}$	0.74
TinyImageNet	Original	95.20	79.91	96.00	79.72	0.98
	Random	$19.33_{\pm 3.40}$	$78.08_{\pm 0.08}$	$23.33_{\pm 0.94}$	$77.86_{\pm 0.09}$	0.82
	Low → High	$75.40_{\pm 1.02}$	$\mathbf{79.44}_{\pm 0.01}$	$84.00_{\pm 0.00}$	$\mathbf{79.20}_{\pm 0.04}$	0.88
	High → Low	$\mathbf{0.00}_{\pm 0.00}$	$72.37_{\pm 0.27}$	$\mathbf{0.00}_{\pm 0.00}$	$72.37_{\pm 0.18}$	0.71

Table 2: Sample unlearning with ResNet18 and CIFAR-10.

Method	Retain Train Acc ↑	Forget Train Acc ↓	Test Acc	MIA
Original	87.44	84.291	92.87	1.00
Random	74.73	31.60	85.41	0.84
Low → High	**76.93**	36.08	**88.24**	0.86
High → Low	73.30	**23.12**	83.14	0.82

5 EXPERIMENTAL RESULTS

5.1 DATASETS AND MODELS

We evaluate gradient ascent-based unlearning on both image and text classification benchmarks. For *image classification*, we perform class-level unlearning on **CIFAR-10** and **Tiny-ImageNet**, where all samples from a randomly selected class form the forget set $\mathcal{F}$ and the remainder form the retain set $\mathcal{R}$. We use **ResNet-18** and **ViT-B/16** architectures pre-trained on ImageNet. For *text classification*, we conduct class-level unlearning on the **20 Newsgroups** dataset with a **RoBERTa-base** model. Additionally, we assess sample-level unlearning on CIFAR-10, where the forget set consists of 500 randomly selected training samples.

5.2 IMPLEMENTATION DETAILS

We fine-tune ResNet-18 and ViT-B/16 from ImageNet-pretrained checkpoints for 10 epochs using cross-entropy loss and Adam optimizer (learning rate 1×10^{-4}). Unlearning proceeds via gradient ascent on $\mathcal{F}$ with batch size 32 (unless noted otherwise). Epistemic uncertainty is estimated via Monte Carlo Dropout (30 stochastic forward passes, dropout rate 0.5 as in training); predictive entropy serves as the primary ordering signal. Other hyperparameters (learning rate, number of epochs/steps) are specified per experiment. Importantly, the forget-set ordering is computed once on the initial pre-unlearning model (using predictive uncertainty, per-sample loss, or gradient norm) and then held fixed throughout gradient ascent. No re-computation occurs during unlearning. This fixed-order protocol is the computationally efficient design we recommend and use throughout the paper.

5.3 EVALUATION METRICS

We measure unlearning efficacy using classification accuracy on the forget set (↓ for effective forgetting) and retain set (↑ for utility preservation). We also report membership inference attack (MIA) scores to quantify residual privacy leakage (higher MIA indicates poorer forgetting). All results are averaged over three independent runs unless stated otherwise.

5.4 CLASS UNLEARNING

Table 1 examines the effect of sample ordering after a single epoch of gradient ascent. While practical unlearning is typically performed over multiple epochs and a single epoch is generally insufficient to achieve complete forgetting, this experiment isolates early optimization dynamics under a fixed budget. The results demonstrate that gradient ascent-based unlearning is *path-dependent*: the ordering of forget samples substantially changes the forgetting–retention trade-off. High-to-low ordering induces the most aggressive forgetting but at the cost of reduced retain accuracy. In contrast, low-to-high ordering forgets more conservatively during the initial steps, resulting in higher forget accuracy after one epoch while almost perfectly preserving retain-set performance. On TinyImageNet, for example, Low→High achieves 84.0% forget test accuracy with 79.20% retain accuracy—substantially better retention than Random (23.3% / 77.86%) and High→Low (0.0% / 72.37%). These single-epoch results therefore reflect different early-stage operating points rather than differences in eventual unlearning capability. As shown in Figure 1, with additional epochs all orderings achieve near-complete forgetting, while low-to-high consistently maintains superior

