ON THE PATH DEPENDENCE OF GRADIENT ASCENT-BASED UNLEARNING

Varun Sampath Kumar, Esmaeil S. Nadimi & Vinay Chakravarthi Gogineni *
Applied AI and Data Science Unit
The Maersk Mc-Kinney Moller Institute
University of Southern Denmark
Odense, Denmark
{vaku,esi,vigo}@mmmi.sdu.dk

ABSTRACT

Machine unlearning enables selective erasure of knowledge associated with specific data points from trained models. In this work, we show that gradient ascent-based machine unlearning is fundamentally path-dependent, i.e, under identical optimization budgets, different orderings of the same forget samples can induce qualitatively different forgetting vs retention trade-offs. Through experiments on CIFAR-10 and TinyImageNet, we demonstrate that structured sample orderings systematically steer unlearning toward distinct extremal regimes, whereas random ordering explores these outcomes only stochastically and with high variance. We further show that this behavior is robust across multiple ordering proxies, including epistemic uncertainty, loss, and gradient norm, indicating that the ordering itself, rather than a specific heuristic are the primary driver of trajectory divergence. These findings reveal sample ordering as a previously overlooked but practically significant degree of freedom in machine unlearning and motivate trajectory-aware design and evaluation of optimization-based unlearning methods.

1 INTRODUCTION

Modern machine learning models are increasingly trained on large-scale datasets that contain sensitive and user-generated information. Once training is complete, the influence of individual data points becomes implicitly encoded within model parameters, making their selective removal non-trivial without retraining from scratch. The enactment of the General Data Protection Regulation (GDPR) Mantelero (2013) by the European Union has made the deletion of user data from machine learning models a critical requirement in practice. Unlike conventional databases, where records can be explicitly removed, eliminating the influence of specific training data from a learned model is substantially more complex, as it requires negating the contribution of the targeted data while preserving utility on the remaining data. Consequently, machine unlearning has emerged as a critical research area aimed at removing the influence of selected data from trained models without incurring the prohibitive cost of full retraining Bourtoule et al. (2021); Xu et al. (2024).

Among existing practical unlearning approaches Golatkar et al. (2020); Kurmanji et al. (2023), many rely on a post-training fine-tuning phase to remove the influence of selected samples. These methods typically apply gradient ascent on the forget set for a limited number of epochs, optionally followed by gradient descent on the retain set to recover any loss in model utility. However, recent evaluations show that such approaches are highly sensitive to hyperparameter choices which includes learning rate, number of epochs, and stopping criteria which significantly weakens their theoretical guarantees. Similarly, recent works highlight the unstable behavior of ascent-based methods when unlearning forget sets whose representations are entangled with retain data Mavrothalassitis et al. (2025). These observations raise a fundamental question: *to what extent is ascent-based unlearning intrinsically well-defined?*

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

Sample ordering has been widely studied in curriculum learning Bengio et al. (2009), where easy-to-hard schedules are known to improve convergence during training. Unlearning, however, poses a fundamentally different objective,i.e., controlled destruction of the forget set while maintaining stability on the retain set. This distinction motivates our study of hardness and its impact on the unlearning trajectory.

We show that gradient-ascent–based unlearning is fundamentally path-dependent, much like gradient descent in training, but with a crucial caveat. Ordering samples that induce larger gradients early causes the optimization trajectory to drift substantially, leading to rapid forgetting at the cost of significant degradation in model utility. In contrast, reversing this order yields slower, more conservative forgetting trajectories that better preserve retained accuracy over extended unlearning horizons. Random ordering, by comparison, explores this trade-off only stochastically and exhibits high variance. These findings reveal sample ordering as an important and often overlooked degree of freedom in gradient ascent-based unlearning. More broadly, our results motivate a shift toward trajectory-aware design and evaluation of unlearning methods, complementing algorithm-centric perspectives with explicit consideration of optimization dynamics.

