Understanding Machine Unlearning Through the Lens of Mode Connectivity

Jiali Cheng, Hadi Amiri
University of Massachusetts Lowell
{jiali_cheng, hadi_amiri}@uml.edu

Abstract

Machine Unlearning aims to remove undesired information from trained models without full retraining from scratch. Despite recent progress, the loss landscape and optimization geometry of unlearning are poorly understood. In this paper, we study machine unlearning through the lens of mode connectivity–the phenomenon that independently trained models can often be connected by smooth low-loss paths in parameter space. We introduce *mode connectivity in unlearning* (MCU) and evaluate it across a range of settings, including curriculum learning, second-order optimization, and connectivity across different unlearning methods. We find that many unlearned models lie in connected basins with smooth retain/forget behavior, while changes in training dynamics can move solutions into different basins. MCU also reveals that models within the same basin can differ substantially on privacy metrics, and that unlearning progresses nonlinearly from the original model to the unlearned model. In addition, linear connectivity suggests that most approximate unlearning methods are mechanistically distinct from retraining. Finally, MCU-based ensembling can improve generalization and robustness to relearning attacks, and MCU smoothness correlates with unlearning difficulty. To our knowledge, this is the first study of machine unlearning through the lens of mode connectivity[1].

1 Introduction

The widespread deployment of machine learning models raises the need for *machine unlearning*–the process of removing specific knowledge from a trained model without affecting other knowledge (Bourtoule et al., 2021a; Liu et al., 2024a). This need is driven by both legal and ethical imperatives, such as removing copyrighted data (Eldan & Russinovich, 2023), as well as practical necessity of purging outdated or incorrect information (Dhingra et al., 2022). As models scale in size and training cost, understanding unlearning methods is becoming an important research frontier in trustworthy machine learning.

Concurrently, the phenomenon of *mode connectivity* has shown that independently trained models can often be connected by low-loss paths in parameter space (Garipov et al., 2018; Qin et al., 2022) (Figure 1a). These findings have important implications for understanding loss landscape, model ensembling, and generalization (Garipov et al., 2018; Zhao et al., 2020).

However, existing studies on mode connectivity have largely focused on image classification tasks (Draxler et al., 2018; Vrabel et al., 2025) with simple optimization regimes. Whether mode connectivity extends to unlearning remains unexplored. More broadly, despite recent advances (Liu et al., 2024b; Hong et al., 2024b), the loss landscape of unlearning remains poorly understood, along with its implications for the behavior and properties of unlearning.

To bridge these gaps, we introduce and formalize the concept of **Mode Connectivity in Unlearning (MCU)**–a framework to study unlearning from the loss landscape perspective. Specifically, we investigate the following research question:

[1]Code: https://github.com/CLU-UML/Mode-Connectivity-in-Unlearning

RQs: *What can MCU reveal about unlearning methods, including loss landscape geometry, mechanistic similarity, and robustness to relearning attacks?*

Unlike prior mode connectivity work, which studies standard training under a single objective, MCU must simultaneously preserve performance on the retain set and induce forgetting on the forget set. Answering this question provides insight into the generalization, stability, and interpretability of unlearning methods. For instance, a smooth path between two unlearning solutions that maintains strong retain behavior and forgetting may indicate shared inductive biases or similar optimization dynamics, while a lack of connectivity may indicate divergent solution structures.

Through extensive experiments on diverse tasks and different training paradigms, we find that unlearning has a smooth loss landscape. Unlearned models with different optimizations tend to converge into the same smooth basin (§5.1–§5.6). Models within the same low-loss basin can function drastically different, especially on privacy-related metrics (§5.2). We discover that along the change from original model to unlearned model in parameter space, unlearning behavior is not evenly distributed, and first suppresses memorization of textual content, while later commits to deeper forgetting (§5.3). MCU also demonstrates that unlearning is generally mechanistically different from retraining. Additionally, MCU-based ensemble can improve generalization and robustness of unlearning (§5.5). Finally, we show that MCU smoothness is a proxy for unlearning difficulty (§5.6). These insights provide new directions for future unlearning research. Our contributions are:

- **Mode Connectivity in Unlearning (MCU)**: We introduce MCU–a framework for studying machine unlearning through loss landscape geometry.
- **Analysis of Unlearning Loss Landscape**: We evaluate MCU across LLM and classification unlearning benchmarks, multiple unlearning algorithms, and diverse training settings including curriculum learning, second-order optimization, and cross-method connectivity.
- **Insights into Unlearning**: MCU reveals mechanistic differences between unlearning and retraining, functional heterogeneity within connected basins, and a link between connectivity smoothness and unlearning difficulty. We also show that MCU-based ensembling improves generalization and robustness to relearning attacks.

Overall, our results suggest that MCU is useful both as a diagnostic tool for understanding unlearning geometry and as a practical tool for improving robustness and model selection.

2 Preliminaries

Notation Let f_{θ_o} be a model trained on dataset $\mathcal{D}$ with task loss L. In addition, assume that $\mathcal{D}$ can be divided into two disjoint sets: the forget set $\mathcal{D}_f$ and the retain set $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$.

Machine Unlearning Machine unlearning aims to remove the influence of the forget set $\mathcal{D}_f$ from the trained model f_{θ_o} and preserve the knowledge of retain set $\mathcal{D}_r$. A good unlearning model f' should achieve high loss on $\mathcal{D}_f$ and low loss on $\mathcal{D}_r$. A commonly used solution is to fine-tune the original model f_{θ_o} to minimize the task loss on $\mathcal{D}_r$ while maximizing the task loss on $\mathcal{D}_f$ (Jia et al., 2024; Liu et al., 2024a; Cheng & Amiri, 2026). For example, GradDiff (Maini et al., 2024) directly implements the above approach:

$$f' = \arg\min_{\theta'} L(\mathcal{D}_r) - L(\mathcal{D}_f). \tag{1}$$

Details of related work and unlearning methods are discussed in Appendix A and B.

Mode Connectivity Let θ_1 and θ_2 denote the weights of two independently trained models on some dataset $\mathcal{D}$ using loss L. The objective of mode connectivity is to find a curve $\phi(t) \to \mathbb{R}^{|\theta|}, t \in [0,1]$ in the parameter space that connects the two minimizers θ_1 and θ_2, where $\phi(0) = \theta_1$ and $\phi(1) = \theta_2$. Curve $\phi(t)$ connecting θ_1 and θ_2 satisfies mode connectivity

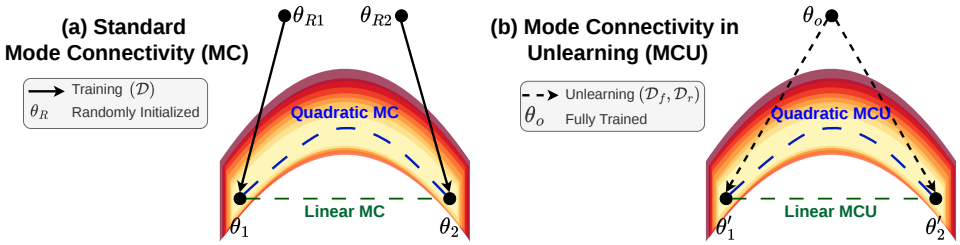


Figure 1: **(a)**: Illustration of standard mode connectivity (MC): MC finds a smooth curve connecting two endpoints that yields consistent low loss on $\mathcal{D}$. **(b)**: Illustration of mode connectivity in unlearning (MCU): unlearning removes knowledge of forget set $\mathcal{D}_f$ from the trained model f_{θ_o} while maintaining knowledge of retain set $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$. MCU finds a smooth curve connecting the two unlearned models θ'_1 and θ'_2 that yields consistent low loss on $\mathcal{D}_r$ and high loss on $\mathcal{D}_f$. See details in § 3.

if the path $\phi(t)$ does not yield "barriers," defined as sudden increase in loss (Garipov et al., 2018; Lubana et al., 2023). Formally, $\forall t \in [0,1]$:

$$L\big(\mathcal{D}; \phi(t)\big) \leq (1-t) \cdot L(\mathcal{D}; \theta_1) + t \cdot L(\mathcal{D}; \theta_2). \tag{2}$$

In the loss landscape, mode connectivity tries to find a low loss path ϕ connecting θ_1 and θ_2 without hitting any barrier. In other words, every set of parameter induced by $\phi(t)$ yield comparable performance to the minimizers θ_1 and θ_2. The parametrization of ϕ determines the shape of the curve connecting the two minimizers θ_1, θ_2. Below, we present two commonly used curve types:

- **Linear:** a linear interpolation of minimizers with no optimization involved, i.e. $\phi(t) = (1-t)\theta_1 + t\theta_2$. Stronger linear connectivity indicates stronger mechanistic similarity of minimizers, such as their inductive biases (Lubana et al., 2023).
- **Quadratic:** a smooth quadratic curve connecting two minimizers, i.e $\phi(t) = (1-t)^2\theta_1 + 2t(1-t)\theta_{12} + t^2\theta_2$, where θ_{12} needs to be trained explicitly.

To find non-linear curve, we can minimize the total accumulated loss along the curve to find the midpoint θ_{12} which will later be used for interpolation. More details of the curve finding optimization are discussed in Appendix D.

3 Mode Connectivity in Unlearning (MCU)

Definition 1 (Mode Connectivity in Unlearning)**.** *As illustrated in Figure 1, let θ'_1 and θ'_2 denote the weights of two unlearned models of applying some unlearning procedure $\mathcal{U}$ on the original model f_{θ_o} with different configurations. MCU holds if there exists a path $\phi_{\theta'_1 \to \theta'_2}(t)$ in parameter space that connects θ'_1 and θ'_2 without yielding barriers. Formally, $\forall t \in [0,1]$*

$$L\big(\mathcal{D}_r; \phi(t)\big) \leq (1-t) \cdot L(\mathcal{D}_r; \theta'_1) + t \cdot L(\mathcal{D}_r; \theta'_2), \tag{3}$$

$$L\big(\mathcal{D}_f; \phi(t)\big) \geq (1-t) \cdot L(\mathcal{D}_f; \theta'_1) + t \cdot L(\mathcal{D}_f; \theta'_2). \tag{4}$$

In MCU, "barriers" on the retain set $\mathcal{D}_r$ refer to the sudden *increase* of task loss L (as in standard mode connectivity), while "barriers" on the unlearn set $\mathcal{D}_f$ refer to the sudden *decrease* of task loss. Eq. 3 ensures that the task loss on the retain set $\mathcal{D}_r$ remains both low and smooth along the mode connectivity curve, indicating consistent model behavior during the unlearning process. Similarly, Eq. 4 enforces a high and smooth loss on the forget set $\mathcal{D}_f$ along the mode connectivity path. In other words, MCU is realized when there exists a continuous path of model weights connecting the endpoints θ'_1 and θ'_2, such that loss remains low on $\mathcal{D}_r$ and high on $\mathcal{D}_f$ along the curve.

Table 1: MCU settings. CL and SO (FO) denote curriculum learning and second-order (first-order) optimization respectively. All settings except "Rand" are novel in mode connectivity.

Endpoints unlearned with	**Standard**	**Both CL**	**Both SO**	**Mixed CL/Non-CL**	**Mixed FO/SO**
Different Randomness	Rand	Rand-CL	Rand-SO	CL-Non-CL	FO-SO
Different unlearning methods	Met	Met-CL	Met-SO	Met-CL-Non-CL	Met-FO-SO

Connection to Standard Mode Connectivity In contrast to standard mode connectivity, MCU must satisfy objectives on both $\mathcal{D}_f$ and $\mathcal{D}_r$. Essentially, MCU examines whether it is possible to find a continuous curve between two endpoints such that there are no significant loss barriers in **two distinct loss landscapes**–a key difference compared to standard mode connectivity, which typically considers only a single task or dataset. Another key difference is that in standard mode connectivity (Garipov et al., 2018), the endpoints θ_1 and θ_2 are obtained by training the model from two random initializations. In contrast, the endpoints θ_1' and θ_2' in MCU are both derived from the exact same trained model θ_o.

3.1 Influence of Training Dynamics on Mode Connectivity in Unlearning

Unlearned models can be trained by various paradigms, such as curriculum learning (CL) (Bengio et al., 2009; Barbulescu & Triantafillou, 2024; Zhao et al., 2024; Cheng & Amiri, 2024; 2025c) and second-order (SO) optimization (Jia et al., 2024; Zhang et al., 2024a). To systematically analyze the impact of these factors, we define several novel experimental conditions and evaluate MCU under these conditions.

For CL, we define two settings: (a): both endpoints are trained with CL (**-CL**); and (b): one endpoint is obtained through CL and the other does not (**CL-Non-CL**). Similar settings apply to SO. These configurations allow us to assess whether changes in sample learning order affect the emergence of MCU.

In addition, we examine whether endpoints derived from different unlearning methods can be smoothly connected. We hypothesize that methods with similar formulation and inner mechanisms are more likely to establish connectivity. This experiment provides a lens into mechanistic similarity of different unlearning methods (Lubana et al., 2023).

To the best of our knowledge, with the exception of the randomness factor, all other factors discussed above are novel within the mode connectivity literature. Multiple configurations can be combined, as summarized in Table 1. Together, they provide diverse and realistic perspectives on the training dynamics of unlearning, broaden the scope of mode connectivity research, and deepen our understanding of the factors that enable or prevent successful unlearning.

4 Experimental Setup

Datasets and Forget Sets We analyze MCU on widely adopted LLM unlearning and classification unlearning benchmarks. For LLM unlearning, we use TOFU (Maini et al., 2024), MUSE (Shi et al., 2025), and WMDP (Li et al., 2024) dataset. For classification, we use three datasets from MU-Bench (Cheng & Amiri, 2024): image classification on CIFAR-10 (Krizhevsky, 2009), biomedical text relation classification on DDI2013 (Segura-Bedmar et al., 2013), and image-text visual entailment on NLVR2 (Suhr et al., 2019). The original models and standard data splits are provided by MU-Bench (Cheng & Amiri, 2024; 2025a) and open-unlearning (Dorna et al., 2025). Specifically, TOFU has 1%, 5%, 10% of forget set. MUSE has `Books` and `News` as forget sets, while WMDP has `Cyber-security` (Cyber) as forget sets. MU-Bench provides 2%, 4%, 6%, 8%, 10% forget set.

Unlearning Methods We use MCU to analyze the following LLM unlearning methods: 1) Gradient Ascent (GA) (Golatkar et al., 2020), 2) GradDiff (GD) (Maini et al., 2024), 3)

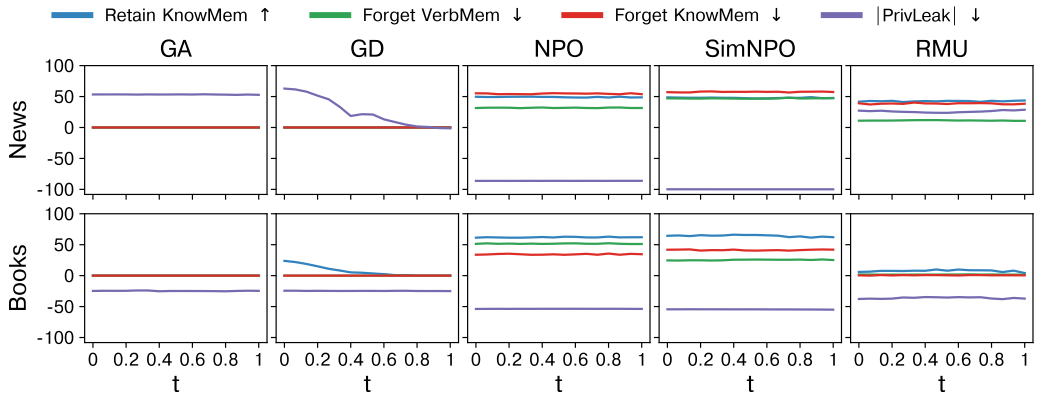


Figure 2: MCU under Rand setting on MUSE dataset. Additional results are shown in Appendix E Figure 9–23.

