Understanding Machine Unlearning Through the Lens of Mode Connectivity

Jiali Cheng [1] **Hadi Amiri** [1]

Abstract

Machine Unlearning aims to remove undesired information from trained models without requiring full retraining from scratch. Despite recent advancements, their underlying loss landscapes and optimization dynamics received less attention. In this paper, we investigate and analyze machine unlearning through the lens of mode connectivity– the phenomenon where independently trained models can be connected by smooth low-loss paths in the parameter space. We define and study mode connectivity in unlearning across a range of overlooked conditions, including connections between different unlearning methods, models trained with and without curriculum learning, and models optimized with first-order and second-order techniques. Our findings show distinct patterns of fluctuation of different evaluation metrics along the curve, as well as the mechanistic (dis)similarity between unlearning methods. To the best of our knowledge, this is the first study on mode connectivity in the context of machine unlearning.

1. Introduction

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such as removing copyrighted data from LLMs (Eldan & Russinovich, 2023), as well as practical necessity of purging outdated or incorrect information (Dhingra et al., 2022).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

[1]University of Massachusetts Lowell, USA. Correspondence to: Jiali Cheng <jiali_cheng@uml.edu>, Hadi Amiri <hadi_amiri@uml.edu>.

Published at the ICML 2025 Workshop on Machine Unlearning for Generative AI.

Concurrently, the phenomenon of *mode connectivity* in deep learning 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), as illustrated in §2, 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 most mode connectivity has focused largely on vision tasks (Draxler et al., 2018; Vrabel et al., 2025), with straight-forward optimization objectives and static data distributions. Its relevance to unlearning– especially in the context of LLMs –remains unexplored. In addition, despite recent advances in LLM unlearning (Liu et al., 2024b;c; Hong et al., 2024b), our understanding of their underlying optimization dynamics is still limited. In particular, it is unclear whether mode connectivity holds during unlearning, and what this reveals about the loss landscape.

This paper introduces and formalizes the concept of **Mode Connectivity in Unlearning (MCU)**–a framework to analyze the structure of the loss landscape during unlearning and to assess whether different unlearning strategies converge to mechanistically similar solutions. Specifically, we investigate the following research questions: **RQ1**: under what training conditions (e.g., curriculum learning, second-order optimization) does mode connectivity emerge during unlearning? **RQ2**: can MCU reveal understandings of unlearning methods, such as mechanistic similarity or differences between different methods?

Answering these questions provides insight into the generalization, stability, and interpretability of unlearning methods. For instance, the existence of a smooth and low-loss path between two unlearning solutions may indicate shared inductive biases or similar optimization dynamics, which suggest that the unlearning methods reside in connected regions of the loss landscape. Conversely, a lack of connectivity may indicate divergent training behaviors and different solution structures.

Through extensive experiments on diverse tasks and different training paradigms, we find that the emergence of MCU is highly influenced by factors such as the choice of unlearning method, the complexity of unlearning task, and data marked for unlearning. We find that unlearned

models trained with fundamentally different optimization techniques can converge to the same low-loss manifold. In addition, although the same manifold can yield models with similar losses, their performances on different evaluation metrics can vary significantly. These findings provide insight into the loss landscape and unlearning dynamics. Our contributions are:

- **Mode Connectivity in Unlearning (MCU)**: We introduce and formalize MCU as a novel framework for studying machine unlearning through the lens of loss landscape (§3). To the best of our knowledge, this is the first study of mode connectivity in the unlearning setting.
- **Novel Experimental Conditions for Mode Connectivity**: We examine MCU under a range of new experimental conditions in mode connectivity, including curriculum learning, second-order optimization, and unlearning methods. These conditions have not previously been investigated in mode connectivity literature.

  These experiments provide new empirical insights into how optimization techniques influence unlearning (§7).
- **Insights into Unlearning Dynamics**: We reveal mechanistic similarities and divergences in the optimization dynamics of unlearning methods, particularly in LLMs. These insights provide new directions for the development of robust unlearning methods (§4.1–5.1).

2. Preliminaries

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

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

$$f' = \arg\min_{\theta'} L(D_r) - L(D_f). \tag{1}$$

Details of additional unlearning methods are discussed in Appendix 9.