To summarize, our main contributions are:

- We show that ascent-based unlearning is fundamentally *path-dependent*: under identical optimization budgets, the ordering of forget samples alone determines the forgetting–retention trajectory.
- We identify a reversed unlearning curriculum, demonstrating that deferring samples inducing large ascent updates enables stable, multi-epoch forgetting with minimal utility loss.
- We validate this principle across datasets and architectures, showing that simple ordering strategies rehabilitate gradient ascent as a competitive and controllable unlearning baseline.

2 Ascent-Based Unlearning and Sample Ordering

We consider the standard supervised learning setting, where a model f_θ is trained on a dataset $\mathcal{D} = \mathcal{F} \cup \mathcal{R}$, consisting of a forget set $\mathcal{F}$ and a retain set $\mathcal{R}$. The goal of machine unlearning is to remove the influence of $\mathcal{F}$ from the trained model while preserving performance on $\mathcal{R}$.

Ascent-based unlearning removes information associated with $\mathcal{F}$ by performing gradient ascent on the forget-set loss. Given a pre-trained model with parameters θ_0, unlearning proceeds via iterative updates

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

where η is a fixed learning rate and $\pi(t)$ denotes the forget sample selected at iteration t according to an ordering π. Existing work almost universally treats π as a nuisance variable, typically implemented as a uniform random shuffle. In contrast, we explicitly separate the ascent update rule from the sample selection sequence and treat the ordering of forget samples as a design choice that shapes the unlearning trajectory.

2.1 Why Sample Ordering Shapes Unlearning Trajectories

The strong empirical sensitivity of ascent-based unlearning to π raises a natural question: why should ordering alone induce qualitatively different forgetting–retention behavior under identical optimization budgets? We address this question through a local second-order analysis provided in Appendix A.1.

Briefly, ascent updates are inherently non-commutative. For any two forget samples z_i and z_j, the difference in the resulting parameter state after applying two ascent updates in opposite orders satisfies

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

where g and H denote the local gradient and Hessian evaluated at θ_0. This interaction between gradient magnitude and local curvature implies that early ascent updates exert a disproportionate influence on the final parameter state.

This observation provides a principled notion of *hardness* in unlearning. Samples that induce large ascent updates characterized by large gradient norms or strong curvature interactions are *hard to unlearn* in the sense that applying them early causes substantial parameter drift and rapid degradation of retained performance. Conversely, deferring such samples attenuates early drift and yields more stable unlearning trajectories.

Within this framework, any per-sample signal correlated with ascent update magnitude, such as loss, gradient norm, or predictive uncertainty, can be used to construct orderings that either amplify or suppress path-dependent drift. Predictive uncertainty therefore serves as a convenient illustrative proxy rather than a necessary component of the analysis. This theoretical perspective motivates our empirical findings: sample ordering is not a statistical artifact, but a systematic mechanism for controlling the forgetting–retention trade-off under fixed optimization budgets.

3 Experiments and Results

We conducted class-level unlearning experiments on CIFAR-10 Krizhevsky (2009) and TinyImageNet Le & Yang (2015), where all samples from a randomly selected class form the forget set and the remaining samples constitute the retain set. We evaluated ResNet-18 He et al. (2016) and ViT-B/16 Dosovitskiy et al. (2021) models, initialized from standard pretrained checkpoints and fine-tuned for 10 epochs using cross-entropy loss with the Adam optimizer (learning rate 1×10^{-4}). Unlearning is performed via gradient ascent on the forget set with a fixed learning rate of 1×10^{-5}, batch size 32, and identical optimization budgets across all runs, varying only the ordering of forget samples. We compared random ordering with structured curricula based on epistemic uncertainty estimated via Monte Carlo Dropout Gal & Ghahramani (2016) (30 passes, dropout rate 0.5), using predictive entropy to construct low-to-high and high-to-low orderings. Performance is evaluated using classification accuracy on forget and retain train and test sets, where effective unlearning corresponds to low forget accuracy and high retain accuracy.

3.1 Order Sensitivity Under Identical Unlearning Budgets

The results presented in Table 1 and Table 2 demonstrates that ascent-based unlearning is fundamentally *path-dependent*,i.e., even with identical optimization budgets, the ordering of forget samples markedly alters the forgetting–retention trade-off. Specifically, 'High $\rightarrow$ Low' ordering induces the most aggressive forgetting at the expense of retain accuracy, whereas 'Low $\rightarrow$ High' achieves moderate forgetting and stronger retention. This highlights sample ordering as a previously overlooked but practically significant lever for steering unlearning behavior.