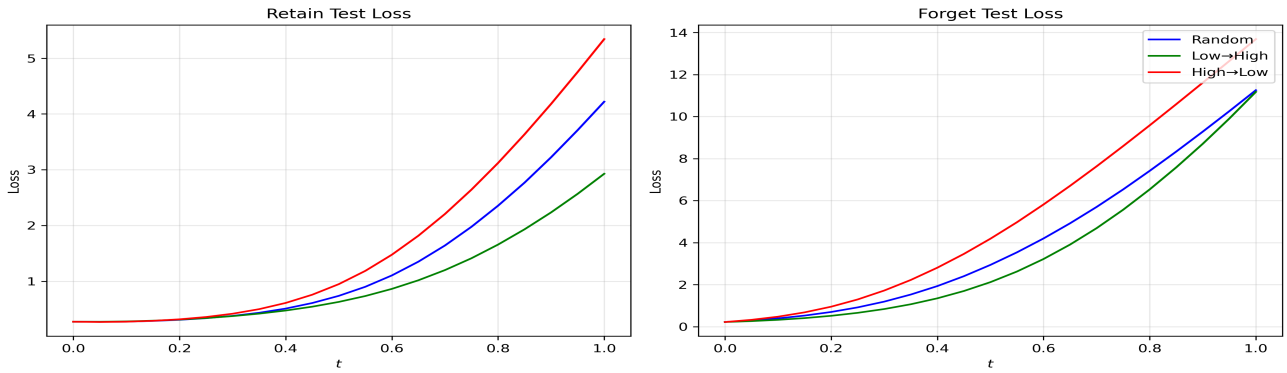

Figure 2: Mode connectivity via linear interpolation. **Left:** Retain test loss along paths from original ($t = 0$) to unlearned models ($t = 1$). High→Low shows a clear barrier, while Low→High remains flat. **Right:** Forget test loss rises monotonically (fastest for High→Low), indicating effective but ordering-dependent erasure. Colors: blue=Random, green=Low→High, red=High→Low.

retain accuracy throughout the trajectory. Together, these findings identify sample ordering as a practical lever for controlling the forgetting–retention trade-off.

5.5 SAMPLE UNLEARNING

To assess whether path dependence persists beyond class-level unlearning, we conduct sample-level experiments on CIFAR-10 with ResNet-18. Here, the forget set $\mathcal{F}$ comprises 500 randomly selected training samples (drawn from a random classes on the dataset), while the retain set $\mathcal{R}$ consists of the remaining samples.

Table 2 presents results under a fixed optimization budget. Random ordering yields intermediate forgetting (forget accuracy $\approx 31.6\%$) with moderate retain-set degradation. Low-to-high uncertainty ordering produces more conservative forgetting (forget accuracy $\approx 36.1\%$) but substantially better retention (retain accuracy $\approx 76.9\%$), making it preferable in utility-sensitive or privacy-constrained applications. Conversely, high-to-low ordering achieves aggressive forgetting (forget accuracy $\approx 23.1\%$) at greater cost to retain performance (retain accuracy $\approx 73.3\%$).

These patterns mirror the class-level findings, albeit with somewhat smaller gaps due to reduced statistical entanglement between forget and retain samples. The results further confirm that forget-sample ordering remains a powerful, systematic control mechanism for navigating the forgetting vs retention trade-off, even in finer-grained unlearning scenarios.

5.6 UNLEARNING TRAJECTORIES

To investigate how sample ordering shapes the *dynamics* of ascent-based unlearning under fixed optimization budgets, we track forget-set and retain-set test accuracy over multiple epochs on CIFAR-10. Figure 1 illustrates the resulting trajectories for three strategies: random, low-to-high uncertainty, and high-to-low uncertainty.

Across all orderings, forget accuracy decreases monotonically with additional unlearning steps, confirming that effective forgetting is asymptotically achievable regardless of ordering. However, the *rate* of forgetting and the associated retain-set degradation depend critically on sample order. High-to-low ordering drives forget accuracy toward near-zero rapidly (within a few epochs) but causes substantial early parameter drift, leading to pronounced retain accuracy loss. In contrast, low-to-high ordering yields a slower, more gradual forgetting curve that preserves retain performance over a much larger number of epochs.

These trajectories provide direct evidence of the fundamental path dependence of gradient ascent-based unlearning: even when the same final forgetting objective is reachable, the intermediate path is different and thus the forgetting vs retention trade-off attained under realistic (constrained) budgets is highly sensitive to forget-sample ordering. In this light, ordering functions as an implicit regularization mechanism that modulates the speed of forgetting relative to collateral retain-set damage, offering a simple yet powerful lever for navigating the Pareto frontier without altering the optimizer, learning rate, or other hyperparameters.