Mode Connectivity Let θ_1 and θ_2 denote the weights of two independently trained models on some dataset 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 if the path $\phi(t)$ does not yield "barriers," defined as sudden increase in loss (Garipov et al., 2018; Lubana et al., 2023). Formally, $\forall t \in [0,1]$:

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

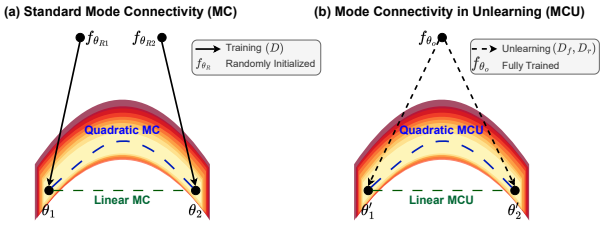


Figure 1: **(a)**: Illustration of standard mode connectivity (MC): MC finds a smooth curve connecting two minimizers that yields consistent low loss on D. **(b)**: Illustration of mode connectivity in unlearning (MCU): unlearning removes knowledge of forget set D_f from the trained model f_{θ_o} while maintaining knowledge of retain set $D_r = D \setminus D_f$. MCU finds a smooth curve connecting the two unlearned models θ_1' and θ_2' that yields consistent low loss on D_r and high loss on D_f. See § 3.

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. $\phi(t) = t\theta_1 + (1-t)\theta_2$. Stronger linear connectivity indicates stronger mechanistic similarity of minimizers, such as their inductive biases (Lubana et al., 2023).
- **Bézier:** 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 11.

3. Mode Connectivity in Unlearning (MCU)

Definition 3.1 (Mode Connectivity in Unlearning (MCU))**.** 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, $\forall t \in [0, 1]$

$$L\big(D_r; \phi(t)\big) \leq t \cdot L(D_r; \theta_1') + (1-t) \cdot L(D_r; \theta_2'), \quad (3)$$

$$L\big(D_f; \phi(t)\big) \geq t \cdot L(D_f; \theta_1') + (1-t) \cdot L(D_f; \theta_2'). \quad (4)$$

In MCU, "barriers" on the retain set D_r refer to the sudden *increase* of task loss L (as in standard mode connectivity), while "barriers" on the unlearn set D_f refer to the sudden *decrease* of task loss. Eq. 3 ensures that the task loss on the retain set 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 D_f along the mode connectivity path. In other words, MCU is realized when there exists a continues path of model weights connecting the minimizers θ_1' and θ_2', such that performance remains high on D_r and low on D_f along the curve.

Connection to Standard Mode Connectivity In contrast to standard mode connectivity, MCU must satisfy objectives on both D_f and D_r. Essentially, MCU examines whether it is possible to find a continues curve between two minimizers 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 minimizers θ_1 and θ_2 are obtained by training the model from two random initializations. In contrast, the minimizers θ_1' and θ_2' in MCU are both derived from the exact same trained model f_{θ_o}.

We present Influence of Training Dynamics on Mode Connectivity in Unlearning in Appendix §7.

4. Results

Details of experimental setup is discussed in Appendix §8.

4.1. RQ1: Emergence of MCU Across Diverse Conditions

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 not only possible but often prevalent–though, its emergence is influenced by unlearning method, training dynamics and the size of forget set D_f.

MCU between independently unlearned models On TOFU, we observe almost perfectly smooth curve with no degradation of unlearning quality for three out of four unlearning methods (GA, GD, and NPO). These curves show consistently high forget quality, model utility, and ROUGE score, which suggest that unlearning solutions reside on a connected low-loss manifold. 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. For RL, we observe significant fluctuations in forget quality, particularly in the middle part. However, forget quality remains above the accepted $\rho > 0.05$ in KS test, indicating connectivity still holds but less smoothly, see Figure 2. These results suggest that while some unlearning methods such as RL may converge to less optimal solutions, other induce flatter loss regions. 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.

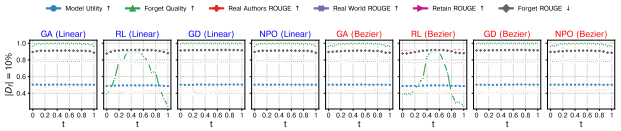


Figure 2: MCU under **Rand** setting on **TOFU dataset**. Additional results are shown in Appendix 12 Figure 7–11 for TOFU and Figure 47–51 for classification tasks.