Table 1: Effect of sample ordering under under batch size 32 and learning rate 1e-5 for 1 epoch on CIFAR-10 (ResNet-18).

Method	Forget Train Acc ↓	Retain Train Acc ↑	Forget Test Acc ↓	Retain Test Acc. ↑
Original	87.38	80.66	96.33	89.87
Random	15.18 ± 0.33	67.11 ± 0.35	33.27 ± 0.62	81.03 ± 0.28
Low $\rightarrow$ High	26.08 ± 0.67	$\mathbf{71.53 \pm 0.18}$	47.93 ± 0.68	$\mathbf{84.69 \pm 0.15}$
High $\rightarrow$ Low	$\mathbf{3.32 \pm 0.07}$	63.48 ± 0.15	$\mathbf{8.23 \pm 0.05}$	77.25 ± 0.08

Table 2: Effect of sample ordering under batch size 32 and learning rate 5e-5 for 1 epoch on TinyImageNet (ViT-B/16).

Method	Forget Train Acc ↓	Retain Train Acc ↑	Forget Test Acc ↓	Retain Test Acc. ↑
Original	95.20	79.91	96.00	79.72
Random	19.33 ± 3.40	78.08 ± 0.08	23.33 ± 0.94	77.86 ± 0.09
Low $\rightarrow$ High	75.40 ± 1.02	$\mathbf{79.44 \pm 0.01}$	84.00 ± 0.00	$\mathbf{79.20 \pm 0.04}$
High $\rightarrow$ Low	$\mathbf{0.00 \pm 0.00}$	72.37 ± 0.27	$\mathbf{0.00 \pm 0.00}$	72.37 ± 0.18

3.2 Unlearning Trajectories Over Epochs

To examine how sample ordering influences the *dynamics* of ascent-based unlearning under a fixed optimization budget, we track forget and retain test accuracy over multiple unlearning epochs.

Figure 1 shows the resulting trajectories on CIFAR-10 for three ordering strategies: Random, Low→High, and High→Low.

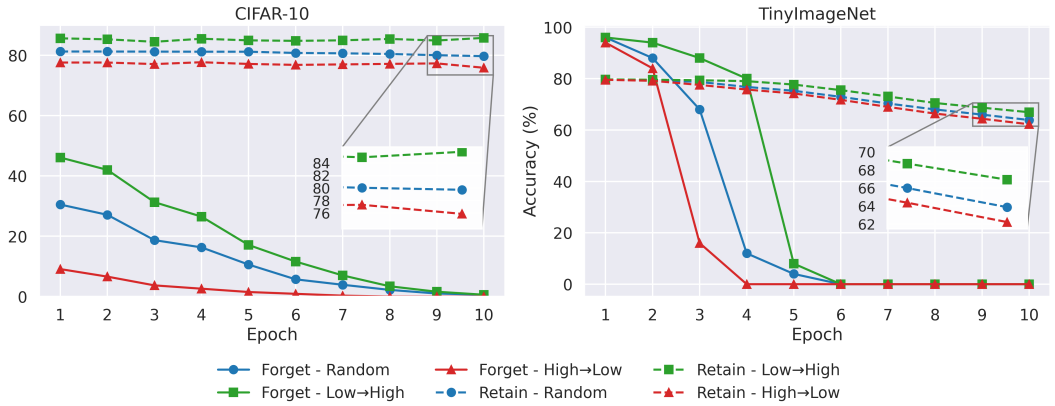

Figure 1: Unlearning trajectories over 10 epochs on CIFAR-10 (left) and TinyImageNet (right).

Across all strategies, forget accuracy decreases monotonically with additional unlearning steps, indicating that all orderings are capable of driving effective forgetting given sufficient optimization. However, the *rate* at which forgetting is achieved and the accompanying degradation in retain performance depends strongly on the ordering of forget samples. High→Low ordering rapidly drives forget accuracy to near zero within a few epochs, but induces substantial early drift that degrades retain accuracy. In contrast, Low→High ordering produces a slower, more conservative forgetting trajectory that preserves retain performance over a larger number of epochs.