Sensitivity analyses across unlearning batch sizes, learning rates, and forget classes reveal consistent directional effects from structured orderings, while random ordering consistently produces higher variance and less predictable outcomes. We also ablated alternative proxies (per-sample loss and exact gradient norm at θ_0), observing qualitatively identical path-dependent behavior. Detailed results including class-wise trends on all CIFAR-10 classes, batch-size sweeps, learning-rate sweeps, and additional visualizations are provided in Appendix A.

Table 3: Effect of sample ordering on 20 Newsgroups (RoBERTa-base, 1 epoch, lr=1×10^{-6}). Results averaged over 3 runs.

Method	Forget Train Acc ↓	Retain Train Acc ↑	Forget Test Acc ↓	Retain Test Acc ↑
Original	95.7	95.2	85.6	71.4
Random	64.8	93.0	52.1	69.5
Low → High	68.6	93.5	56.1	70.5
High → Low	60.7	91.6	50.0	65.5

Table 4: Comparison with unlearning baselines for class unlearning on CIFAR-10 with ResNet-18.

Method	Forget Test Acc ↓	Retain Test Acc ↑
Original	89.3	84.467
Gradient Ascent (Random)	0	68.067
Random Labelling	0.3	68.3
BDSH	4.7	73.456
JiT	29.5	59.622
Gradient Ascent (Low → High, proposed)	**0**	**76.811**

5.7 MODE CONNECTIVITY ANALYSIS

To probe the landscape-level consequences of path dependence in ascent-based unlearning, we perform mode connectivity analysis via linear parameter interpolation between the original pre-unlearning model (θ_0) and each post-unlearning model ($\theta_{\text{unlearned}}$), following Garipov et al. [2018]. We evaluate retain-set and forget-set test loss along the straight-line path

$$\theta(t) = (1-t)\theta_0 + t\theta_{\text{unlearned}}, \quad t \in [0,1], \qquad (10)$$

using 21 evenly spaced interpolation points.

Figure 2 shows the resulting curves on CIFAR-10 (ResNet-18) for retain test loss (left) and forget test loss (right), across random (blue), low-to-high (green), and high-to-low (red) orderings. All paths begin near the original low-loss region (≈ 0.3–0.5) at $t = 0$. By $t = 1$, losses increase, reflecting unlearning progress.

On the retain set (left panel), high-to-low ordering induces the steepest loss rise and highest endpoint ($\gtrsim 5$), with a pronounced barrier emerging around $t \approx 0.4$–0.6. This sharp disconnection from the original basin aligns with amplified early second-order drift (Eq. (8)) and the rapid retain degradation seen in Tables 1 and Fig. 1. In contrast, low-to-high ordering produces a much flatter curve with lower endpoint loss (≈ 2.8–3.0) and minimal barrier, confirming that deferring large-gradient samples preserves connectivity and stable retention. Random ordering shows intermediate behavior with moderate fluctuations.

On the forget set (right panel), test loss rises monotonically across all orderings (indicating progressive erasure), with high-to-low rising fastest (aggressive forgetting) and low-to-high slowest (controlled, gradual erasure). The forget paths remain relatively smooth, consistent with deliberate destabilization of forget knowledge.

These connectivity results provide landscape-level evidence that sample ordering during ascent not only governs unlearning dynamics but also shapes the post-unlearning geometry: aggressive (high-to-low) curricula push models into isolated, high-barrier regions at high retain cost, while conservative (low-to-high) curricula maintain connectivity to the original basin with minimal collateral damage. This reinforces the mechanistic role of ordering in controlling path dependence and underscores the value of trajectory- and geometry-aware design and evaluation for optimization-based unlearning methods.

5.8 TEXT CLASSIFICATION UNLEARNING

To assess the generality of path dependence beyond vision tasks, we extend experiments to text classification—a domain where machine unlearning is especially relevant for removing copyrighted material, harmful content, or privacy-sensitive data. We perform class-level unlearning on the 20 Newsgroups dataset Mitchell [1997], a standard 20-class topic classification benchmark. The forget set $\mathcal{F}$ comprises all training samples from a randomly selected class, with the retain set $\mathcal{R}$ consisting of the remaining samples.