On classification dataset, results vary. GA results in the smoothest MCU curves, both linear and quadratic, particularly for small $|D_f| = 1\%$. 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 47), 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 suggests that simpler connectivity may appear but hard to detect. We hypothesize that BT has a more rugged loss landscape than other methods, possibly because of its indirect loss formulation based on representations rather than explicit tasks loss. These results highlight the difference in loss landscape of unlearning methods. More details are discussed in Appendix 12.1.

Curriculum Learning (CL) Has Varying Effects on MCU When both endpoints (minimizers) are unlearned with CL (Rand-CL), we observe different contributions from CL across different datasets. On TOFU, GA, GD, or NPO result in connectivity patterns that are equally as performant as Non-CL minimizers, see Figures 8 and 7). This is while RL yields minimizers with slightly lower forget quality but smoother mode connectivity curves, for both

linear and quadratic. This suggests that CL-based unlearning may converge to different regions than non-CL-based unlearning, resulting in comparable and sometimes-better performance. On classification tasks, finding MCU in CL space (Figure 48) is much easier than non-CL space (Figure 47) for BT. However, CL has trivial contribution on RL and SU. These results imply that CL can guide optimization toward flatter regions of the loss landscape, depending on the model and method.

Second-Order Optimization (SO) Has Varying Effects on MCU Similar to CL, when both minimizers are unlearned with SO, we observe different effects incurred by SO on generative and discriminative tasks, and on different size of forget set $|D_f|$. On TOFU, So-based unlearning results in more pronounced barriers (Rand-SO in Figure 9) compared to standard unlearning (Rand in Figure 7) and CL-based unlearning (Rand-CL in Figure 7). These results can be attributed to SO optimization, which takes larger steps in parameter space than FO optimization (Liu et al., 2024a). This may sometimes lead to different low-loss manifolds. On classification datasets, SO generally leads to a smoother manifold for all methods, where linear and quadratic connectivity are easier to emerge for all methods, even when $|D_f|$ is large. Still, SO is insufficient for RL and SU when unlearning large forget sets, e.g. $|D_f| \geq 8\%$.

MCU between CL and Non-CL MCU may emerge between models trained with and without CL. On TOFU, GD show smooth connectivity for both small and large D_f (1% or 10% of data). However, when $|D_f| = 5\%$, we observe sudden decrease of loss (a "cliff") in forget quality. Similarly, with other unlearning methods, we observe such cliffs connecting two separate manifolds with small size of forget set. When $|D_f| = 10\%$, the MCU curve undergoes several barriers and cliffs. 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. This suggests that CL-based and non-CL-based unlearning are likely to converge to roughly the same low loss manifolds, yielding models of comparable unlearning efficacy. On TOFU, the manifold is smoother. While on classification tasks, there is slightly more small barriers for the loss of D_f.

MCU between FO and SO MCU across FO and SO minimizers show more diverse patterns. On TOFU, Figure 11, GD but not other unlearning methods show smooth MCU. Similarly, on classification datasets (Figure 51), smooth MCU hardly emerges. These results suggest that FO and SO optimization can drive the same unlearning method converge to different low-loss manifold. These minima are not connected by smooth pathways, demonstrated by failure of quadratic MCU. There is no consistent results that FO or SO is better.

Connectivity differs on D_r and D_f On retain set D_r, smooth connectivity are more likely to occur and easier to find. On the other hand, the loss landscape over the forget set can be highly irregular and prone to fluctuations. This could be because D_r is typically much larger than D_f, which leads to a more stable optimization signal and a smoother curve in the retain region of the loss landscape.

Smooth manifold can occur between low quality unlearned models In most cases, we observe that smooth manifold appear when both minimizers achieve low loss, which is consistent with early works on mode connectivity (Draxler et al., 2018; Garipov et al., 2018). However, in unlearning, worse performing minimizers do not necessarily mean smoother manifold in loss landscape. One example is the comparison between Rand (Figure 7) and Rand-CL (Figure 8) when unlearning with RL on TOFU dataset. While on RL, CL yield minimizers with slightly lower forget quality when $|D_f = 1\%|$, $|D_f = 5\%|$, and equally low forget quality when $|D_f = 10\%|$. However, worse performing minimizers do not necessarily mean smoother manifold in loss landscape. Specifically, we find that RL yields a mode connectivity curve with less fluctuations, i.e. less variation in forget quality, on both linear and quadratic curve.