Negative Preference Optimization (NPO) (Zhang et al., 2024b), 4) SimNPO (Fan et al., 2024b), and 5) RMU (Li et al., 2024). For classification tasks we use: 1) Gradient Ascent (GA) (Golatkar et al., 2020), 2) Random Labeling (RL) (Graves et al., 2021), 3) Bad Teaching (BT) (Chundawat et al., 2023), and 4) Saliency Unlearning (SU) (Fan et al., 2024c). These methods cover a diverse set of unlearning paradigms and are commonly used in existing works. Appendix B provides additional details.

Evaluation To evaluate MCU, we sample multiple points by varying the interpolation weight $t \in [0, 1]$ with small step size. Each value of t induces a set of model weights according to the parametrization of the curve $\phi_\theta(t)$ (§ 2). Following previous mode connectivity work on language models (Qin et al., 2022), we sample 16 points with equal step size in $[0, 1]$. For each induced model $\theta_\phi(t)$, we evaluate its performance using the standard benchmark-specific metrics. On TOFU, we use Forget Quality (↑) and Model Utility (↑) which is a p-value of Kolmogorov-Smirnov test (KS-Test). On MUSE, we use Forget VerbMem (↓), Forget KnowMem (↓), Retain KnowMem (↑), and privacy leakage ($|\text{PrivLeak}|$, ↓). On WMDP, we use accuracy on forget set (↓) and accuracy on MMLU evaluation set (↑). Appendix C provides additional details and metrics. On classification unlearning tasks, we use accuracy on test set D_t(↑), accuracy on forget set D_f(↓), accuracy on retain set D_r(↑), and Zero-Retrain Forgetting score ZFR (↑) which measures the prediction similarity of D_f between unlearned and original models.

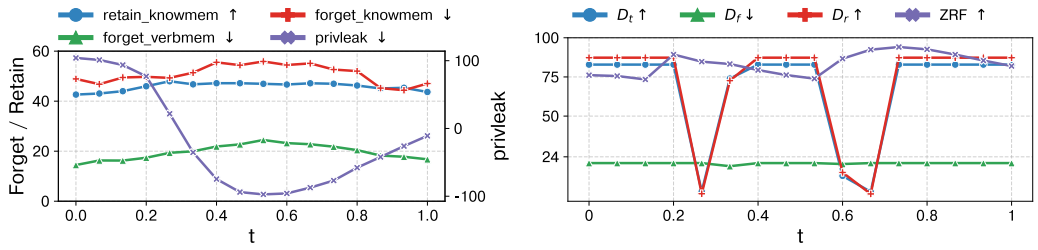
5 Results

With MCU, we study what geometric regimes arise in unlearning, and what they reveal about the solution space (§5.1–5.6). Then we leverage MCU to analyze unlearning from different perspectives, including 1) parameter displacement from original model (§5.3), 2) mechanistic similarity of unlearning methods (§5.4), 3) robustness to attacks (§5.5), and 4) unlearning difficulty (§5.6).

5.1 Unlearning Has Smooth Loss Landscape

We investigate the conditions under which mode connectivity in unlearning (MCU) emerges across different models, datasets, and optimization strategies. Our results show that MCU is often prevalent.

We observe smooth curves with no fluctuation of unlearning quality (Forget VerbMem & KnowMem) and utility retention (Retain KnowMem) on MUSE when both endpoints are unlearned with different random seeds, shown in Figure 2. MCU curves on NPO, SimNPO show consistently high forget quality (low forget ROUGE) and utility retention



(a) MCU on MUSE-News with RMU Unlearning (b) MCU on DDI with BT Unlearning

Figure 3: Functionally heterogeneous basin. Models within the same basin in loss landscape can have different functional performance in different dimensions. (a) On MUSE with RMU, performance on $\mathcal{D}_f$ and $\mathcal{D}_r$ remains similar, but PrivLeak has large fluctuations within the range $[-100, 100]$. (b) On DDI with BT, performance on $\mathcal{D}_f$ and ZRF remains stable, whereas $\mathcal{D}_t$ and $\mathcal{D}_r$ vary substantially.

Table 3: Average barrier of $\mathcal{D}_f$ metrics of different MCU configurations.

Barrier	Rand	Rand-CL	Rand-SO	CL-Non-CL	FO-SO
Mean	1.2	0.7	0.2	2.0	1.8
Max	3.0	2.2	1.2	8.8	8.1

(high retain KnowMem), which suggests that unlearning solutions reside on a connected low-loss manifold. This means we find a basin with consistent high unlearning efficacy. In other cases on GA, GD, RMU, smooth curves can appear between low quality unlearned models, i.e. a basin of low unlearning performance. This observation aligns with findings in standard mode connectivity (Draxler et al., 2018), where minima are not isolated but from a single connected manifold of low loss in parameter space, see Figure 2. The existence of mode connectivity paths suggests that modern neural networks have enough parameters such that they can achieve good predictions while a big part of the network undergoes structural changes.

Scaling We study the change of loss landscape geometry along two scaling axes: 1) increasing forget set, and 2) increasing model size. We define two metrics to quantify the barrier of MCU curve. 1) *Max barrier* measures the largest interior rise of the curve above the worse endpoint. Given a scalar curve $f_{\theta(t)}, t \in [0,1]$, the max barrier is $B_{\max} = \max_{t\in[0,1]} f_{\theta(t)} - \max\{f_{\theta_0}, f_{\theta_1}\}$. 2) *Mean barrier* measures the average positive excess above the worse endpoint, i.e. $B_{\text{mean}} = \int_0^1 \max\left(f_{\theta(t)} - \max\{f_{\theta_0}, f_{\theta_1}\}, 0\right) dt$. We only compute the average barrier on all $\mathcal{D}_f$ metrics, since the curve on $\mathcal{D}_r$ has minimal fluctuations in most cases.

Table 2: Scaling of MCU smoothness.

$\lvert\mathcal{D}_f\rvert$	Max Barrier			Mean Barrier		
	1B	3B	8B	1B	3B	8B
1%	18.1	9.6	7.5	4.1	3.0	2.8
5%	8.2	5.9	3.0	0.9	0.8	0.8
10%	2.9	1.7	0.0	0.3	0.1	0.0

Table 2 shows a consistent pattern: MCU becomes more smooth as the forget set or the model scales up. However, the smoothness comes at the cost of bad unlearning quality. This suggests that scaling unlearning more knowledge on larger models is more challenging.

CL and SO Make Unlearning Converge to Different Basins We investigate if varying training paradigms can make unlearning methods converge to different basins. Note training paradigms do not change the unlearning algorithm (loss function). Therefore, the loss landscape remains unchanged for each algorithm.

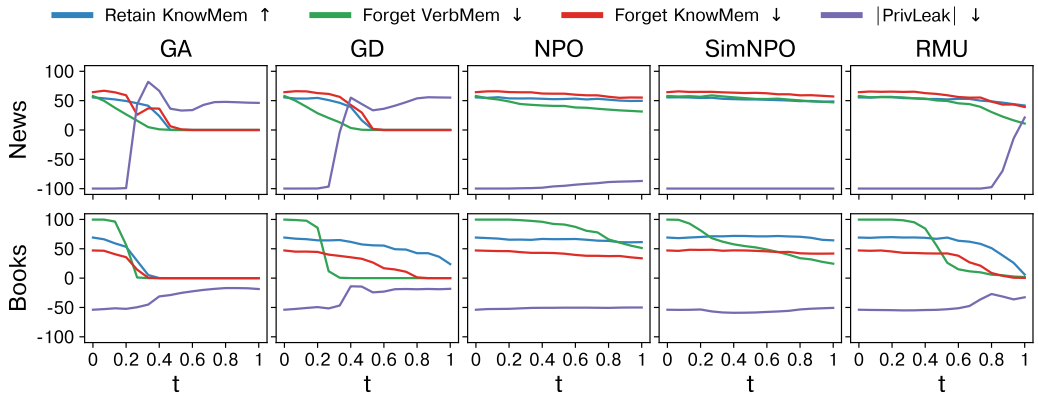


Figure 4: Linear interpolation between the original model θ_o and the unlearned model θ_u on MUSE dataset. Additional results are shown in Appendix E Figure 9–23.

These results in Table 3 suggest that under CL-Non-CL and FO-SO, the MCU curves show large max barrier, indicating that two endpoints are likely to reside in two different and disconnected basins. When both endpoints are optimized with CL or SO, this behavior does not appear. This indicates that CL and SO can help unlearning converge to different solutions, highlighting potential future research directions.

5.2 Functionally Heterogeneous Basin

We notice functional smoothness can differ significantly across different dimensions. Even minimizers reside in the same basin with similar low loss, their functional performances can vary significantly across different dimensions.

For example, with RMU on MUSE-News, we observe a relatively smooth MCU curve, with trivial performance variation on Forget VerbMem, Forget KnowMem, and Retain KnowMem. However, performance on PrivLeak can fluctuate significantly along the same MCU curve, ranging from +100 to -100, see Figure 3(a). This indicates that although close to each other in parameter space, some minimizers are extremely prone to attackers and may leak information, while others may be much more robust. Another example is with BT on DDI. We observe stable accuracy on $\mathcal{D}_f$ along the MCU curve, but significant fluctuations on $\mathcal{D}_r$, $\mathcal{D}_t$ and moderate fluctuation on ZRF, see Figure 3(b). This indicates that although close to each other in parameter space, some minimizers maintain good amount of knowledge from the original model ($\mathcal{D}r$ and $\mathcal{D}_t$), while others may have completely forgotten.

On retain set $\mathcal{D}_r$, smooth connectivity are more likely to occur and easier to find. This could be because $\mathcal{D}_r$ is typically much larger than $\mathcal{D}_f$, which leads to a more stable optimization signal and a smoother curve in the retain region of the loss landscape.

This behavior demonstrates that although residing in the same basin with similar loss, the unlearning behaviors can differ significantly in different functional dimensions, particularly for privacy related metrics. This also shows a limitation of current evaluation protocols for unlearning and motivates the need for richer and intrinsic evaluation on model parameters (Hong et al., 2024a).

5.3 Probing the Unlearning Direction

To probe whether parameter displacement along the unlearning direction is meaningful, we apply linear interpolation between the original model θ_o and the unlearned model θ_u, i.e. $\theta(t) = (1-t)\theta_o + t\theta_u, \quad t \in [0,1]$. This probe results in Figure 4 reveals several important properties of unlearning. First, the clear asymmetry between the retain set $\mathcal{D}_r$ and the forget set $\mathcal{D}_f$ indicates the asymmetry of loss landscape.

Non-linearly Distributed Unlearning We observe that the unlearning behavior does not accumulate linearly as t increases. Unlearning does not emerge until deviation is accumulated to certain threshold. In other words, model may remain functionally close to the original model for much of the path, and only near specific regions does strong forgetting activate.

Suppress First, Unlearn Later A recurring pattern is that lexical overlap metrics (e.g. forget ROUGE) improves earlier and more smoothly than overall unlearning efficacy metrics (e.g. forget quality). Similar pattern appear on MUSE in Figure 4, where verbatim memorization (Forget VerbMem) decreases first and knowledge memorization (Forget KnowMem) decreases later. This suggests that unlearning follows a *first suppress, then unlearn* manner, where models first stop reproducing target content in surface form, while remaining relatively competent on the underlying knowledge to forget. As unlearning proceeds to a certain extent, models exhibit deeper level forgetting, such as change of the generated distributions. From this perspective, our analysis provides a complementary geometric lens for understanding how superficial unlearning emerges (Hong et al., 2024a).

5.4 MCU Reveals Mechanistic (Dis)similarity between Unlearning Methods

Lubana et al. (2023) have found that if linear MC exists between two endpoints, they share internal mechanisms or rely on similar features when making predictions. We use this to analyze the mechanistic similarity between retraining and unlearning, and between two different unlearning methods. We find that utility retention is a poor proxy for mechanistic similarity. Most methods (except for GA) maintain relatively stable utility between retraining and unlearning.

Retraining vs. Unlearning In this setting, we probe the linear MCU between a model retrained from scratch ($\theta_1' = \theta_{\text{Retrain}}$) and an unlearned model ($\theta_2' = \theta_{\text{Unlearn}}$). We find that most unlearning methods are not smoothly connected to retraining under forgetting metrics, even when endpoint utility is comparable. This indicates that approximate unlearning often reaches functionally valid endpoints through mechanisms that are distinct from retraining, rather than by approximating the retraining solution along a shared linear subspace. Among the evaluated methods, only NPO demonstrates mechanistic similarity to retraining, shown in Figure 5.

Different Unlearning Methods In this setting, we probe the linear MCU between two different unlearned models. Figure 8 in Appendix E shows the pair-wise linear MCU between all methods. We can see that GA shows smooth MCU on forget set (knowmem and verbmem) with all other unlearning methods, indicating that all methods share similar internal mechanisms of handling forget set samples. They do differ in retain set significantly (Row 1 $t \to 1$), since GA does not have retain mechanism.

We observe high similarity on retain knowledge between NPO and SimNPO, an improved version of NPO. However, they differ from each other on forget knowledge and extraction strengths. Both of them are significantly different from GD, since GD simply minimizes loss on D_r and maximizes loss on D_f, while NPO and SimNPO involves implicit reward modeling.

5.5 MCU-based Ensemble Improves Generalization and Robustness

Generalization Prior work has found that mode connectivity indicates generalization, where minimizers converge to a minima with smooth loss landscape and constant low error, potentially leading to improved performance (Garipov et al., 2018; Wang et al., 2023). Previous experiments demonstrate that intermediate minimizers along the MCU curve may outperform both endpoints. We then propose a generalization method, sampling $N = 3$ minimizers along the curve with equal distance. We average the parameters of the sampled models and two endpoints, resulting in a new merged unlearner. Results on WMDP show that our ensemble strategy can outperform the endpoints in both unlearning

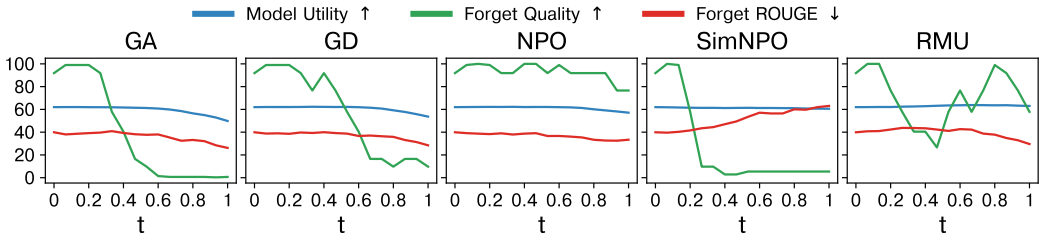


Figure 5: Mechanistic dissimilarity between retraining and unlearning on TOFU. Additional results are shown in Appendix E, Figures 14–18, Figures 25–33.

Table 4: MCU-based ensemble improves generalization.

$\mathcal{D}_f$ Acc (↓)	NPO	+ MCU Ensemble
63.7	24.1	21.5

Table 5: MCU smoothness is a proxy for unlearning difficulty.