We fine-tune a RoBERTa-base model Liu et al. [2019] for 10 epochs using cross-entropy loss and Adam optimizer (learning rate 1×10^{-5}). Unlearning proceeds via gradient ascent on $\mathcal{F}$ with learning rate 1×10^{-6}, batch size 32, and fixed budget (1 epoch), varying only forget-sample ordering. As before, we compare random ordering against low-to-high and high-to-low curricula based on epistemic uncertainty (Monte Carlo Dropout with 30 stochastic forward passes, dropout rate 0.5; predictive entropy as ordering signal).

Performance is measured via classification accuracy on forget and retain train/test splits (low forget accuracy = effective unlearning; high retain accuracy = utility preservation). Results, averaged over three runs, appear in Table 3. Consistent with image-domain findings, ascent-based unlearning exhibits pronounced path dependence. High-to-low ordering drives the most aggressive forgetting (forget test accuracy $\approx 50.0\%$) but incurs greater retain degradation (retain test accuracy $\approx 65.5\%$). Low-to-high ordering yields more conservative forgetting (forget test accuracy $\approx 56.1\%$) while substantially better preserving retain performance (retain test accuracy $\approx 70.5\%$). Random ordering produces intermediate results with noticeably higher variance across runs.

These results confirm that ordering-induced trajectory divergence is not domain-specific but a general characteristic of gradient ascent-based unlearning. Even in text classification setting where samples are sequential and models capture long-range dependencies by deferring high-uncertainty samples enables stable, utility-preserving forgetting. This underscores the practical value of trajectory-aware ordering strategies across modalities and model architectures.

5.9 COMPARISON WITH OTHER RETAIN DATA-FREE UNLEARNING APPROACHES

To further contextualize the effectiveness of our proposed sample ordering strategy, we compare it against several established unlearning baselines on CIFAR-10 class unlearning with ResNet-18. We focus on retain-data-free methods, as the gradient ascent-based approach operates solely on the forget set, aligning with practical scenarios where retain data may be inaccessible due to privacy constraints or storage limitations. The baselines include: (i) Random Labelling Golatkar et al. [2020], a simple heuristic that relabels forget samples to random classes during fine-tuning; (ii) BDSH (Boundary Shrinking) Chen et al. [2023], a zero-shot method that shrinks decision boundaries around forget samples without retain data access; and (iii) JiT (Just-in-Time) Unlearning Foster et al. [2025], an information-theoretic zero-shot approach that minimizes gradient effects via local perturbations around forget samples. All methods are evaluated on test-set accuracy, averaged over three runs.

Table 4 presents the results. Our low-to-high uncertainty ordering achieves perfect forgetting (0% forget test accuracy) while preserving the highest retain test accuracy (76.81%) among all compared methods. This performs above vanilla random-ordering GA by 8.7% in retain utility without sacrificing forgetting efficacy. Compared to other retain-data-free baselines, our approach achieves BDSH by 3.4% in retain accuracy (with stronger forgetting) and substantially improves over JiT, which exhibits weaker forgetting and the lowest retain performance. These gains highlight how trajectory-aware ordering elevates simple GA to a more favorable position on the forgetting-retention Pareto frontier, offering a lightweight, hyperparameter-free enhancement that competes with more specialized zero-shot techniques.

6 CONCLUSION

In this work, we identified sample ordering during gradient ascent as a critical yet previously underexplored degree of freedom in machine unlearning. Through theoretical analysis and experiments across image and text classification benchmarks, we demonstrated that ascent-based unlearning is fundamentally *path dependent*: under identical optimization budgets, different orderings of the same forget samples yield markedly different forgetting–retention trade-offs.

We showed that gradient magnitude (and correlated proxies such as predictive uncertainty) is the primary driver of trajectory divergence. Samples inducing large ascent updates amplify early second-order interactions and parameter drift when encountered early, leading to aggressive forgetting but substantial retention degradation. Conversely, deferring such high-gradient samples via a low-to-high ordering yields more stable trajectories, enabling effective forgetting with minimal utility loss (albeit over more steps). This ordering strategy consistently improves the Pareto frontier in the forget vs retain performance space relative to random shuffling, which explores the trade-off stochastically and with higher variance.