Effect of forget set size We examine the impact of increasing the size of the forget set D_f on MCU behavior, Figure 7. For GA, GD, and NPO, both forget quality and model utility remain stable along the linear and quadratic curves. In contrast, RL show more complex behaviors. At $|D_f| = 1\%$, forget quality degrades near the curve center. At $|D_f| = 5\%$, performance remains relatively stable. At $|D_f| = 10\%$, forget quality substantially increases in the middle of the curve. These patterns are consistent on both linear and quadratic curves. The results suggest that RL is prone to converging to poor local minima in TOFU.

5. Conclusion

We introduce and formalize mode connectivity in unlearning (MCU) as a framework for understanding the loss landscape and optimization dynamics of machine unlearning. Our experiments across diverse tasks, unlearning methods, and training configurations show that MCU provides insights into method similarity, training stability, and unlearning difficulty.

Ethics Statement

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.

Reproducibility Statement

Our experiments are based on public benchmarks and implementations from TOFU (Maini et al., 2024) and MU-Bench (Cheng & Amiri, 2024). We will release our code and data splits to the research community.

References

Barbulescu, G.-O. and Triantafillou, P. 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.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. 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.

Benton, G., Maddox, W., Lotfi, S., and Wilson, A. G. G. Loss surface simplexes for mode connecting volumes and fast ensembling. In Meila, M. and Zhang, T. (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.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159, 2021a. doi: 10.1109/SP40001.2021.00019.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *IEEE Symposium on Security and Privacy (SP)*, 2021b.

Chen, Z., Cheng, J., Tolomei, G., Liu, S., Amiri, H., Wang, Y., Nag, K., and Lin, L. Frog: Fair removal on graphs. *arXiv preprint arXiv:2503.18197*, 2025.

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

Cheng, J. and Amiri, H. Multidelete for multimodal machine unlearning. In Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., and Varol, G. (eds.), *Computer Vision – ECCV 2024*, pp. 165–184, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-72940-9.

Cheng, J. and Amiri, H. Linguistic blind spots of large language models. *arXiv preprint arXiv:2503.19260*, 2025b.

Cheng, J. and Amiri, H. Tool unlearning for tool-augmented llms. *arXiv preprint arXiv:2502.01083*, 2025c.

Cheng, J., Dasoulas, G., He, H., Agarwal, C., and Zitnik, M. 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.

Choi, D., Choi, S., Lee, E., Seo, J., and Na, D. 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.

Chundawat, V. S., Tarun, A. K., Mandal, M., and Kankanhalli, M. 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.

Dhingra, B., Cole, J. R., Eisenschlos, J. M., Gillick, D., Eisenstein, J., and Cohen, W. W. 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/.

Draxler, F., Veschgini, K., Salmhofer, M., and Hamprecht, F. Essentially no barriers in neural network energy landscape. In Dy, J. and Krause, A. (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.

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

Fan, C., Liu, J., Hero, A., and Liu, S. Challenging forgets: Unveiling the worst-case forget sets in machine unlearning, 2024a.

Fan, C., Liu, J., Lin, L., Jia, J., Zhang, R., Mei, S., and Liu, S. Simplicity prevails: Rethinking negative preference optimization for llm unlearning. *arXiv preprint arXiv:2410.07163*, 2024b.

Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D., and Liu, S. 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.

Frankle, J., Dziugaite, G. K., Roy, D., and Carbin, M. Linear mode connectivity and the lottery ticket hypothesis. In III, H. D. and Singh, A. (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.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (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.

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

Graves, L., Nagisetty, V., and Ganesh, V. 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.

Hoang, T., Rana, S., Gupta, S., and Venkatesh, S. 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.

Hong, Y., Yu, L., Yang, H., Ravfogel, S., and Geva, M. Intrinsic evaluation of unlearning using parametric knowledge traces. *arXiv preprint arXiv:2406.11614*, 2024a.

Hong, Y., Zou, Y., Hu, L., Zeng, Z., Wang, D., and Yang, H. Dissecting fine-tuning unlearning in large language models. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (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/.