Difficulty	Default	Hard	Easy
Max Bar.	18.1	18.5	10.3
Mean Bar.	4.1	6.3	3.2

effectiveness (forget accuracy), see Table 4. This suggests that interpolated models may achieve a better trade-off between forgetting and retaining, and presents a promising directions for ensembling or model selection using mode connectivity. Other examples include RL on NLVR2.

Robustness to Relearning It is well established that a smoother loss landscape is associated with greater robustness in deep learning models (Zhang et al., 2017; Foret et al., 2021; Zhang et al., 2024c; Li et al., 2025). Specifically, minimizers along smooth mode connectivity curves are typically more robust than those obtained by standard training. Building on this insight, our findings suggest that MCU can identify unlearned models that are more resilient to adversarial threats, such as relearning attacks (Hu et al., 2024; Chen et al., 2026b).

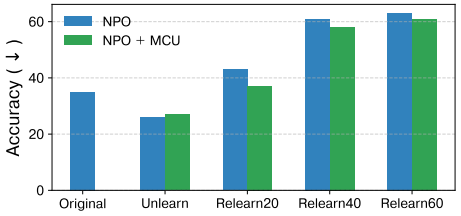


Figure 6: MCU finds minimizers with smooth loss landscape and robust to relearning attack.

Following prior work on unlearning robustness (Fan et al., 2025; Chen et al., 2026a), we attack an unlearned model for WMDP Cyber by fine-tuning it on generic Wikipedia text and subsequently evaluate its unlearning effectiveness post-attack. As shown in Figure 6, models unlearned via standard NPO methods are susceptible to relearning, with their forgetting effects quickly deteriorating. In contrast, models obtained through MCU - by averaging across multiple minimizers - demonstrate markedly greater robustness to such attacks. We attribute this improvement to the smoother loss landscape induced by the averaging process. This observation is consistent with recent work showing that minimizing loss landscape sharpness (encouraging smoothness) during optimization enhances the robustness to adversarial attacks for LLM unlearning (Fan et al., 2025).

5.6 MCU as A Proxy for Unlearning Difficulty

Many prior work has discovered different forget sets can induce substantially different levels of unlearning difficulty. In particular, challenging forget sets may be adversarial (Fan et al., 2024a), closely adjacent to the test set (Cheng et al., 2023; Chen et al., 2025a), highly memorized (Zhao et al., 2024), paraphrased by LLMs (Łucki et al., 2024; Lynch et al., 2024;

Cheng & Amiri, 2025b), mechanistically difficult as characterized by circuits (Cheng et al., 2026), or samples with high forget-retain overlaps (Chen et al., 2025b; Reisizadeh et al., 2026; Chen et al., 2026b;a).

We hypothesize that the smoothness of the MCU basin can serve as a proxy for unlearning difficulty from the perspective of loss landscape geometry. In particular, smoother MCU may reflect a flatter and more navigable loss landscape, whereas irregular or disconnected MCU may indicate sharper geometry, more difficult optimization, and a more challenging forget-retain tradeoff.

To test this, we leverage the easy and hard forget set splits from Cheng et al. (2026) as ground truth difficulty and computed the MCU smoothness of both splits. We find that both the mean and maximum barriers are larger on the hard split than on the easy split. As shown in Table 5, the MCU curve has more barriers with larger maximum barrier on the hard split, while on the easy split, MCU curve is smoother.

6 Conclusion

We introduce mode connectivity in unlearning (MCU) as a framework for understanding the loss landscape and optimization dynamics of machine unlearning. To the best of our knowledge, this is the first work that studies the loss landscape of unlearning with mode connectivity across various settings. We find that the emergence of mode connectivity can be influenced by task complexity, forget set size, and optimization strategies like curriculum learning and second-order methods.

Our experiments across diverse tasks, unlearning methods, and training configurations show that MCU provides a useful lens for analyzing the loss landscape of machine unlearning, and provides insights into mechanistic similarity, robustness, and unlearning difficulty. We further show that MCU can be used as a diagnostic tool and open new directions for improving unlearning methods.

Ethical Considerations

This work focuses on improving the transparency and reliability of machine unlearning, which is motivated by ethical considerations such as user data privacy, regulatory compliance, and the right to be forgotten. All experiments are conducted on publicly available datasets, and no personally identifiable or sensitive data is used.

References

George-Octavian Barbulescu and Peter Triantafillou. To each (textual sequence) its own: Improving memorized-data unlearning in large language models. In *Proceedings of the 40th International Conference on Machine Learning*, 2024.

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.

Gregory Benton, Wesley Maddox, Sanae Lotfi, and Andrew Gordon Gordon Wilson. Loss surface simplexes for mode connecting volumes and fast ensembling. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 769–779. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/benton21a.html.

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)*, pp. 141–159, 2021a. doi: 10.1109/SP40001.2021.00019.

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

Ziheng Chen, Jiali Cheng, Hadi Amiri, Kaushiki Nag, Lu Lin, Sijia Liu, Gabriele Tolomei, and Xiangguo Sun. Frog: Fair removal on graph. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 415–424, 2025a.

Ziheng Chen, Jin Huang, Jiali Cheng, Yuchan Guo, Mengjie Wang, Lalitesh Morishetti, Kaushiki Nag, and Hadi Amiri. Future: Flexible unlearning for tree ensemble. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pp. 4680–4684, 2025b.

Ziheng Chen, Jiali Cheng, Zezhong Fan, Hadi Amiri, Diyuan Wu, Gabriele Tolomei, and Yang Zhang. Tracer: Token reassignment for concept erasure in generative recommendation. *arXiv preprint arXiv:2606.07688*, 2026a.

Ziheng Chen, Jiali Cheng, Zezhong Fan, Hadi Amiri, Yunzhi Yao, Xiangguo Sun, and Yang Zhang. Cure: Circuit-aware unlearning for llm-based recommendation. *arXiv preprint arXiv:2604.04982*, 2026b.

Jiali Cheng and Hadi Amiri. Mu-bench: A multitask multimodal benchmark for machine unlearning. *arXiv preprint arXiv:2406.14796*, 2024.

Jiali Cheng and Hadi Amiri. Multidelete for multimodal machine unlearning. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 165–184, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-72940-9.

Jiali Cheng and Hadi Amiri. Tool unlearning for tool-augmented llms. *arXiv preprint arXiv:2502.01083*, 2025b.

Jiali Cheng and Hadi Amiri. Speech Unlearning. In *Interspeech 2025*, pp. 3209–3213, 2025c. doi: 10.21437/Interspeech.2025-2412.

Jiali Cheng and Hadi Amiri. Machine unlearning across tasks and modalities. In Sijia Liu, Yang Liu, and Nathalie Baracaldo (eds.), *Machine Unlearning for Governance of Foundation Models*, pp. 227–240. Springer Nature Switzerland, Cham, 2026. ISBN 978-3-032-17282-2. doi: 10.1007/978-3-032-17282-2_15. URL https://doi.org/10.1007/978-3-032-17282-2_15.

Jiali Cheng, George Dasoulas, Huan He, Chirag Agarwal, and Marinka Zitnik. GNNDelete: A general strategy for unlearning in graph neural networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=X9yCkmT5Qrl.

Jiali Cheng, Ziheng Chen, Chirag Agarwal, and Hadi Amiri. A mechanistic perspective and circuit-guided difficulty metric for unlearning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 10950–10964, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.532. URL https://aclanthology.org/2026.findings-acl.532/.

Dasol Choi, Soora Choi, Eunsun Lee, Jinwoo Seo, and Dongbin Na. Towards efficient machine unlearning with data augmentation: Guided loss-increasing (gli) to prevent the catastrophic model utility drop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 93–102, June 2024.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7210–7217, Jun. 2023. doi: 10.1609/aaai.v37i6.25879. URL https://ojs.aaai.org/index.php/AAAI/article/view/25879.

Bhuwan Dhingra, Jeremy R. Cole, Julian Martin Eisenschlos, Daniel Gillick, Jacob Eisenstein, and William W. Cohen. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273, 2022. doi: 10.1162/tacl_a_00459. URL https://aclanthology.org/2022.tacl-1.15/.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zachary C Lipton, J Zico Kolter, and Pratyush Maini. Openunlearning: Accelerating llm unlearning via unified benchmarking of methods and metrics. *arXiv preprint arXiv:2506.12618*, 2025.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1309–1318. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/draxler18a.html.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*, 2023.

Chongyu Fan, Jiancheng Liu, Alfred Hero, and Sijia Liu. Challenging forgets: Unveiling the worst-case forget sets in machine unlearning, 2024a.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*, 2024b.

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*, 2024c. URL https://openreview.net/forum?id=gn0mIhQGNM.

Chongyu Fan, Jinghan Jia, Yihua Zhang, Anil Ramakrishna, Mingyi Hong, and Sijia Liu. Towards llm unlearning resilient to relearning attacks: A sharpness-aware minimization perspective and beyond. *arXiv preprint arXiv:2502.05374*, 2025.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3259–3269. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/frankle20a.html.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/be3087e74e9100d4bc4c6268cdbe8456-Paper.pdf.

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

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11516–11524, May 2021. doi: 10.1609/aaai.v35i13.17371. URL https://ojs.aaai.org/index.php/AAAI/article/view/17371.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Tuan Hoang, Santu Rana, Sunil Gupta, and Svetha Venkatesh. Learn to unlearn for deep neural networks: Minimizing unlearning interference with gradient projection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4819–4828, January 2024.

Yihuai Hong, Lei Yu, Haiqin Yang, Shauli Ravfogel, and Mor Geva. Intrinsic evaluation of unlearning using parametric knowledge traces. *arXiv preprint arXiv:2406.11614*, 2024a.

Yihuai Hong, Yuelin Zou, Lijie Hu, Ziqian Zeng, Di Wang, and Haiqin Yang. Dissecting fine-tuning unlearning in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3933–3941, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.228. URL https://aclanthology.org/2024.emnlp-main.228/.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. *arXiv preprint arXiv:2406.13356*, 2024.

Jiabao Ji, Yujian Liu, Yang Zhang, Gaowen Liu, Ramana Rao Kompella, Sijia Liu, and Shiyu Chang. Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference. *arXiv preprint arXiv:2406.08607*, 2024.

Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=0jZH883i34.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. SOUL: Unlocking the power of second-order optimization for LLM unlearning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4276–4292, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.245. URL https://aclanthology.org/2024.emnlp-main.245/.

Aly Kassem, Omar Mahmoud, and Sherif Saad. Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4360–4379, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.265. URL https://aclanthology.org/2023.emnlp-main.265/.

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=xlr6AUDuJz.

Tian Li, Tianyi Zhou, and Jeff Bilmes. Tilted sharpness-aware minimization. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=bwfiFv8KG2.

Zhanran Lin, Puheng Li, and Lei Wu. Exploring neural network landscapes: Star-shaped and geodesic connectivity. *arXiv preprint arXiv:2404.06391*, 2024.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*, 2024a.

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. Protecting privacy in multimodal large language models with mllmu-bench. *arXiv preprint arXiv:2410.22108*, 2024b.

Ekdeep Singh Lubana, Eric J Bigelow, Robert P. Dick, David Krueger, and Hidenori Tanaka. Mechanistic mode connectivity. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 22965–23004. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/lubana23a.html.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An adversarial perspective on machine unlearning for ai safety. *arXiv preprint arXiv:2409.18025*, 2024.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in llms. *arXiv preprint arXiv:2402.16835*, 2024.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. Tofu: A task of fictitious unlearning for llms, 2024.

Yujia Qin, Cheng Qian, Jing Yi, Weize Chen, Yankai Lin, Xu Han, Zhiyuan Liu, Maosong Sun, and Jie Zhou. Exploring mode connectivity for pre-trained language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6726–6746, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.451. URL https://aclanthology.org/2022.emnlp-main.451/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

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. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7043–7058, 2026.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In Suresh Manandhar and Deniz Yuret (eds.), *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 341–350, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/S13-2056/.

Amrith Setlur, Benjamin Eysenbach, Virginia Smith, and Sergey Levine. Adversarial unlearning: Reducing confidence along adversarial directions. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=cJ006qBE8Uv.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TArmA033BU.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6418–6428, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1644. URL https://aclanthology.org/P19-1644/.

Enayat Ullah, Tung Mai, Anup Rao, Ryan A. Rossi, and Raman Arora. Machine unlearning via algorithmic stability. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4126–4142. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/ullah21a.html.

Jakub Vrabel, Ori Shem-Ur, Yaron Oz, and David Krueger. Input space mode connectivity in deep neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3qeOy7HwUT.

Ren Wang, Yuxuan Li, and Sijia Liu. Exploring diversified adversarial robustness in neural networks via robust mode connectivity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2346–2352, June 2023.

Shaokui Wei, Mingda Zhang, Hongyuan Zha, and Baoyuan Wu. Shared adversarial unlearning: Backdoor mitigation by unlearning shared adversarial examples. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=zqOcW3R9rd.

Yinjun Wu, Edgar Dobriban, and Susan Davidson. DeltaGrad: Rapid retraining of machine learning models. In *Proceedings of the International Conference on Machine Learning*, 2020.

Binchi Zhang, Yushun Dong, Tianhao Wang, and Jundong Li. Towards certified unlearning for deep neural networks. In *Forty-first International Conference on Machine Learning*, 2024a. URL https://openreview.net/forum?id=1mf1ISuyS3.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024b. URL https://openreview.net/forum?id=MXLBXjQkmb.

Yihao Zhang, Hangzhou He, Jingyu Zhu, Huanran Chen, Yifei Wang, and Zeming Wei. On the duality between sharpness-aware minimization and adversarial training. *arXiv preprint arXiv:2402.15152*, 2024c.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=QAbhLBF72K.

Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgwzCEKwH.

A Related work

Machine Unlearning Early unlearning methods span across efficiently retraining (Bourtoule et al., 2021b; Wu et al., 2020), model pruning (Jia et al., 2023), manipulating gradients (Ullah et al., 2021; Hoang et al., 2024), adversarial unlearning (Setlur et al., 2022; Wei et al., 2023), and data augmentation (Choi et al., 2024). Unlearning on LLMs recently draws more attention (Eldan & Russinovich, 2023; Ji et al., 2024; Kassem et al., 2023; Cheng & Amiri, 2025b). However, there is less attention on mechanistically understanding the loss landscape of machine unlearning methods.

Mode Connectivity Furthermore, several studies have shown that independently trained minimizers can be connected by low loss paths, a phenomenon known as mode connectivity (Draxler et al., 2018; Garipov et al., 2018; Frankle et al., 2020), across both vision and language models (Qin et al., 2022). During pruning, linear mode connectivity emerges only at early stage of training. This connectivity has been extended to multi-dimensional manifolds(Benton et al., 2021), and alternative topologies such as star-shaped and geodesic connectivity(Lin et al., 2024). Mode connectivity can also lead to more effective (Garipov et al., 2018) or adversarially robust (Zhao et al., 2020; Wang et al., 2023) models if ensembling along the curve. Vrabel et al. (2025) discover that mode connectivity can happen in input space. Existing works on mode connectivity focus on the learning process. There is no prior work that investigates mode connectivity in machine unlearning.

B Details of Unlearning Methods

Below, we present the details of the unlearning methods used in our study.

Gradient Ascent Gradient Ascent (GA) (Golatkar et al., 2020) performs gradient ascent on D_f without any mechanism to maintain utility on the retain set D_r.

Random Labeling Random Labeling (RL) (Golatkar et al., 2020) fine-tunes f_{θ_o} on D_f with corrupted labels and the original D_r (or a fraction of it if the entire D_r is too large). This method aims to inject errors to the forget set.