More broadly, our findings highlight the importance of trajectory- and geometry-aware design in unlearning methods. By explicitly accounting for optimization dynamics rather than treating them as noise, we obtain more controllable, lower-variance, and practically robust unlearning outcomes across modalities and architectures.

7 ACKNOWLEDGEMENTS

This work received funding from the Novo Nordisk Foundation (Project Reference Number: NNF24OC0095455).

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605585161. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

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 Pri-*

vacy (SP), pages 141–159, 2021. doi: 10.1109/SP40001.2021.00019.

Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7766–7775, 2023. doi: 10.1109/CVPR52729.2023.00750.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gn0mIhQGNM.

Jack Foster, Kyle Fogarty, Stefan Schoepf, Zack Dugue, Cengiz Oztireli, and Alexandra Brintrup. An information theoretic approach to machine unlearning. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 8803–8812, Red Hook, NY, USA, 2018. Curran Associates Inc.

Vinay Chakravarthi Gogineni and Esmaeil S. Nadimi. Efficient knowledge deletion from trained models through layer-wise partial machine unlearning. *Journal of Machine Learning Research*, 26(245):1–33, 2025. URL http://jmlr.org/papers/v26/24-0469.html.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

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 Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *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)*, pages 11278–11294, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.563. URL https://aclanthology.org/2025.naacl-long.563/.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 1957–1987. Curran Associates, Inc., 2023.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL https://arxiv.org/abs/1907.11692.

A. Mantelero. The eu proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Computer Law and Security Review*, 29(3):229–235, 2013.

Ioannis Mavrothalassitis, Pol Puigdemont, Noam Levi, and Volkan Cevher. Ascent fails to forget. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 45330–45365. Curran Associates, Inc., 2025.

Jiaxing Miao, Liang Hu, Qi Zhang, Lai Zhong Yuan, and Usman Naseem. Cufg: Curriculum unlearning guided by the forgetting gradient, 2025.

Tom Mitchell. Twenty Newsgroups. UCI Machine Learning Repository, 1997. DOI: https://doi.org/10.24432/C5C323.

Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1), August 2023. ISSN 0360-0300. doi: 10.1145/3603620. URL https://doi.org/10.1145/3603620.

Jiatong Yu, Yinghui He, Anirudh Goyal, and Sanjeev Arora. On the impossibility of retrain equivalence in machine unlearning, 2025. URL https://arxiv.org/abs/2510.16629.

On the Path Dependence of Gradient Ascent-Based Unlearning (Supplementary Material)

Varun Sampath Kumar[1] **Esmaeil S Nadimi**[1] **Vinay Chakravarthi Gogineni**[1]

[1]Applied AI and Data Science Unit, The Maersk McKinney Moller Institute, University of Southern Denmark, Denmark

A ADDITIONAL SIMULATION RESULTS

We report on class-wise forget and retain test accuracy for CIFAR-10 class unlearning. Each class is unlearned independently and experiments are repeated over three random seeds.

Table 5: Class-wise forget and retain test accuracy (%) on CIFAR-10 (mean over 3 runs).

Class	**Forget Test Accuracy**			**Retain Test Accuracy**		
	Random	Low → High	High → Low	Random	Low → High	High → Low
0	23.1	43.3	5.9	84.5	88.5	79.5
1	34.9	49.3	8.3	81.5	85.3	77.2
2	26.1	40.4	8.4	87.5	90.2	83.0
3	28.8	41.1	7.7	89.6	91.5	86.5
4	21.5	38.2	6.0	84.9	88.5	81.1
5	24.5	34.8	5.2	84.9	88.2	82.8
6	16.7	32.8	3.5	86.2	88.9	83.1
7	20.6	41.4	3.8	84.2	88.2	80.1
8	23.0	40.5	5.7	83.5	86.8	80.9
9	28.3	46.5	6.3	84.3	87.6	81.7

The results in Table 5 illustrate that unlearning effectiveness varies significantly across classes and ordering strategies. High → Low consistently achieves the lowest forget accuracies, demonstrating the most aggressive unlearning, but often at the expense of lower retain accuracies. In contrast, Low → High preserves retention while still reducing forget accuracy moderately, and Random ordering lies in between. These patterns reinforce our earlier observations: sample ordering alone is sufficient to control the forgetting–retention trade-off, and this effect persists at a class-wise granularity. Importantly, the consistent trends across classes suggest that trajectory-aware strategies, such as ordering by epistemic uncertainty, can provide predictable and controllable unlearning behavior for all categories.