Ji, J., Liu, Y., Zhang, Y., Liu, G., Kompella, R. R., Liu, S., and Chang, S. Reversing the forget-retain objectives: An efficient llm unlearning framework from logit difference. *arXiv preprint arXiv:2406.08607*, 2024.

Jia, J., Liu, J., Ram, P., Yao, Y., Liu, G., Liu, Y., Sharma, P., and Liu, S. Model sparsity can simplify machine unlearning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=0jZH883i34.

Jia, J., Zhang, Y., Zhang, Y., Liu, J., Runwal, B., Diffenderfer, J., Kailkhura, B., and Liu, S. SOUL: Unlocking the power of second-order optimization for LLM unlearning. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (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/.

Kassem, A., Mahmoud, O., and Saad, S. Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks in language models. In Bouamor, H., Pino, J., and Bali, K. (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/.

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

Lin, Z., Li, P., and Wu, L. Exploring neural network landscapes: Star-shaped and geodesic connectivity. *arXiv preprint arXiv:2404.06391*, 2024.

Liu, H., Li, Z., Hall, D. L. W., Liang, P., and Ma, T. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=3xHDeA8Noi.

Liu, S., Yao, Y., Jia, J., Casper, S., Baracaldo, N., Hase, P., Yao, Y., Liu, C. Y., Xu, X., Li, H., et al. Rethinking

machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*, 2024b.

Liu, Z., Dou, G., Jia, M., Tan, Z., Zeng, Q., Yuan, Y., and Jiang, M. Protecting privacy in multimodal large language models with mllmu-bench. *arXiv preprint arXiv:2410.22108*, 2024c.

Lubana, E. S., Bigelow, E. J., Dick, R. P., Krueger, D., and Tanaka, H. Mechanistic mode connectivity. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (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.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. Learning word vectors for sentiment analysis. In Lin, D., Matsumoto, Y., and Mihalcea, R. (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015/.

Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z. C., and Kolter, J. Z. Tofu: A task of fictitious unlearning for llms, 2024.

Qin, Y., Qian, C., Yi, J., Chen, W., Lin, Y., Han, X., Liu, Z., Sun, M., and Zhou, J. Exploring mode connectivity for pre-trained language models. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (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/.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. 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.

Segura-Bedmar, I., Martínez, P., and Herrero-Zazo, M. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). In Manandhar, S. and Yuret, D. (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/.

Setlur, A., Eysenbach, B., Smith, V., and Levine, S. Adversarial unlearning: Reducing confidence along adversarial directions. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=cJ006qBE8Uv.

Ullah, E., Mai, T., Rao, A., Rossi, R. A., and Arora, R. Machine unlearning via algorithmic stability. In Belkin, M. and Kpotufe, S. (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.

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

Wang, R., Li, Y., and Liu, S. 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.

Wei, S., Zhang, M., Zha, H., and Wu, B. 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.

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

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

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

Zhao, K., Kurmanji, M., Bărbulescu, G.-O., Triantafillou, E., and Triantafillou, P. 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.

Zhao, P., Chen, P.-Y., Das, P., Ramamurthy, K. N., and Lin, X. Bridging mode connectivity in loss landscapes and adversarial robustness. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJgwzCEKwH.

5.1. RQ2: Understanding Machine Unlearning with MCU

We use MCU to gain insights into how unlearning methods behave, when they align, and how task complexity and training dynamics affect their convergence in the loss landscape.

MCU uncovers mechanistic (dis)similarity between unlearning methods Mode connectivity reflects whether two minimizers are mechanistically similar–reside in a shared region of the loss landscape–often indicating shared inductive biases or similar optimization behavior (Lubana et al., 2023). With Rand setting on TOFU, smooth MCU curves occur between GA and GD, NPO when $|D_f| = 1\%$, indicating that GD and NPO are essentially running GA. This is likely because unlearning just $|D_f| = 1\%$ does not have significant effect on preserving performance on the large $|D_r|$. Therefore unlearning $|D_f|$ dominants the optimization process. As $|D_f|$ varies, the similarity may change. RL and GD are much similar when $|D_f| = 5\%$, while becoming dissimilar again when $|D_f| = 10\%$.