Saliency Unlearning SalUn (SU) (Fan et al., 2024c) first finds parameters that are salient to unlearning D_f. Next, it performs Random Labeling but only updates the salient parameters.

Bad Teaching Bad Teaching (BT) (Chundawat et al., 2023) forces the unlearned model to predict D_r similarly to the original model and to predict D_f similarly to an incompetent model (e.g. a randomly initialized model). It minimizes the KL-Divergence between prediction logits $\mathbb{KL}(f'(D_r)||f(D_r))$ on D_r and maximizes KL-Divergence between prediction logits $\mathbb{KL}(f'(D_f)||f_d(D_f))$ on D_f, where f_d is the incompetent model, e.g. a randomly initialized model.

Gradient Difference GradDiff (GD) (Maini et al., 2024) minimizes task loss on D_r and maximizes task loss on D_f.

Negative Preference Optimization NPO (Zhang et al., 2024b) is built upon the DPO (Rafailov et al., 2023) algorithm to post-train LLMs. In the original DPO, each query q corresponds to a winning response y_w to prioritize and a losing response y_l to suppress. NPO uses only the losing response with no winning response.

C Details of Evaluation Metrics

We provide detailed descriptions of the evaluation metrics used in our analysis. On MU-Bench (Cheng & Amiri, 2024) tasks, we follow the original paper to adopt accuracy as the

evaluation metric. In addition, we employ Zero-Retrain Forgetting score (↑) (Chundawat et al., 2023), which measures the similarity of prediction logits on D_f between the unlearned model and a random model.

TOFU (Maini et al., 2024) evaluates the unlearned model using p-value of Kolmogorov-Smirnov test for Model Utility (↑) and Forget Utility (↑), which measure the similarity of probability distributions between the unlearned and retrained model. Additionally, we also include verbatim evaluation using ROUGE-L recall score on Retain Authors (↑), Forget Authors (↓), Real Authors (↑), and World Knowledge (↑).

MUSE (Shi et al., 2025) evaluates the unlearned model using verbatim memorization on forget set (forget_verbmem ↓) and knowledge memorization on forget and retain set (forget_knowmem ↓, retain_knowmem ↑), by probing the unlearned model with a series of question related to forget set. All of these scores are measured by ROUGE-L.

On WMDP (Li et al., 2024) evaluates the unlearned model using accuracy on WMDP forget set. It also evaluates the general utility of the unlearned model using the MMLU benchmark (Hendrycks et al., 2021).

D Details of Curve Finding Process

To find the curve that connects θ_1 and θ_2, we can first compute the average loss along the curve:

$$\hat{\ell}(\theta) = \frac{\int L(\phi_\theta)\, d\phi_\theta}{\int d\phi_\theta}. \tag{5}$$

The numerator $\int L(\phi_\theta)\, d\phi_\theta$ is the line integral of the loss L along the curve ϕ_θ. It sums up the loss values at all points along the curve, weighted by the length of the curve in the parameter space. Intuitively, it measures the total accumulated loss along the curve, accounting for how long the curve is in regions with high or low loss.

The denominator $\int d\phi_\theta$ is the total length of the curve in the parameter space. It normalizes the numerator by the total length, ensuring that the result does not depend on the specific parameterization of the curve (e.g., stretching or shrinking segments artificially).

Minimizing the above loss ensures that the path between the two sets of weights corresponds to models with consistently high accuracy.

The integrals can be rewritten in terms of the parameter $t \in [0, 1]$ as

$$\hat{\ell}(\theta) = \int_0^1 L(\phi_\theta(t)) q_\theta(t)\, dt, \tag{6}$$

$$q_\theta(t) = \frac{\|\phi'_\theta(t)\|}{\int_0^1 \|\phi'_\theta(t)\|\, dt}. \tag{7}$$

$$\mathbb{E}_{t\sim[0,1]}\hat{\ell}(\theta) = \int_0^1 L(\phi_\theta(t)) q_\theta(t)\, dt. \tag{8}$$

E Additional Results

We present detailed results on TOFU in Figure 9–18 and on classification datasets in Figure 19–33.

E.1 MCU under independently unlearned minimizers

On TOFU, we find almost perfectly smooth curve with no degradation of unlearning quality on 3 out of 4 unlearning methods (GA, GD, and NPO). Along the curves, all model weights yield consistent unlearning quality, measured by a series of evaluation metrics, including forget quality, model utility, and ROUGE score. On the other hand when using method RL, the model weights along the curve is of consistently high quality in model utility but have slightly different forget quality. Specifically, in the middle part of the curve, we observe a drop of 0.1 point in forget quality and an increase of 0.05 point in forget ROUGE (↓). However, since forget quality is the p-value of KS test, any value greater than 0.05 is considered as good unleared model, see Figure 9 for details. As the size of forget set increases, indicated by different rows in Figure 9, there is trivial variation of forget quality and model utility along the linear and quadratic curve on GA, GD, and NPO. On RL, we notice interesting behaviors. When $|D_f| = 1\%$, forget quality degrades in the middle of the curve. When $|D_f| = 5\%$, forget quality does not change significantly. When $|D_f| = 10\%$, forget quality significantly increases in the middle of the curve. These behaviors are consistent on both linear and quadratic curves. We attribute these to the fact that RL is not an appropriate unlearning method for TOFU, which stuck in local optima and cannot ultimately converge to the low loss valley.

Therefore, we can find that the loss landscape of most unlearning methods on TOFU has essentially a flat low-loss valley where barriers, i.e. sudden performance degradation, rarely appear. This implies that, similar to learning (Draxler et al., 2018), minima of unlearning are perhaps best seen as points on a single connected manifold of low loss, rather than as the bottoms of distinct valleys for each individual unlearning method. The existence of mode connectivity paths suggests that modern neural networks have enough parameters such that they can achieve good predictions while a big part of the network undergoes structural changes. However, some unlearning methods may not converge to the low loss manifold, such as RL on TOFU dataset.

On classification dataset, we observe different patterns across different unlearning methods. On GA, it is generally easier to observe smooth MCU curve, both linear and quadratic, with small variation in forget set performance when $|D_f| = 1\%$. Due to the similarity in design, RL and SU have very similar MCU patterns. Both types of curve yield models with degraded forget set performance (↓) in the middle part of the curve (green line in Figure 19), with more prominent degradation on linear than quadratic curves. On BT, there is a strong linear MCU but the curve finding process fails to converge to meaningful quadratic MCU. This demonstrates that simpler connectivity may appear but hard to detect. We hypothesize that BT has a more rugged loss landscape than other methods, likely because it computes loss based on representations not directly on tasks loss. These results highlight the difference in loss landscape of unlearning methods.

CL and Non-CL On classification datasets, GA shows strong linear and quadratic MCU. RL and SU show quadratic (but not linear) connectivity, with slight degradation of forget set performance. This indicates that CL-based and Non-CL-based methods can converge to the same low loss manifold. On BT when $|D_f| = 4\%$, although there is almost no variation in forget set performance on linear curve, there is a major drop on retain set performance at the middle of the curve. Since MCU considers both forget set and retain set performance, this is not an emergence of MCU.

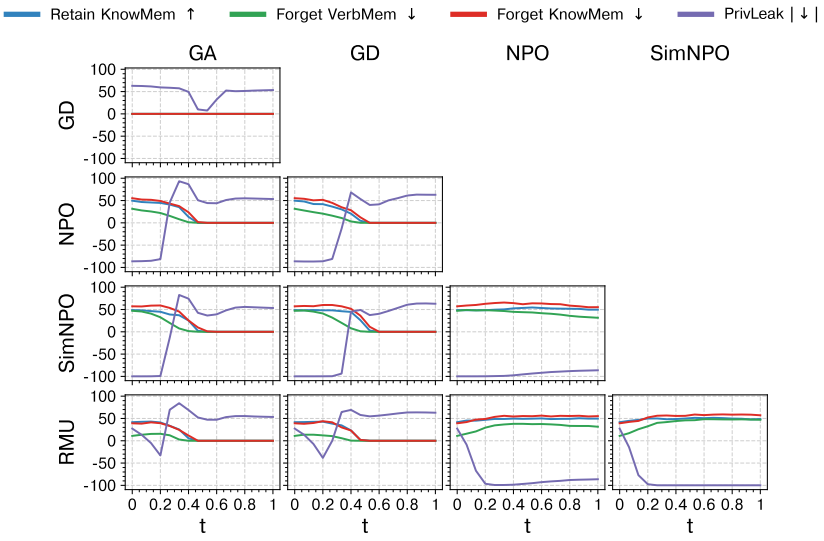


Figure 7: MCU under **Met** setting on **MUSE News dataset**. Methods on rows and columns correspond to θ_1' and θ_2' respectively. MCU is symmetric. Additional results are shown in Appendix E, Figures 14–18, Figures 25–33.

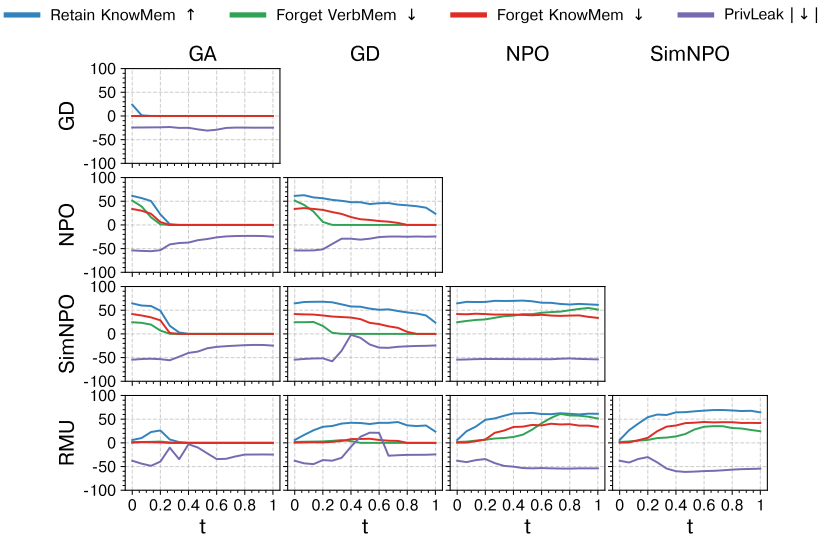


Figure 8: MCU under **Met** setting on **MUSE Books dataset**. Methods on rows and columns correspond to θ_1' and θ_2' respectively. MCU is symmetric. Additional results are shown in Appendix E, Figures 14–18, Figures 25–33.

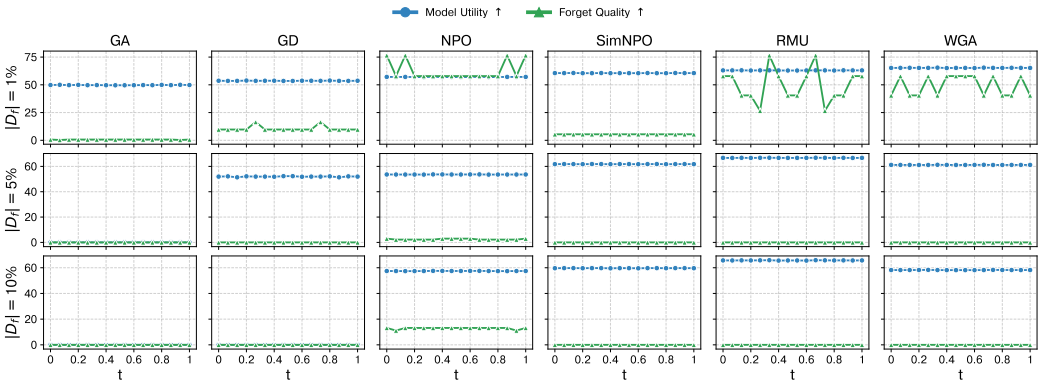


Figure 9: MCU under **Rand** setting on **TOFU dataset**.

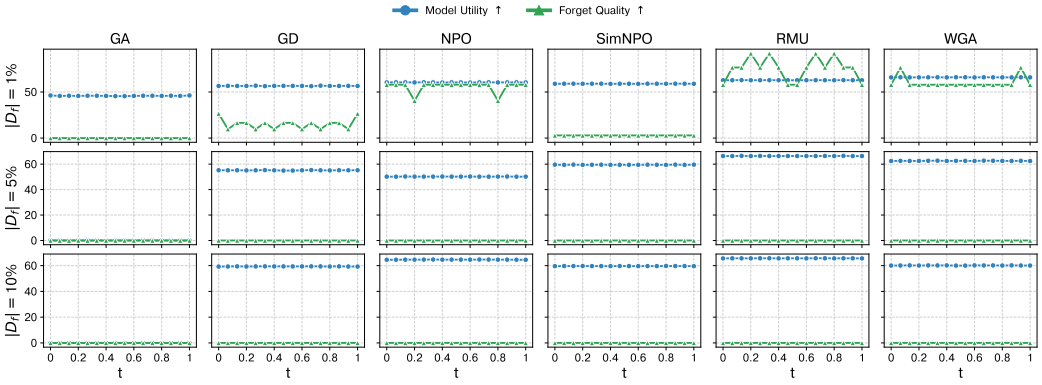


Figure 10: MCU under **Rand-CL** setting on **TOFU dataset**.

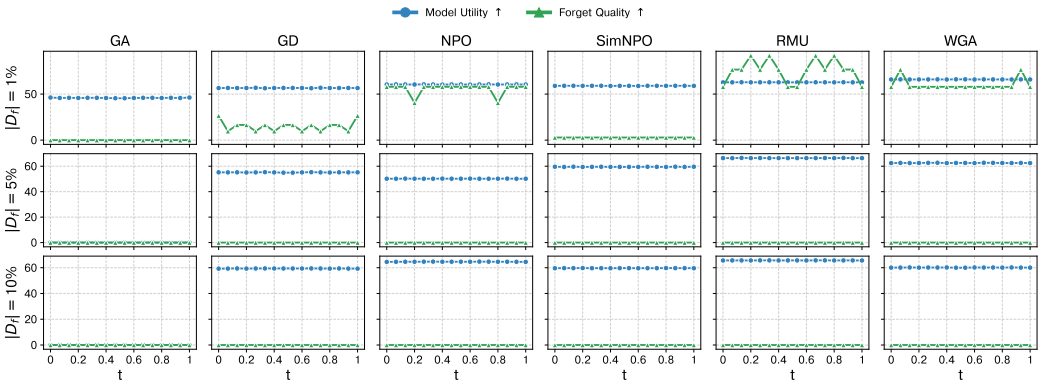


Figure 11: MCU under **Rand-SO** setting on **TOFU dataset**.

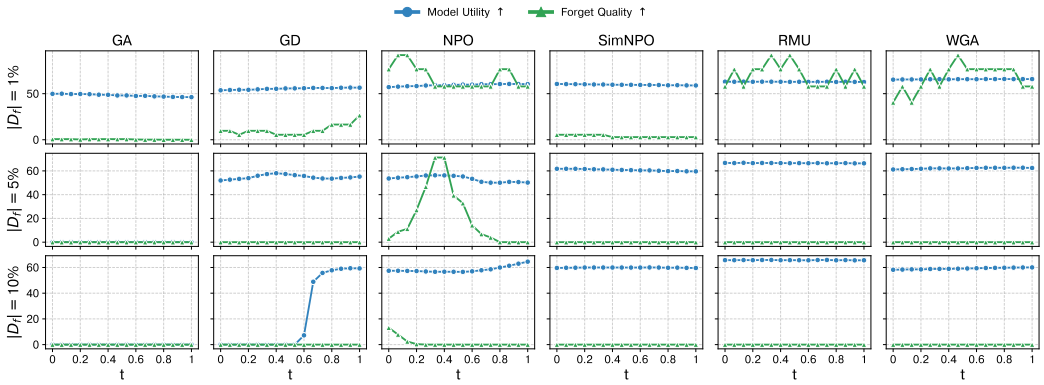


Figure 12: MCU under **CL-Non-CL** setting on **TOFU dataset**.