These trajectories demonstrate that ascent-based unlearning is fundamentally path-dependent: even when the final forgetting objective is attainable under all orderings, the intermediate trajectory and thus the forgetting–retention trade-off realized under practical unlearning budgets is highly sensitive to sample ordering. From this perspective, sample ordering acts as an implicit regularization mechanism that controls the speed of forgetting relative to collateral damage on retained data, providing a simple and effective lever for navigating the forgetting–retention frontier without modifying the underlying optimization procedure. In practical settings where unlearning budgets are constrained by utility or deployment requirements, these differences in trajectory shape can lead to qualitatively different outcomes despite identical asymptotic objectives.

We have conducted extensive sensitivity analyses with respect to the unlearning batch size, learning rate, and choice of forget class. Across all settings, structured sample orderings consistently induce predictable and directional effects on forgetting and retention, whereas random ordering exhibits high variance and unpredictable outcomes. In addition to epistemic-uncertainty–based curricula, we evaluated alternative ordering proxies based on per-sample loss and gradient norm, and observed the same qualitative path-dependent behavior. Detailed results, including class-wise trends across all CIFAR-10 classes, batch-size sweeps and learning-rate sweeps are reported in Appendix A.2.

4 Conclusion

In this work, we identified sample ordering as a previously overlooked factor in ascent-based machine unlearning. Through experiments across datasets and architectures, we showed that unlearning is fundamentally path-dependent: even under identical optimization budgets, different orderings of the same forget samples yield different forgetting–retention trade-offs. Sample ordering alone can expose extremal behaviors under fixed budgets, ranging from aggressive forgetting with retention degradation to more conservative forgetting with strong retention. While random ordering explores this space stochastically, structured orderings, for example, by epistemic uncertainty provides a simple and predictable way to shape unlearning dynamics without modifying the underlying algorithm. More broadly, our results highlight the need for trajectory-aware design and evaluation of unlearning methods.

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, pp. 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, Baihe Zhang, Daphne Shi, and Nicholas Carlini. Machine unlearning. *IEEE Symposium on Security and Privacy (S&P)*, 2021.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless machine: Forgetting in deep learning. *arXiv preprint arXiv:2003.02960*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 1957–1987. Curran Associates, Inc., 2023.

Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge, 2015. URL https://api.semanticscholar.org/CorpusID:16664790.

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

Ioannis Mavrothalassitis, Pol Puigdemont, Noam Itzhak Levi, and Volkan Cevher. Ascent fails to forget. *arXiv preprint arXiv:2509.26427*, 2025.

Hongsheng Xu et al. Machine unlearning: A survey. *ACM Computing Surveys*, 57(1):1–36, 2024.

A APPENDIX

A.1 LOCAL ANALYSIS OF ORDER SENSITIVITY AND UNCERTAINTY-GUIDED UNLEARNING

This appendix provides a local, mechanistic analysis explaining (i) why ascent-based unlearning procedures are inherently order-sensitive, and (ii) why ordering forget samples by predictive uncertainty serves as an effective practical heuristic. Our goal is not to establish global guarantees, but to isolate a dominant second-order mechanism that qualitatively explains observed empirical behavior.

A.1.1 SETUP AND NOTATION

Let $\theta_0 \in \mathbb{R}^d$ denote the parameters of a trained neural network. For a data sample $z = (x, y)$, let $\ell(\theta; z)$ be a twice continuously differentiable loss. Given a forget set $\mathcal{F} = \{z_1, \ldots, z_n\}$, ascent-based unlearning applies updates of the form

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta \ell(\theta_t; z_{\pi(t)}), \tag{3}$$

where $\eta > 0$ is a fixed step size and π denotes an ordering over forget samples.

We analyze how the final parameters depend on π under a fixed optimization budget and identical hyperparameters.