Same low-loss manifold can yield different performance While MCU confirms that two unlearning solutions lie in a connected low-loss region, this does not always translate to similar performance acorss all evaluation metrics. For examples, with BT on DDI, we observe stable accuracy on D_f along the MCU curve, but significant fluctuations in ZRF, Figure 6. This 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).

MCU reflects stability of unlearning methods The distinct patterns of MCU can be attributed to the stability of each unlearning method. RL and NPO have certain instability on LLM unlearning (Fan et al., 2024b). With these methods, both linear and quadratic loss landscape of D_f is more irregular, Figures 7–11. This instability can be more prominent with SO optimization (Figure 9), but alleviated with CL (Figure 8), indicating that training dynamics influence not just performance but also the shape of the loss surface.

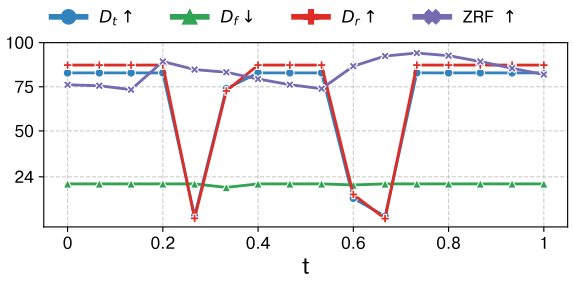


Figure 6: MCU on **DDI dataset**.

Unlearning method may be insufficient to find MCU In some cases, linear MCU exists while quadratic MCU yields much worse curves, see Figure 47 BT method. This is because quadratic MCU requires the underlying unlearning method to find the curve, where some methods fall short during optimization. We observe this phenomenon on BT most frequently, which indicates that BT suffer from significant computation cost and low convergence speed.

Different forget sets induce different loss landscapes We notice that for the same task, different forget sets can yield completely different loss landscapes, which suggests increasing divergence between the solutions and less overlap in the loss-loss regions.

MCU discovers more effective models than minimizers We find intermediate models sampled along the MCU curve may outperform both endpoints in terms of unlearning metrics. 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. Examples include RL on TOFU and NLVR2.

MCU reflects task complexity and unlearning difficulty Failure to observe MCU means that the loss landscape is highly irregular, which may indicate high difficulty of an unlearning task. We observe that MCU appears more often on simpler tasks such as DDI, where even methods like GA–which typically over-forget–show smooth connectivity. In contrast the absence of MCU often indicate the difficulty of unlearning. For example, certain unlearning ratios and methods result in disconnected loss landscape with steep transitions, which suggest high tasks-specific difficulty. These patterns align with prior findings that link unlearning difficulty to factors like sample memorization (Barbulescu & Triantafillou, 2024) and interdependence of forget-retain sets (Zhao et al., 2024), which suggest tasks complexity and data structure considerably shape the loss landscape.

MCU finds challenging forget sets Prior unlearning works identify challenging forget sets using bi-level optimization (Fan et al., 2024a), proximity to test set (Cheng et al., 2023), and LLM-generated stress-test set (Cheng & Amiri,

2025c). We propose that failure to find MCU can be an additional indicator of the difficulty of unlearning, grounded in the geometric properties of the loss landscape. For example, TOFU consistenyly results in smoother MCU that more complex classification datasets, which is consistent with findings in (Jia et al., 2024).

MCU highlights when CL and SO are (in)effective MCU helps determine whether CL or SO meaningfully alter unlearning dynamics and help converge to different manifold. For example, CL often fails to improve performance on TOFU and may even degrade it for RL. Similarly, MCU under FO-SO setting discovers when SO is (in)sufficient. For example, SO is helpful for GA and NPO but not RL on TOFU with $|D_f| = 1\%$ and 5%.

6. 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), updating representations (Cheng & Amiri, 2025a; Chundawat 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; Chen et al., 2025). Unlearning on LLMs recently draws more attention (Eldan & Russinovich, 2023; Ji et al., 2024; Kassem et al., 2023; Cheng & Amiri, 2025c). 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.