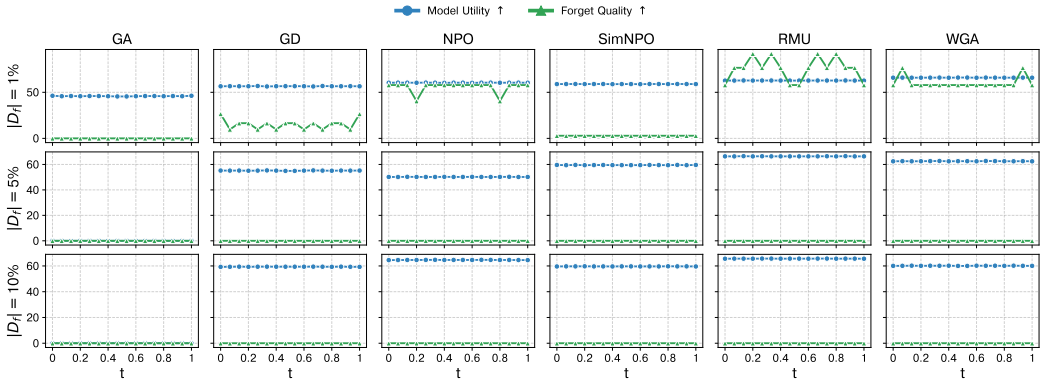
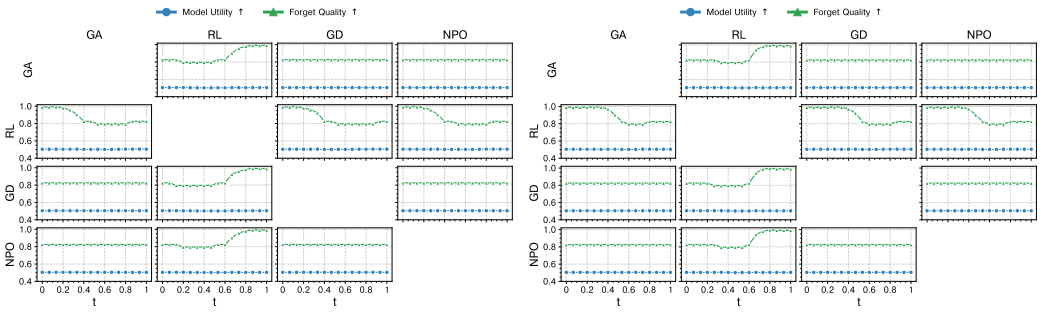
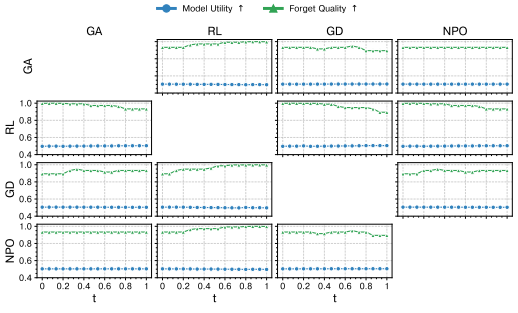


Figure 13: MCU under **FO-SO** setting on **TOFU dataset**.

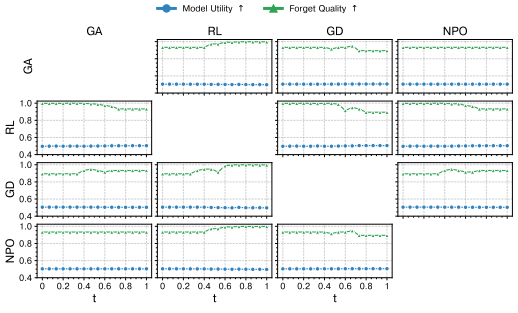


(a) Linear MCU when $|D_f| = 1\%$

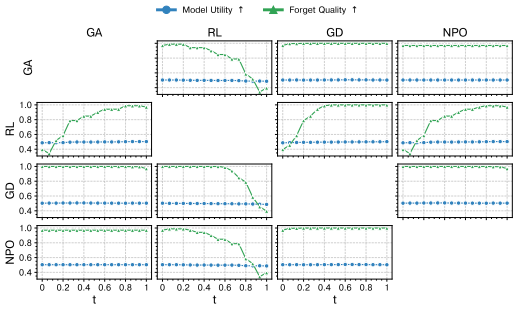
(b) Quadratic MCU when $|D_f| = 1\%$



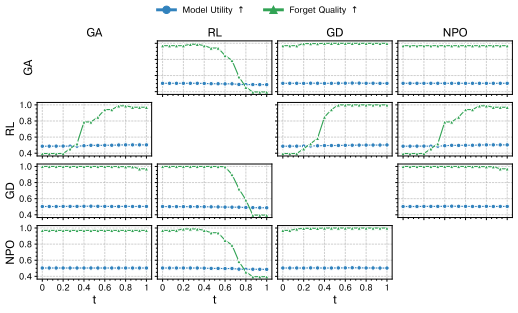
(c) Linear MCU when $|D_f| = 5\%$



(d) Quadratic MCU when $|D_f| = 5\%$

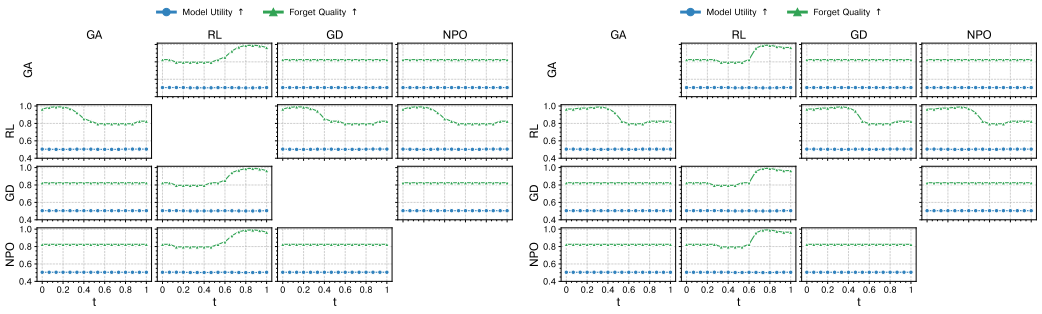


(e) Linear MCU when $|D_f| = 10\%$



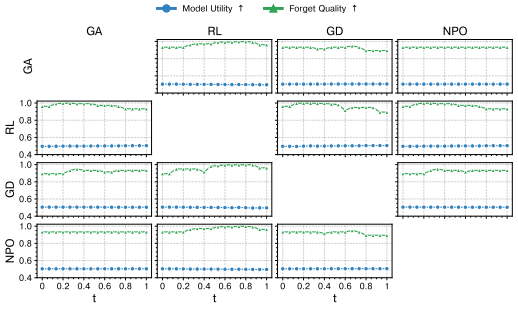
(f) Quadratic MCU when $|D_f| = 10\%$

Figure 14: MCU under **Met** setting on **TOFU dataset**.

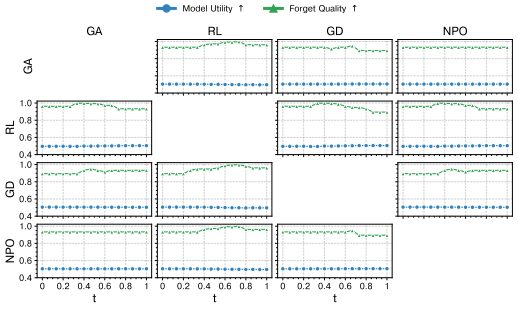


(a) Linear MCU when $|D_f| = 1\%$

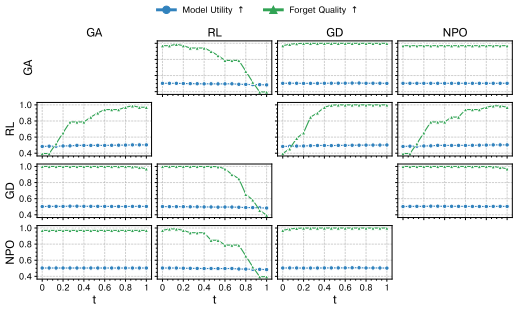
(b) Bezier MCU when $|D_f| = 1\%$



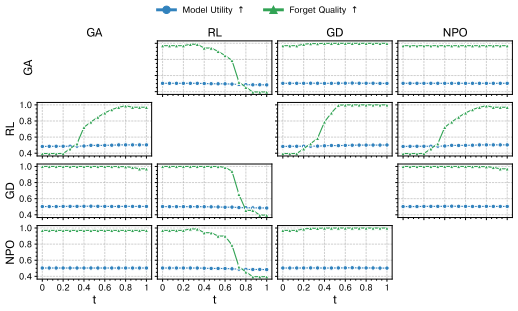
(c) Linear MCU when $|D_f| = 5\%$



(d) Bezier MCU when $|D_f| = 5\%$

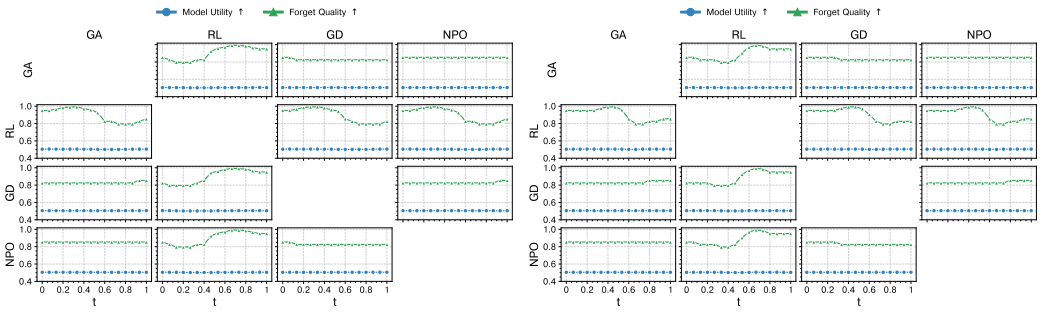


(e) Linear MCU when $|D_f| = 10\%$



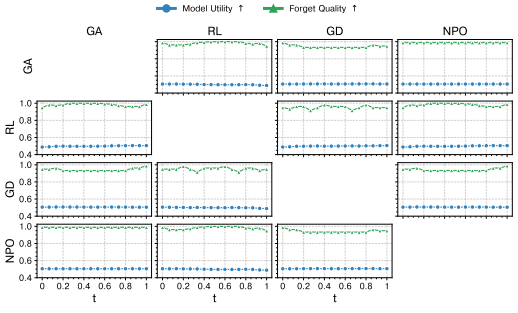
(f) Bezier MCU when $|D_f| = 10\%$

Figure 15: MCU under **Met-CL** setting on **TOFU dataset**.

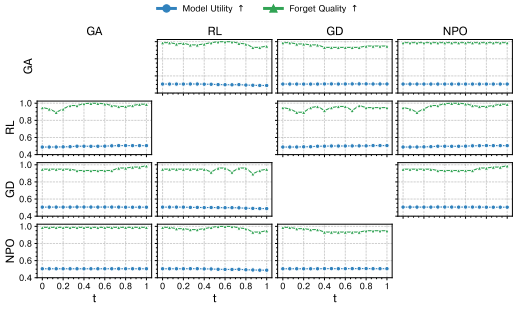


(a) Linear MCU when $|D_f| = 1\%$

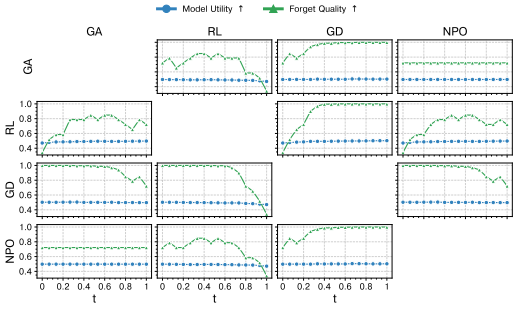
(b) Bezier MCU when $|D_f| = 1\%$



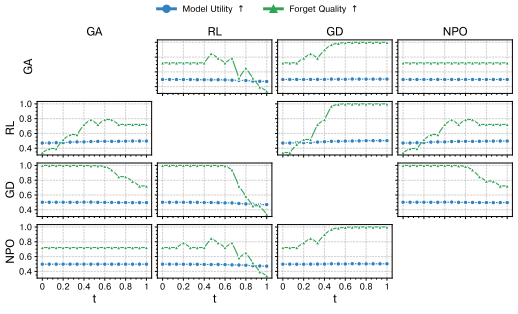
(c) Linear MCU when $|D_f| = 5\%$



(d) Bezier MCU when $|D_f| = 5\%$

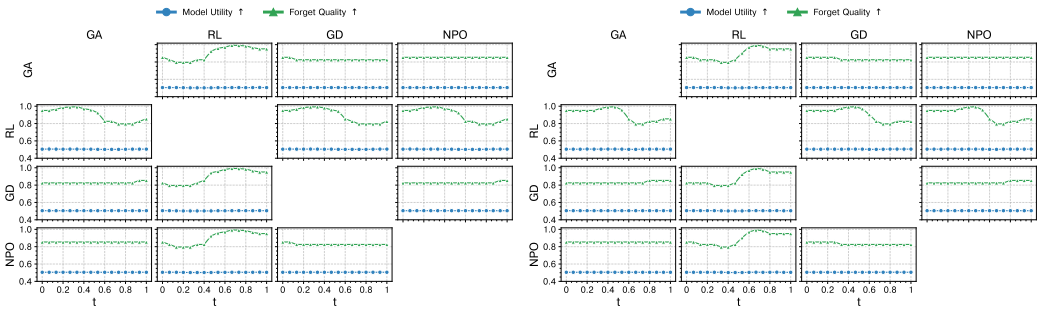


(e) Linear MCU when $|D_f| = 10\%$



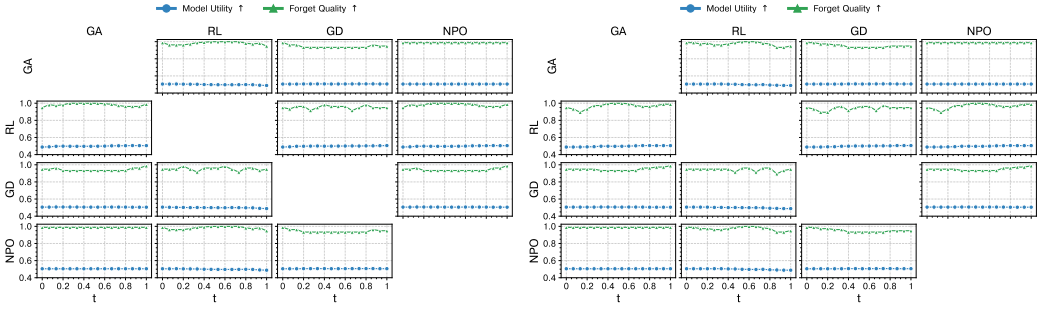
(f) Bezier MCU when $|D_f| = 10\%$

Figure 16: MCU under **Met-SO** setting on **TOFU dataset**.