A.1.2 LOCAL SECOND-ORDER APPROXIMATION

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

$$\nabla \ell(\theta; z_i) \approx g_i + H_i(\theta - \theta_0), \tag{4}$$

where

$$g_i = \nabla \ell(\theta_0; z_i), \qquad H_i = \nabla^2 \ell(\theta_0; z_i).$$

This approximation is local and intended to capture dominant second-order interactions over short unlearning trajectories. Beyond this regime, higher-order terms introduce additional non-commutative effects rather than eliminating path dependence.

A.1.3 NON-COMMUTATIVITY OF ASCENT UPDATES

Consider two forget samples z_i and z_j. Applying ascent on z_i followed by z_j yields

$$\theta_{ij} = \theta_0 + \eta g_i + \eta \left(g_j + H_j(\theta_0 + \eta g_i - \theta_0)\right) \tag{5}$$

$$= \theta_0 + \eta(g_i + g_j) + \eta^2 H_j g_i. \tag{6}$$

Reversing the order gives

$$\theta_{ji} = \theta_0 + \eta g_j + \eta \left(g_i + H_i(\theta_0 + \eta g_j - \theta_0)\right) \tag{7}$$

$$= \theta_0 + \eta(g_j + g_i) + \eta^2 H_i g_j. \tag{8}$$

The resulting difference is

$$\theta_{ij} - \theta_{ji} = \eta^2 (H_j g_i - H_i g_j). \tag{9}$$

Equation equation 9 shows that ascent-based unlearning updates are generally non-commutative. Order sensitivity arises from second-order interactions between gradients and local curvature, and disappears only in degenerate cases (e.g., linear models or vanishing curvature).

A.1.4 MAGNITUDE OF ORDER-INDUCED DRIFT

Taking norms in Eq. equation 9 yields

$$\|\theta_{ij} - \theta_{ji}\| \leq \eta^2 \left(\|H_j\| \, \|g_i\| + \|H_i\| \, \|g_j\|\right). \tag{10}$$

This bound highlights that sensitivity to ordering is governed by interactions between gradient magnitudes and local curvature. In particular, updates involving large gradients tend to induce larger path-dependent drift, especially when applied early in the unlearning trajectory. While stochastic descent is also non-commutative, ascent amplifies path-dependent drift by systematically moving away from a locally stabilizing basin.

A.1.5 Predictive Uncertainty and Gradient Magnitudes

For standard classification losses, the magnitude of the ascent gradient depends on the model's prediction confidence. In particular, under common settings such as softmax classifiers with cross-entropy loss, uncertain or misclassified samples typically induce larger and more variable ascent gradients than confidently predicted samples.

As a result, predictive uncertainty serves as a noisy but informative proxy for the expected magnitude and instability of g_i. We emphasize that uncertainty is not a definition of memorization, but a correlated and readily available signal that biases ascent updates toward more aggressive early parameter displacements when applied first.

A.1.6 Cumulative Effects and Ordering

For a general forget set, sequential ascent updates accumulate second-order interaction terms of the form

$$\sum_{t<s} \eta^2 H_{\pi(s)} g_{\pi(t)}. \tag{11}$$

Ordering forget samples by decreasing predictive uncertainty tends to induce larger gradients earlier in the trajectory, amplifying non-commutative effects and leading to less stable unlearning behavior. Conversely, high-confidence samples are applied in early updates toward smaller gradients, thereby attenuating the magnitude of early interaction terms.This reduces cumulative path-dependent drift under a fixed optimization budget.

A.1.7 Implications for Retain Performance

Let $\mathcal{R}$ denote the retain dataset with aggregate loss $\mathcal{L}_{\mathcal{R}}(\theta)$. When θ remains close to θ_0, a second-order expansion yields

$$\mathcal{L}_{\mathcal{R}}(\theta) \approx \mathcal{L}_{\mathcal{R}}(\theta_0) + \frac{1}{2}(\theta - \theta_0)^\top H_{\mathcal{R}}(\theta - \theta_0). \tag{12}$$

This approximation illustrates why excessive parameter drift during unlearning is often associated with degradation in retain performance. By reducing early drift without modifying the underlying optimization procedure, uncertainty-guided ordering helps preserve retain accuracy while achieving effective forgetting.