A.1 SENSITIVITY ANALYSIS

A.1.1 Effect of Unlearning Batch Size

We analyze the sensitivity of ascent-based unlearning to the batch size used during gradient ascent. Batch sizes range from very small (1, 2, 4) to large (256, 1024), with the learning rate adjusted to ensure stable optimization.

The results in Table 6 show that uncertainty-guided ordering consistently outperforms random ordering across batch sizes, with the advantage being most pronounced at small and moderate scales. For batch sizes from 1 to 32, the uncertainty-

Table 6: Effect of batch size on CIFAR-10 (ResNet-18) forget and retain test accuracy (%).

Batch Size	Forget Test Accuracy			Retain Test Accuracy		
	Random	Low → High	High → Low	Random	Low → High	High → Low
1	40.0	32.4	10.5	82.1	83.7	81.9
4	29.4	50.7	6.3	80.4	85.9	77.4
8	40.0	51.8	6.8	80.8	85.2	76.8
16	40.3	48.6	6.0	81.1	84.6	75.2
32	34.6	56.5	9.4	79.7	85.0	76.9
256	39.9	49.7	23.3	81.3	83.4	79.0
1024	66.3	70.5	63.3	86.9	87.2	86.7

decreasing ordering (High → Low) achieves markedly lower forget test accuracy than both random and uncertainty-increasing (Low → High) orderings, demonstrating more effective unlearning. By contrast, the uncertainty-increasing ordering maintains strong retain performance but yields substantially higher forget accuracy, indicating weaker unlearning. As batch size grows to 256 and especially 1024, the gaps between all three orderings shrink dramatically; at batch size 1024 the forget and retain accuracies become nearly indistinguishable. This pattern indicates that ordering effects are strongest when updates remain highly sequential and weaken as updates are aggregated into larger mini-batches.

This empirical behavior is explained by a natural batched extension of the local second-order analysis in Section 4.2. For mini-batch gradient ascent the parameter update takes the form

$$\theta_{t+1} = \theta_t + \eta G_k(\theta_t), \tag{11}$$

where $G_k(\theta)$ denotes the batch-averaged gradient. Expanding around a reference point θ_0 gives

$$G_k(\theta) \approx g_k + H_k(\theta - \theta_0), \tag{12}$$

with g_k and H_k the batch-averaged gradient and Hessian, respectively. For two consecutive mini-batches i and j, a second-order Taylor expansion of the composed updates yields

$$\theta_{ij} \approx \theta_0 + \eta(g_i + g_j) + \eta^2 H_j g_i, \tag{13}$$

$$\theta_{ji} \approx \theta_0 + \eta(g_i + g_j) + \eta^2 H_i g_j. \tag{14}$$

The resulting order-induced drift is therefore

$$\theta_{ij} - \theta_{ji} \approx \eta^2(H_j g_i - H_i g_j), \tag{15}$$

which has the same structural form as the single-sample case but with gradients and Hessians averaged inside each mini-batch. This averaging attenuates both the magnitude and the variance of the gradient–curvature interaction terms that drive path dependence. Consequently, ordering effects diminish as batch size increases, exactly as observed in Table 6.

A.1.2 Effect of Learning Rate

Table 7: Effect of learning rate on CIFAR-10 forget and retain test accuracy (%).

Learning Rate	Forget Test Accuracy			Retain Test Accuracy		
	Random	Low → High	High → Low	Random	Low → High	High → Low
10^{-5}	0.4	0.9	0.0	80.6	85.4	77.4
10^{-6}	32.8	45.2	6.5	81.6	85.0	76.6
10^{-7}	40.8	52.1	10.9	81.6	84.8	76.9
10^{-8}	36.7	53.8	12.6	81.0	85.0	77.7

As shown in Table 7, In the intermediate regime—where ascent-based unlearning is typically applied in practice—sample ordering induces pronounced differences. Uncertainty-decreasing ordering consistently achieves aggressive forgetting, uncertainty-increasing ordering remains conservative, and random ordering exhibits high variance. This indicates that ordering effects are not artifacts of a specific learning rate choice.