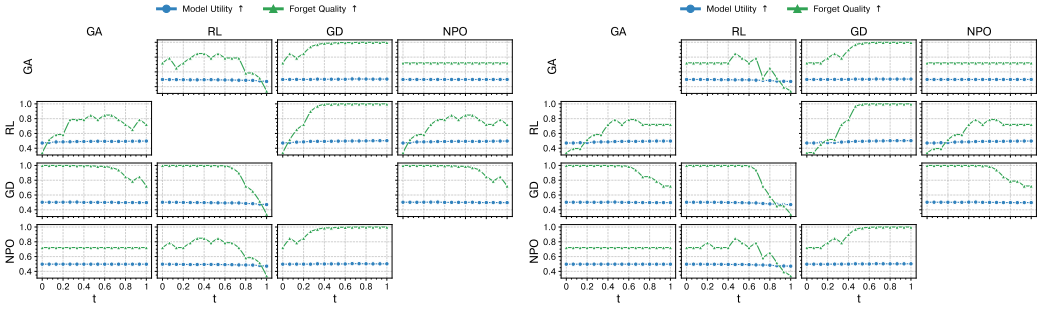
(a) Linear MCU when $|D_f| = 1\%$

(b) Bezier MCU when $|D_f| = 1\%$



(c) Linear MCU when $|D_f| = 5\%$

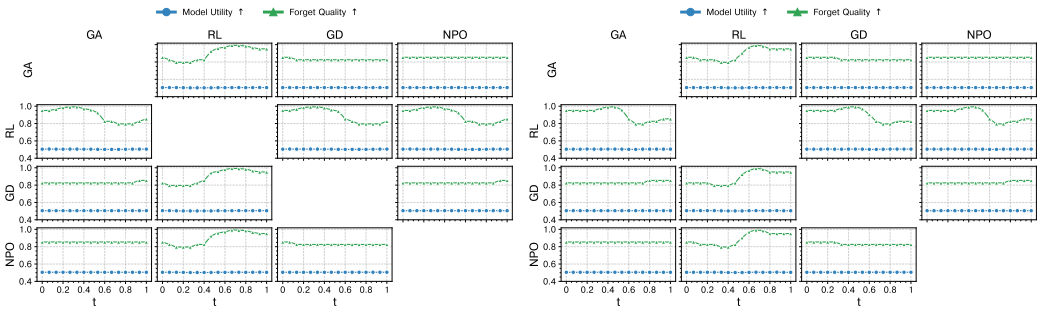
(d) Bezier MCU when $|D_f| = 5\%$



(e) Linear MCU when $|D_f| = 10\%$

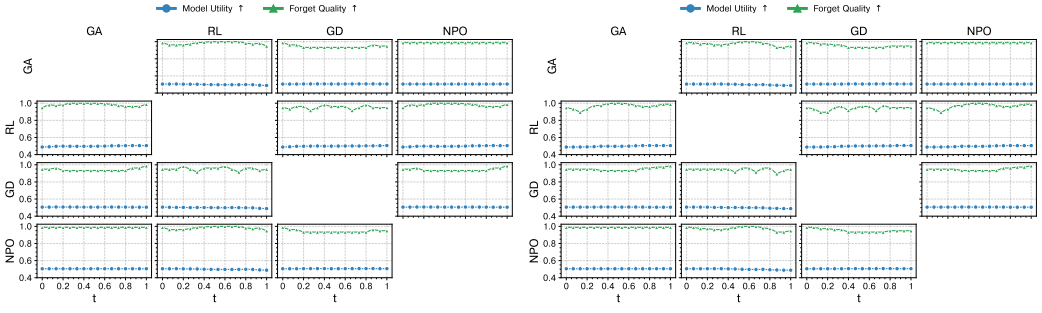
(f) Bezier MCU when $|D_f| = 10\%$

Figure 17: MCU under **Met-CL-Non-CL** setting on **TOFU dataset**.



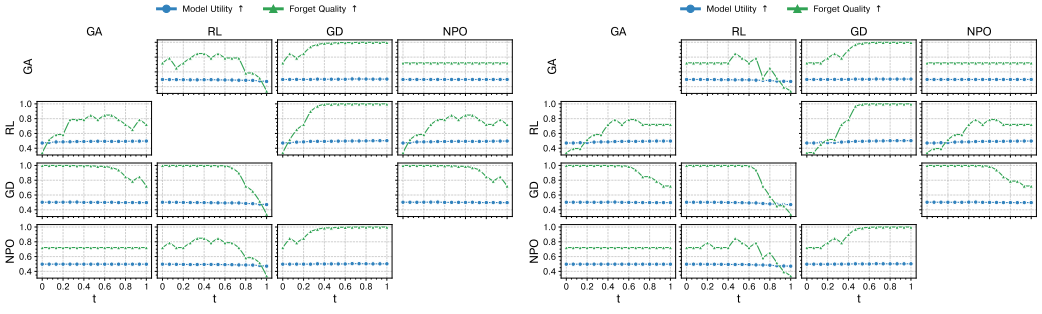
(a) Linear MCU when $|D_f| = 1\%$

(b) Bezier MCU when $|D_f| = 1\%$



(c) Linear MCU when $|D_f| = 5\%$

(d) Bezier MCU when $|D_f| = 5\%$



(e) Linear MCU when $|D_f| = 10\%$

(f) Bezier MCU when $|D_f| = 10\%$

Figure 18: MCU under **Met-FO-SO** setting on **TOFU dataset**.

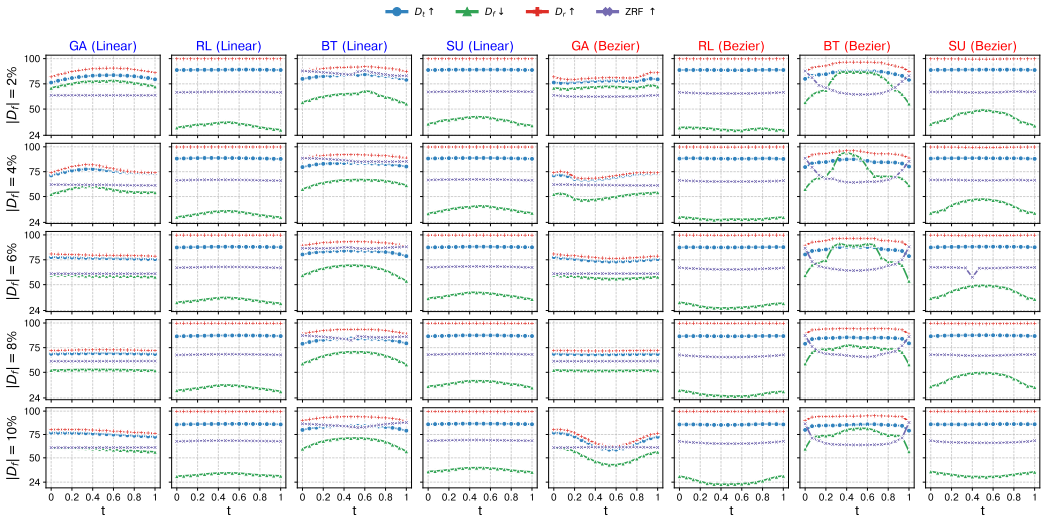


Figure 19: MCU under **Rand** setting on **classification dataset**.

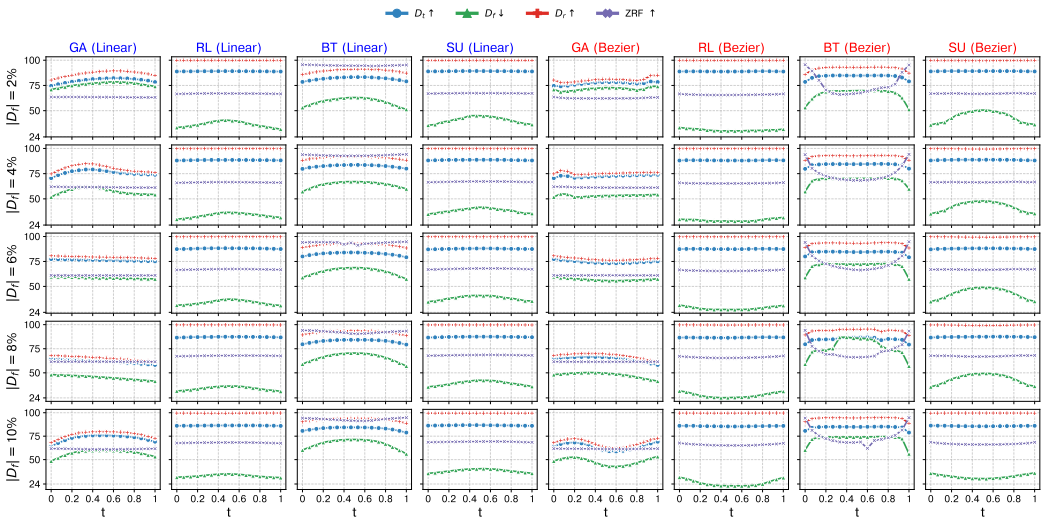


Figure 20: MCU under **Rand-CL** setting on **classification dataset**.

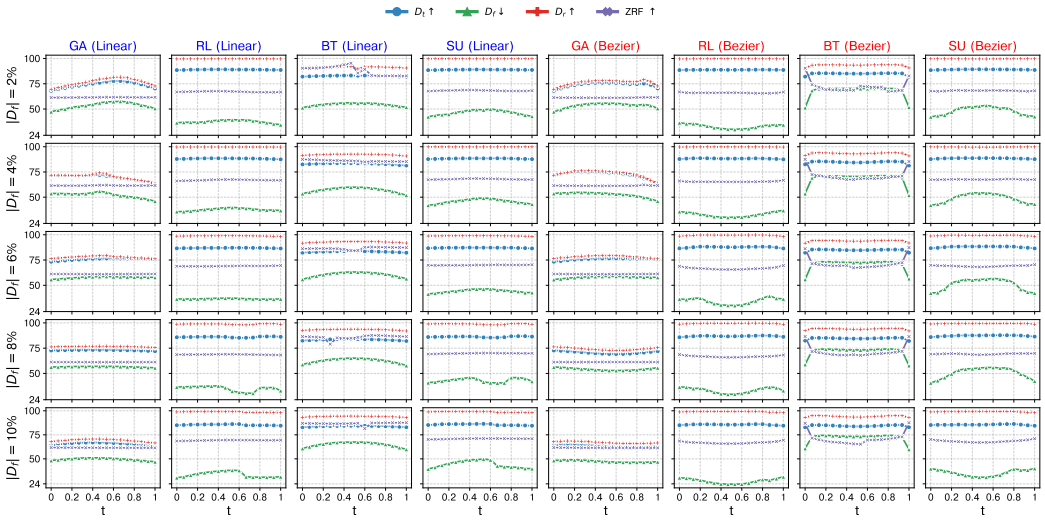


Figure 21: MCU under **Rand-SO** setting on **classification dataset**.

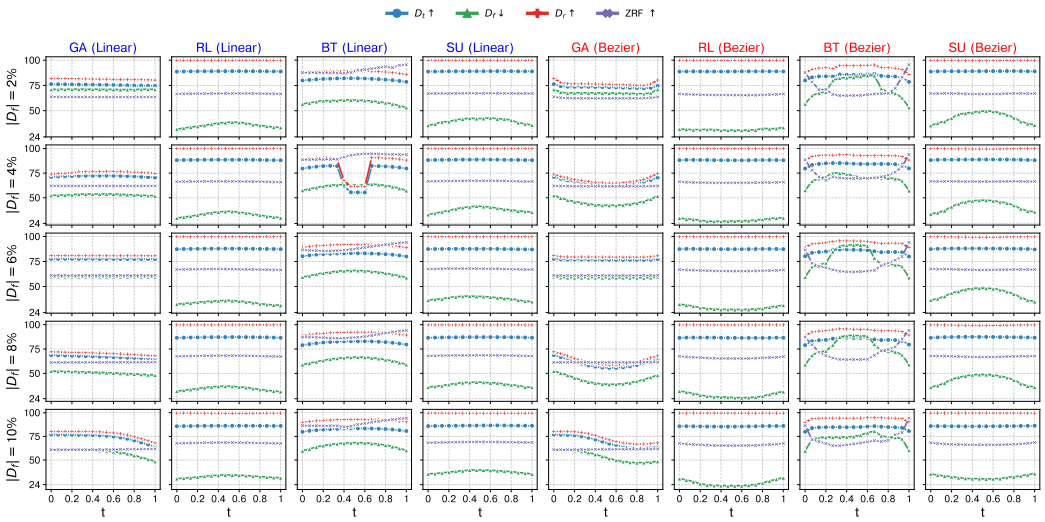


Figure 22: MCU under **CL-Non-CL** setting on **classification dataset**.

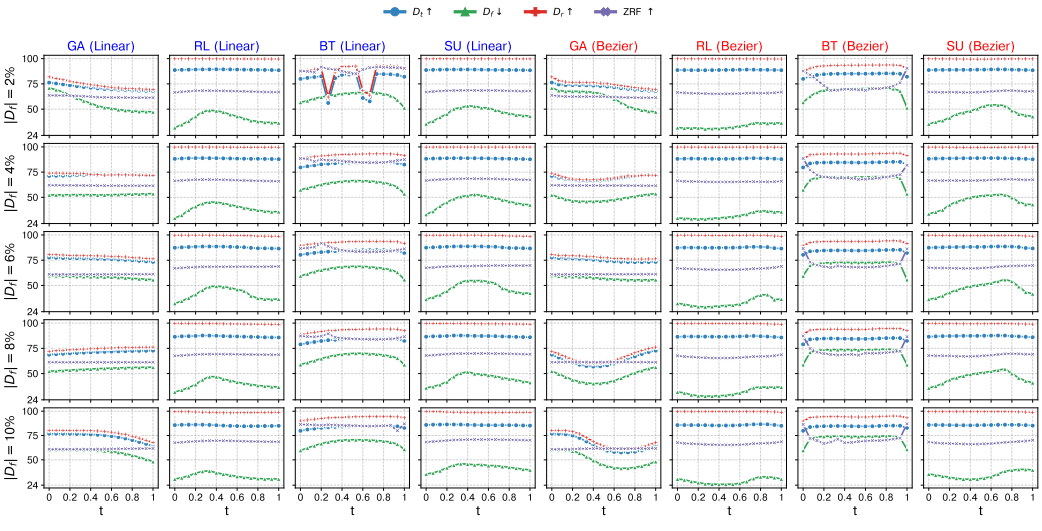
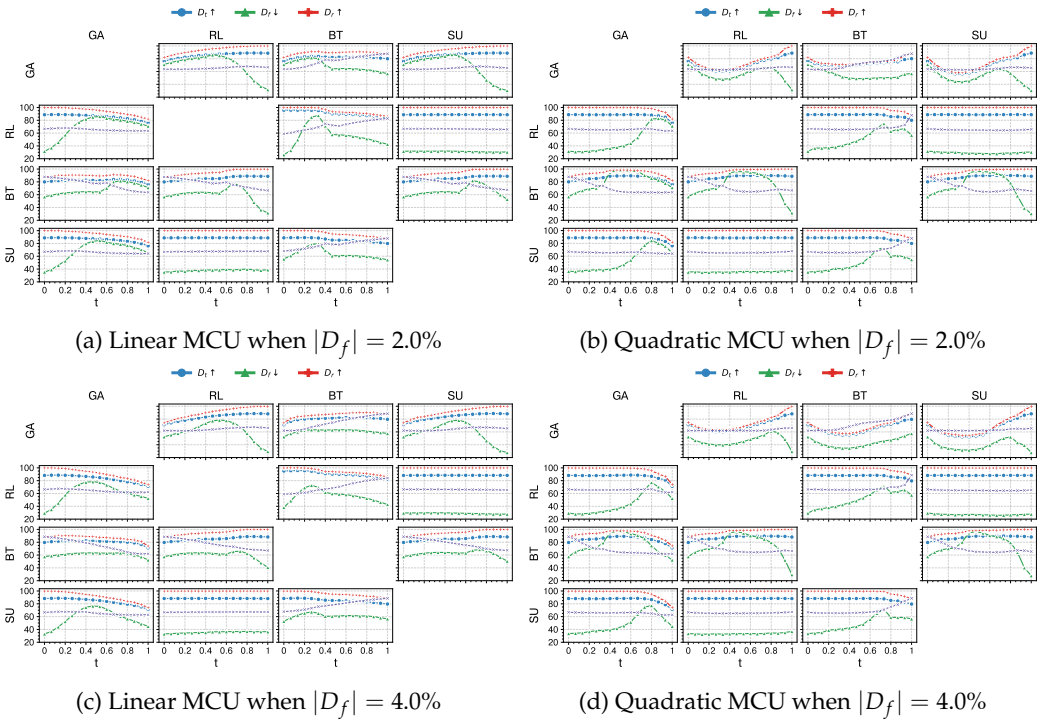


Figure 23: MCU under **FO-SO** setting on **classification dataset**.