A.1.8 Summary

In summary, ascent-based unlearning is inherently path-dependent due to second-order interactions between gradients and local curvature. Predictive uncertainty provides a simple, zero-cost heuristic that correlates with gradient magnitude and can be used to order forget samples. Uncertainty-guided ordering attenuates early path-dependent drift and yields more stable forgetting–retention trade-offs in practice.

A.2 Additional Experimental 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.

The results in Table 3 illustrate that unlearning effectiveness varies significantly across classes and ordering strategies. High $\rightarrow$ Low consistently achieves the lowest forget accuracies, demonstrating the most aggressive unlearning, but often at the expense of lower retain accuracies. In contrast, Low $\rightarrow$ 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.

Table 3: 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

Table 4: Effect of sample ordering under batch size 32 and learning rate 5e-6 for 1 epoch on CIFAR-10 (ViT-B/16).

Method	Forget Train Acc.	Retain Train Acc.	Forget Test Acc.	Retain Test Acc.
Original	92.44	91.48	98.50	97.40
Random Ordering	8.06 ± 0.10	81.35 ± 0.54	14.85 ± 0.72	93.51 ± 0.29
Low → High	45.15 ± 0.03	$\mathbf{89.76 \pm 0.02}$	64.80 ± 0.81	$\mathbf{96.88 \pm 0.03}$
High → Low	$\mathbf{0.00 \pm 0.00}$	52.74 ± 1.70	$\mathbf{0.00 \pm 0.00}$	69.90 ± 1.96

A.3 Sensitivity to hyperparameters

A.3.1 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 indicate that across batch sizes, uncertainty-guided ordering maintains a clear advantage over random ordering. For small and moderate batch sizes, uncertainty-decreasing ordering achieves substantially lower forget accuracy, while uncertainty-increasing ordering preserves retain performance but fails to forget. As batch size increases, the differences between orderings diminish, consistent with the reduced sequentiality of updates and weaker path dependence in the large-batch regime. This trend supports our theoretical analysis, which predicts that order sensitivity arises from sequential gradient–curvature interactions that are attenuated when updates are aggregated.

A.3.2 Effect of Learning Rate

We further examine the effect of learning rate on order sensitivity by sweeping unlearning rates across several orders of magnitude. For very large learning rates, all orderings lead to catastrophic degradation of both forget and retain performance, reflecting unstable optimization. For very small learning rates, unlearning progress is limited regardless of ordering.

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.4 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

Table 5: Effect of sample ordering under batch size 32 and learning rate 5e-6 for 1 epoch on Tiny-ImageNet (ResNet18).

Method	Forget Train Acc.	Retain Train Acc.	Forget Test Acc.	Retain Test Acc.
Original	93.20	80.38	92.00	73.62
Random Ordering	2.50 ± 0.30	70.16 ± 0.14	2.85 ± 0.21	63.73 ± 0.04
Low $\rightarrow$ High	7.90 ± 0.50	$\mathbf{72.16 \pm 0.05}$	6.00 ± 0.00	$\mathbf{65.28 \pm 0.06}$
High $\rightarrow$ Low	$\mathbf{0.00 \pm 0.00}$	68.71 ± 0.05	$\mathbf{0.00 \pm 0.00}$	62.40 ± 0.13

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 $\rightarrow$ High	High $\rightarrow$ Low	Random	Low $\rightarrow$ High	High $\rightarrow$ 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

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.

A.4.1 Ordering Proxies.

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

Learning Rate	**Forget Test Accuracy**			**Retain Test Accuracy**		
	Random	Low $\rightarrow$ High	High $\rightarrow$ Low	Random	Low $\rightarrow$ High	High $\rightarrow$ 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

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

A.4.2 DISCUSSION.

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.

These results indicate that the path-dependent effects identified in this work are driven primarily by the *ordering of forget samples itself*, rather than by the specific proxy used to construct the ordering. Epistemic uncertainty is therefore best viewed as one convenient signal among several that correlate with update magnitude, rather than as a necessary component of the proposed analysis.