A.1.3 Sensitivity to the Choice of Ordering Proxy

In the main paper, we use epistemic uncertainty estimated via Monte Carlo (MC) Dropout as an illustrative proxy for ordering forget samples. While this choice is motivated by its connection to predictive confidence, the path-dependent behavior identified in this work does not rely on uncertainty estimates specifically. To assess the robustness of our findings, we evaluate alternative and simpler ordering criteria based on per-sample loss and gradient norm. All ordering signals are computed once at the initial parameter state θ_0 (i.e., before unlearning), and the same unlearning budget and optimization hyperparameters are used across all experiments. We consider the following proxies:

- **Loss-based ordering.** Samples are ordered by their cross-entropy loss $\ell(\theta_0; z)$.
- **Gradient-norm-based ordering.** Samples are ordered by the ℓ_2 norm of the per-sample gradient $\|\nabla_\theta \ell(\theta_0; z)\|_2$.
- **Epistemic uncertainty (MC Dropout).** Uncertainty is estimated using predictive entropy computed from $M = 30$ stochastic forward passes with dropout enabled (dropout rate $p = 0.5$).

For each proxy, we evaluate both *Low* $\rightarrow$ *High* and *High* $\rightarrow$ *Low* orderings, corresponding to processing low- and high-magnitude samples early in the unlearning trajectory, respectively.

Table 8 reports results for class-level unlearning on CIFAR-10 with a ResNet-18 model using a single epoch of gradient ascent on the forget set. Random ordering is included as a baseline. All results are averaged over three independent runs.

Table 8: Effect of different ordering proxies on CIFAR-10 (ResNet-18). All settings use identical unlearning budgets; only the ordering of forget samples differs.

Ordering Proxy	Direction	Forget Test Acc.	Retain Test Acc.
Original	–	96.50	90.60
Random	–	34.90	82.10
Loss	Low $\rightarrow$ High	51.30	89.26
Loss	High $\rightarrow$ Low	5.70	82.04
Gradient Norm	Low $\rightarrow$ High	54.00	84.90
Gradient Norm	High $\rightarrow$ Low	8.10	79.50
MC Dropout	Low $\rightarrow$ High	47.00	84.84
MC Dropout	High $\rightarrow$ Low	8.20	77.31

Across all ordering proxies, we observe the same qualitative behavior. Orderings that apply high-magnitude samples early in the unlearning trajectory (*High* $\rightarrow$ *Low*) induce aggressive forgetting at the cost of reduced retain accuracy, whereas orderings that prioritize low-magnitude samples (*Low* $\rightarrow$ *High*) preserve retained performance but result in more conservative forgetting. Random ordering yields intermediate behavior with higher variance.

A.1.4 Extended Membership Inference Attack (MIA) Analysis

The elevated membership inference attack (MIA) scores reported in Tables 1 and 2 should not be interpreted as evidence of overfitting in the unlearned models. These experiments use only a single epoch of gradient ascent, a deliberately constrained optimization budget chosen to study the effect of sample ordering during the early stages of unlearning. Under this limited budget, the model has not yet fully erased forget-set information, which is consistent with the forget-set accuracies remaining substantially above 0%. Because a single epoch of gradient ascent is generally insufficient for complete forgetting, the elevated MIA values primarily reflect incomplete unlearning rather than memorization or overfitting.

Gradient ascent-based unlearning is typically run for multiple epochs to ensure effective removal of forget-set information. To evaluate performance under a more standard unlearning budget, we conduct additional experiments with five epochs of gradient ascent on CIFAR-10 class unlearning using ResNet-18 (same hyperparameters as the main experiments: $\eta = 5 \times 10^{-6}$, batch size 32). Results are summarized in Table 9.

After five epochs, all ordering strategies achieve complete forgetting (forget-set accuracy = 0%). Correspondingly, MIA scores drop substantially and fall well below the random-guess threshold of 0.5, confirming effective removal of membership

Table 9: Unlearning performance after five epochs of gradient ascent on CIFAR-10 class unlearning using ResNet-18.

Method	Forget Test Acc. ↓	Retain Test Acc. ↑	MIA ↓
Original	96.90	94.23	0.835
Random Order	0.00	79.00	0.227
High → Low	0.00	76.03	0.247
Low → High (Ours)	**0.00**	**84.97**	**0.172**

information. Among the evaluated strategies, the proposed uncertainty-decreasing ordering (Low → High) achieves the best utility-forgetting–privacy trade-off, obtaining both the highest retain accuracy and the lowest forget accuracy and MIA score.