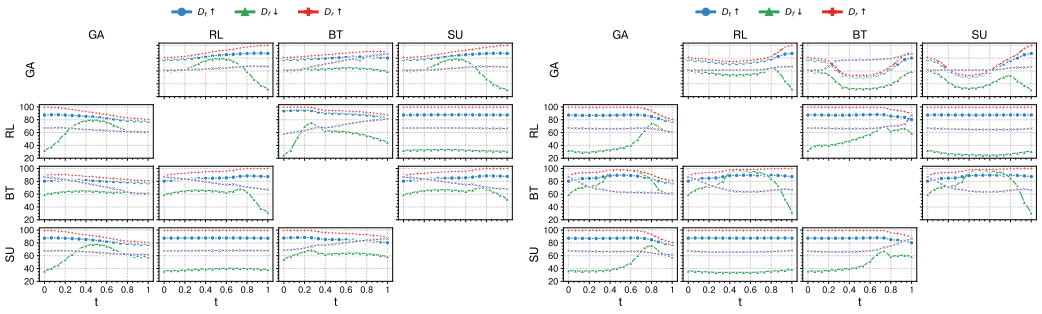


(a) Linear MCU when $|D_f| = 2.0\%$

(b) Quadratic MCU when $|D_f| = 2.0\%$

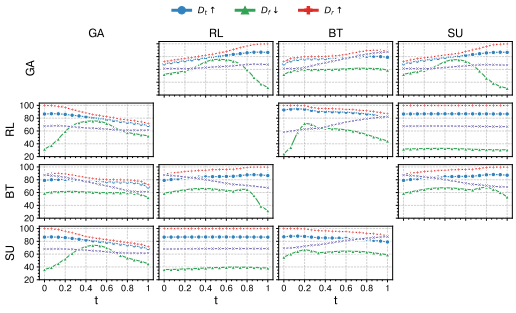
(c) Linear MCU when $|D_f| = 4.0\%$

(d) Quadratic MCU when $|D_f| = 4.0\%$

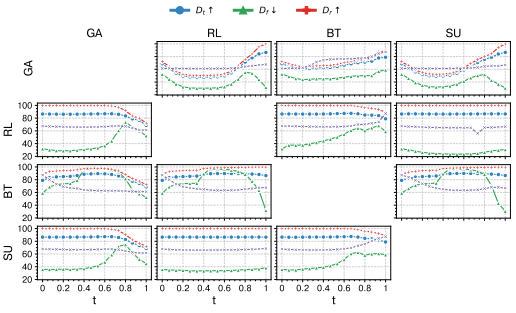


(a) Linear MCU when $|D_f| = 6.0\%$

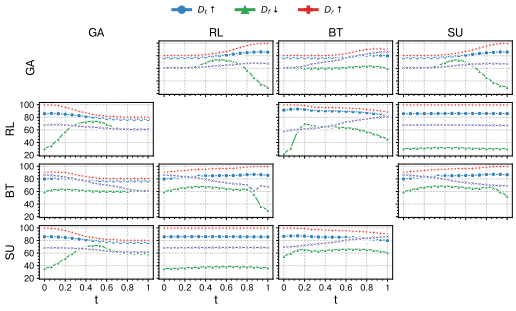
(b) Quadratic MCU when $|D_f| = 6.0\%$



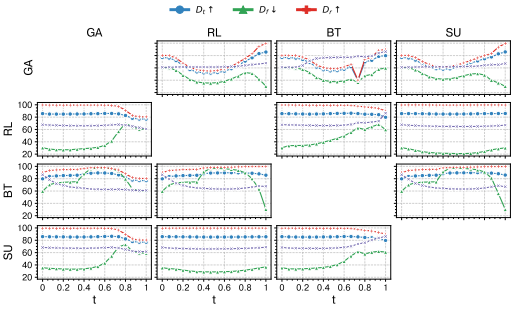
(c) Linear MCU when $|D_f| = 8.0\%$



(d) Quadratic MCU when $|D_f| = 8.0\%$

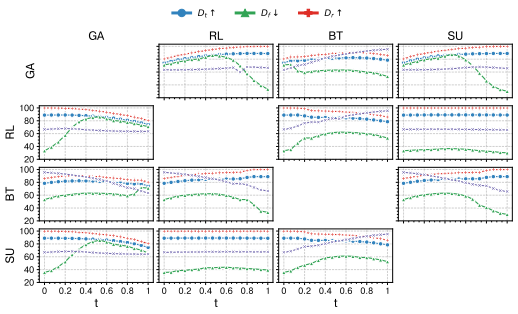


(e) Linear MCU when $|D_f| = 10.0\%$

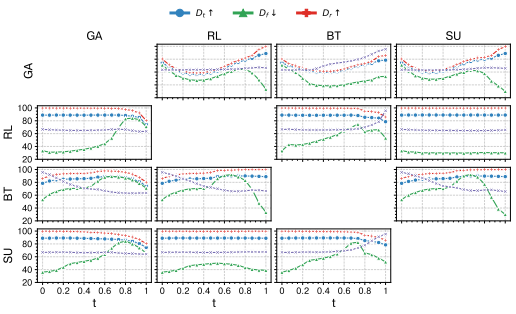


(f) Quadratic MCU when $|D_f| = 10.0\%$

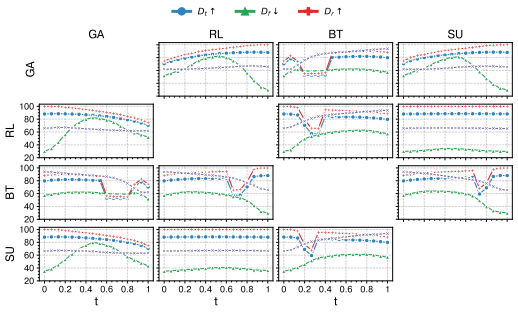
Figure 25: MCU under **Met** setting on **classification datasets**.



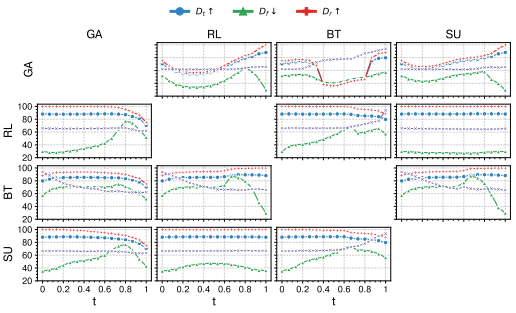
(a) Linear MCU when $|D_f| = 2.0\%$



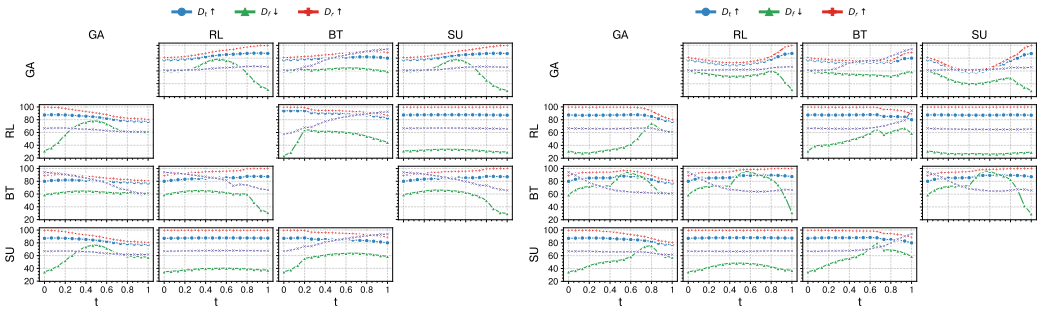
(b) Quadratic MCU when $|D_f| = 2.0\%$



(c) Linear MCU when $|D_f| = 4.0\%$

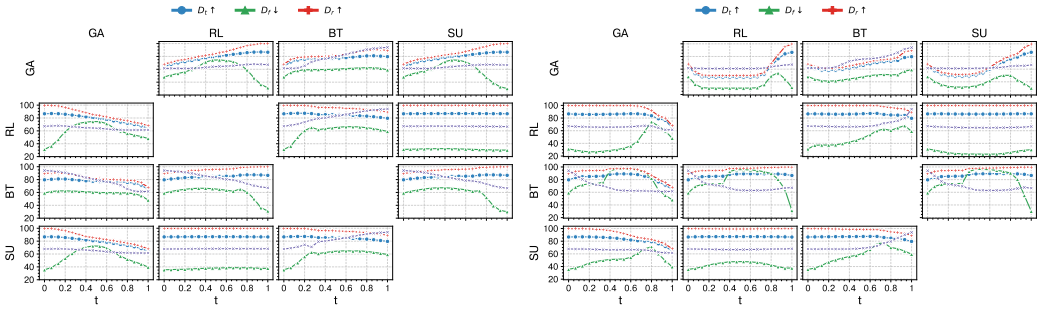


(d) Quadratic MCU when $|D_f| = 4.0\%$



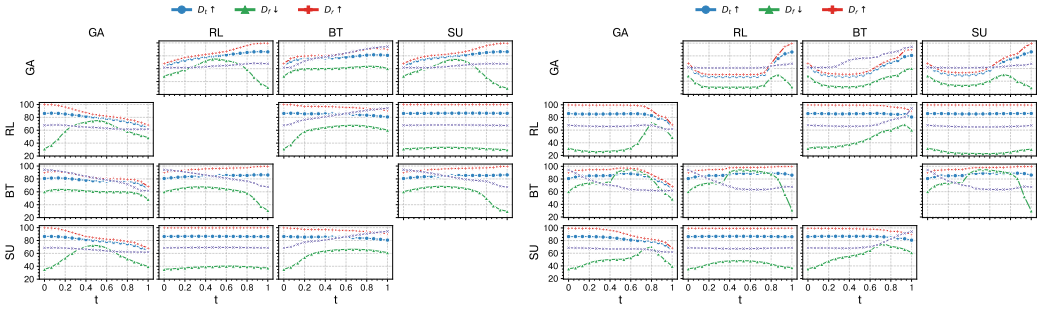
(a) Linear MCU when $|D_f| = 6.0\%$

(b) Quadratic MCU when $|D_f| = 6.0\%$



(c) Linear MCU when $|D_f| = 8.0\%$

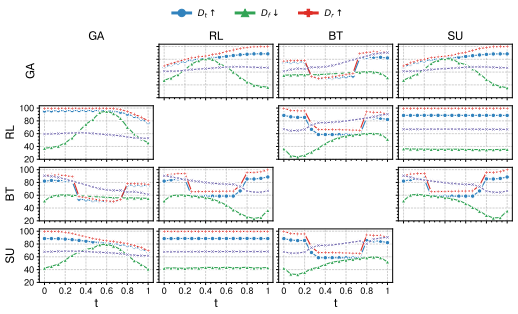
(d) Quadratic MCU when $|D_f| = 8.0\%$



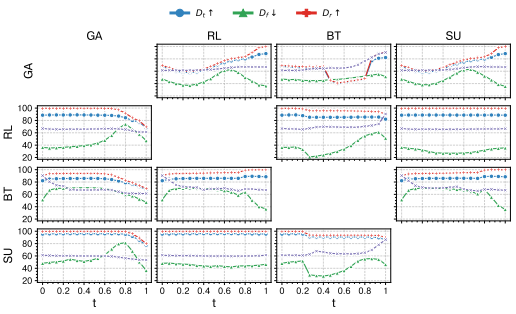
(e) Linear MCU when $|D_f| = 10.0\%$

(f) Quadratic MCU when $|D_f| = 10.0\%$

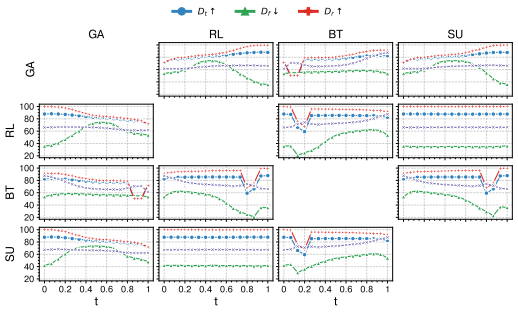
Figure 27: MCU under **Met-CL** setting on **classification datasets**.



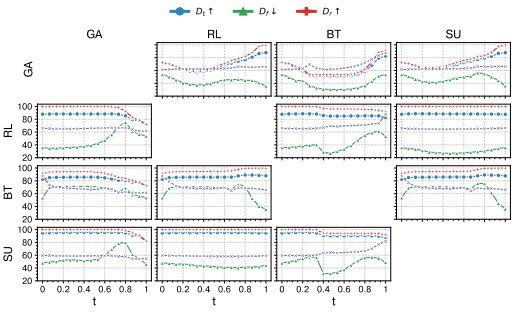
(a) Linear MCU when $|D_f| = 2.0\%$



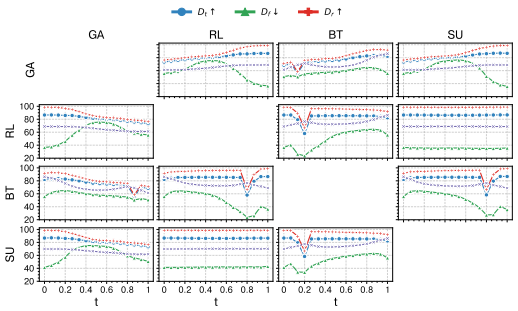
(b) Quadratic MCU when $|D_f| = 2.0\%$



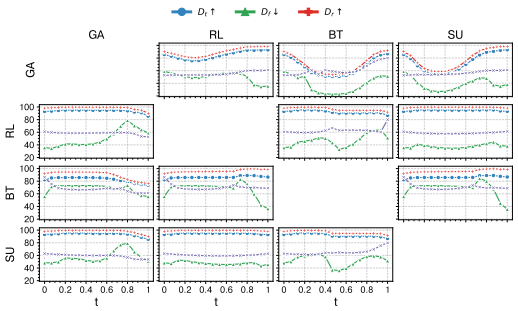
(c) Linear MCU when $|D_f| = 4.0\%$



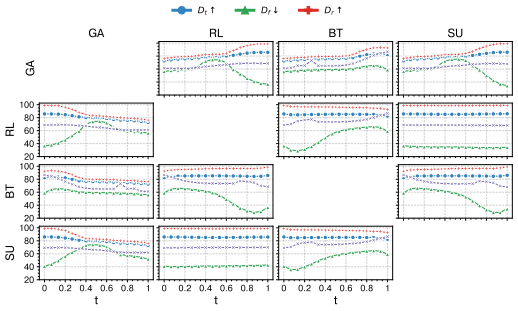
(d) Quadratic MCU when $|D_f| = 4.0\%$



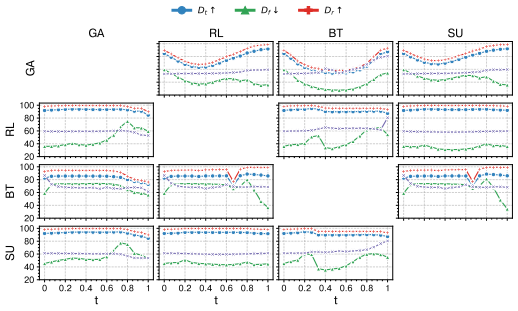
(a) Linear MCU when $|D_f| = 6.0\%$



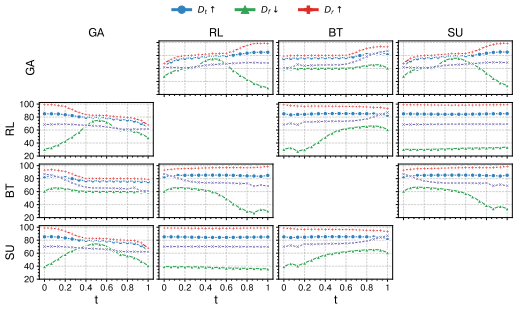
(b) Quadratic MCU when $|D_f| = 6.0\%$



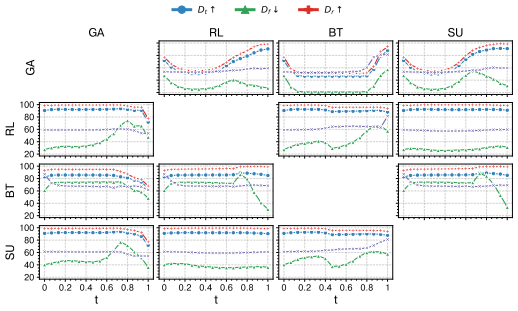
(c) Linear MCU when $|D_f| = 8.0\%$



(d) Quadratic MCU when $|D_f| = 8.0\%$

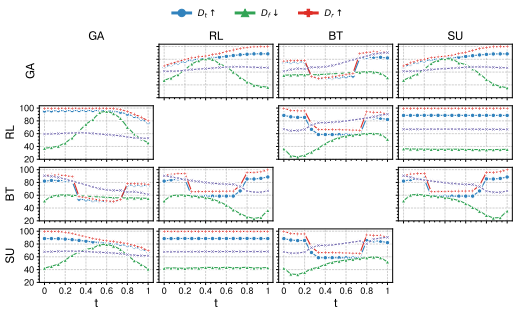


(e) Linear MCU when $|D_f| = 10.0\%$

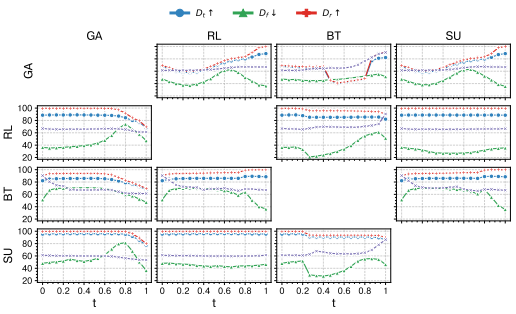


(f) Quadratic MCU when $|D_f| = 10.0\%$

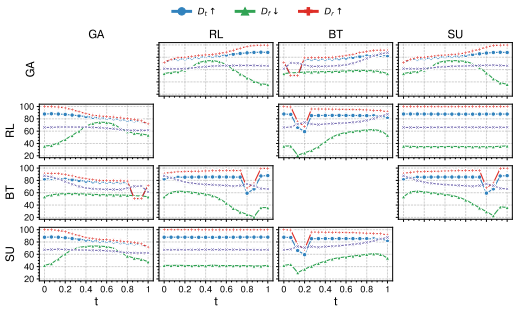
Figure 29: MCU under **Met-SO** setting on **classification datasets**.



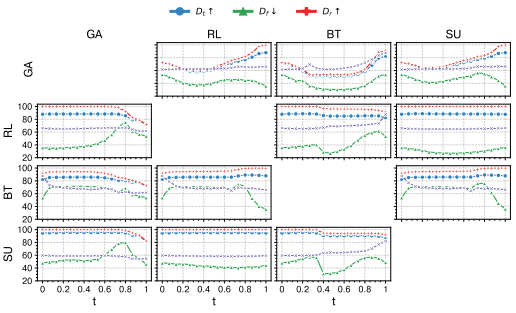
(a) Linear MCU when $|D_f| = 2.0\%$



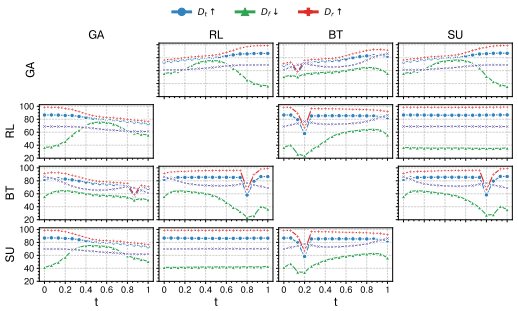
(b) Quadratic MCU when $|D_f| = 2.0\%$



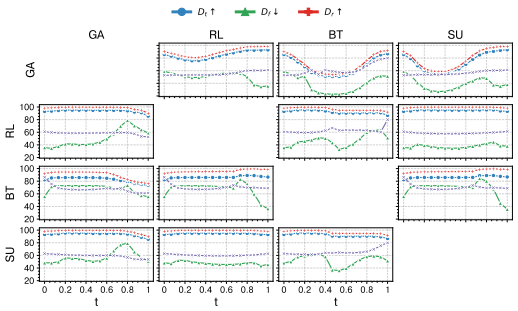
(c) Linear MCU when $|D_f| = 4.0\%$



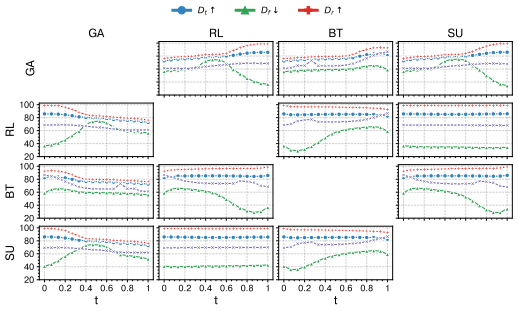
(d) Quadratic MCU when $|D_f| = 4.0\%$



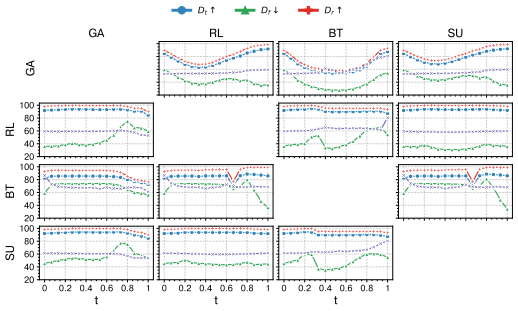
(a) Linear MCU when $|D_f| = 6.0\%$



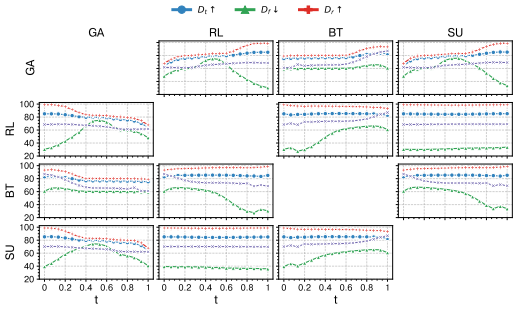
(b) Quadratic MCU when $|D_f| = 6.0\%$



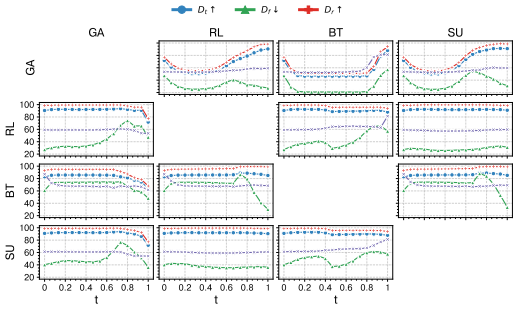
(c) Linear MCU when $|D_f| = 8.0\%$



(d) Quadratic MCU when $|D_f| = 8.0\%$

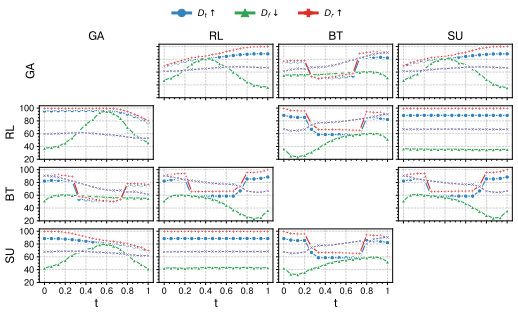


(e) Linear MCU when $|D_f| = 10.0\%$

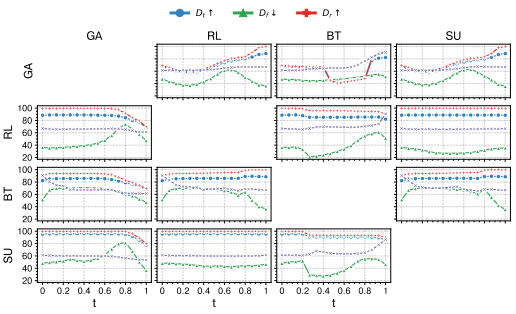


(f) Quadratic MCU when $|D_f| = 10.0\%$

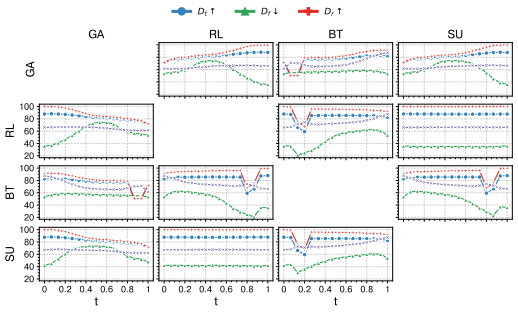
Figure 31: MCU under **Met-CL-Non-CL** setting on **classification datasets**.



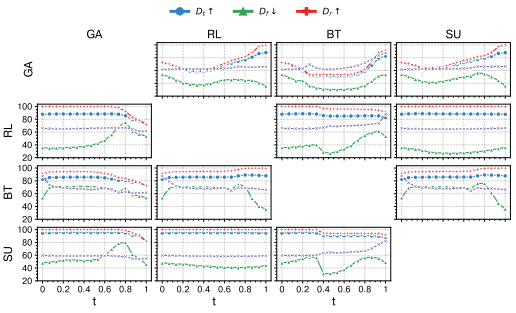
(a) Linear MCU when $|D_f| = 2.0\%$



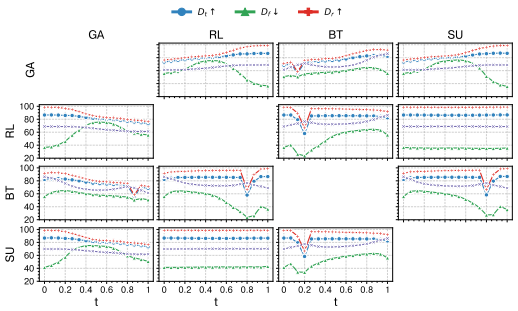
(b) Quadratic MCU when $|D_f| = 2.0\%$



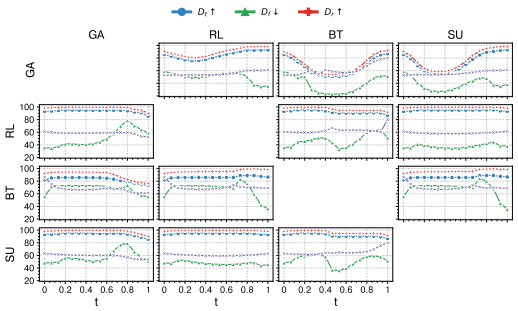
(c) Linear MCU when $|D_f| = 4.0\%$



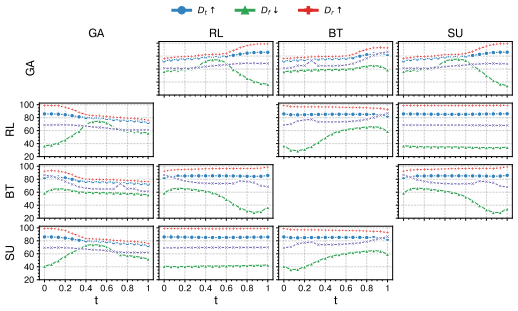
(d) Quadratic MCU when $|D_f| = 4.0\%$



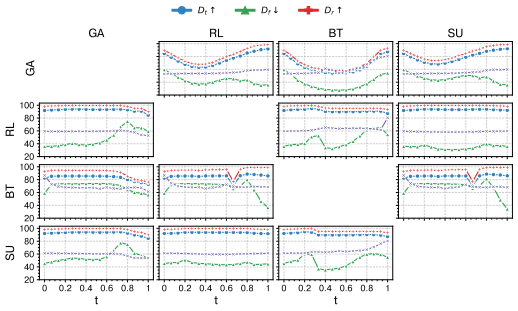
(a) Linear MCU when $|D_f| = 6.0\%$



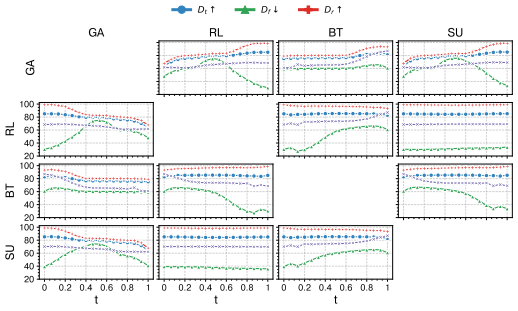
(b) Quadratic MCU when $|D_f| = 6.0\%$



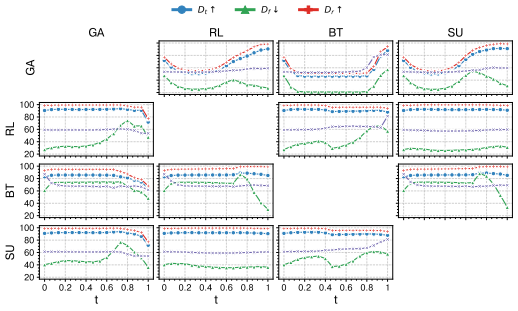
(c) Linear MCU when $|D_f| = 8.0\%$



(d) Quadratic MCU when $|D_f| = 8.0\%$



(e) Linear MCU when $|D_f| = 10.0\%$



(f) Quadratic MCU when $|D_f| = 10.0\%$

Figure 33: MCU under **Met-FO-SO** setting on **classification datasets**.