7. Influence of Training Dynamics on Mode Connectivity in Unlearning

Machine unlearning methods can be trained by techniques like curriculum learning (CL) (Barbulescu & Triantafillou, 2024; Zhao et al., 2024; Cheng & Amiri, 2024) and second-order (SO) optimization (Jia et al., 2024; Zhang et al., 2024a). MCU provides a principled approach to understand how these factors shape the optimization landscape in unlearning–particularly whether a barrier-free path exists between unlearned models in parameter space. To systematically analyze this, we define several novel experimental conditions based on training paradigms, optimization techniques, and data. These conditions allow us to investigate how the structure of the loss landscape responds to randomness, learning curricula, optimization strategies, and properties of forget set. Below, we describe each condition and its role in analyzing MCU. A summary of all conditions and their combinations is provided in Table 1.

Robustness Against Randomness We begin with a classical mode connectivity setup: training two minimizers (unlearned models in our case) independently from different random seeds. This examines whether MCU is robust to stochasticity in optimization.

Sensitivity to Training Curricula Curriculum learning (Bengio et al., 2009), where training data is introduced to a learner in a specific order, contributes to the efficacy of unlearning (Zhao et al., 2024; Cheng & Amiri, 2024). We investigate two CL-based settings: (a): models for both minimizers are trained with curriculum learning, **-CL** in Table 1; and (b): one minimizer is obtained through curriculum learning and the other does not, **CL-Non-CL** in Table 1. These configurations allow us to assess whether changes in sample learning order affect the emergence of MCU.

Connectivity across Optimization Orders While most unlearning methods use first-order (FO) gradients, recent studies demonstrate that unlearning can benefit from second-order information (curvature via Hessian) in case of LLMs (Jia et al., 2024) and discriminative tasks (Cheng & Amiri, 2024). We analyze MCU under two different SO settings: (a): both minimizers (models) are trained with second-order optimization, **SO-SO** in Table 1; and (b): one minimizer is trained with first order optimization and the other trained with second-order optimization, **FO-SO** in Table 1. This analysis helps us understand how different optimization dynamics shape the unlearning loss landscape.

Configuration	Standard	Both CL	Both SO	Mixed CL/Non-CL	Mixed FO/SO
Minimizers unlearned with different randomness	Rand	Rand-CL	Rand-SO	CL-Non-CL	FO-SO
Minimizers unlearned with different unlearning methods	Met	Met-CL	Met-SO	Met-CL-Non-CL	Met-FO-SO

Table 1: Experimental settings to study MCU. Each configuration varies in optimization strategy, training curriculum, or unlearning method. Rand: minimizers trained with different random seeds, CL and SO: curriculum learning and second-order optimization respectively, Met: minimizers trained using different unlearning methods. All settings except for "Rand" are novel in the context of mode connectivity.

Compatibility across Unlearning Method All unlearning methods share the common objective of removing knowledge of D_f while retaining knowledge of D_r. We examine whether minimizers derived from different unlearning methods can be smoothly connected? We hypothesize that methods with similar formulation, inner-working, and objectives are more likely to establish connectivity. This experiment provides a lens into methodological similarity and compatibility of different unlearning algorithms.

Experimental Novelty To the best of our knowledge, with the exception of the randomness factor (see "Robustness Against Randomness" above), 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.

8. Experimental Setup

Datasets and Forget Sets We analyze MCU on generation and classification tasks. For LLM unlearning, we use the TOFU dataset (Maini et al., 2024). 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 (Maas et al., 2011). The original models and standard data splits are from TOFU and MU-Bench, where 1%, 5%, 10% of data is unlearned from TOFU and 2%, 4%, 6%, 8%, 10% of data is unlearned from MU-Bench.

Unlearning Methods We use four LLM unlearning methods for TOFU: Gradient Ascent (GA) (Golatkar et al., 2020), Random Labeling (RL) (Graves et al., 2021), GradDiff (GD) (Maini et al., 2024), and Negative Preference Optimization (NPO) (Zhang et al., 2024b). For classification tasks we use: Gradient Ascent (GA) (Golatkar et al., 2020), Random Labeling (RL) (Graves et al., 2021), Bad Teaching (BT) (Chundawat et al., 2023), and Saliency Unlearning (SU) (Fan et al., 2024c). These methods cover a diverse set of unlearning paradigms and are commonly used in existing works. Appendix 9 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). We evaluate each induced unlearned model based on the following standard evaluation metrics. On classification tasks, we use accuracy on test set $D_t(\uparrow)$, accuracy on forget set $D_f(\downarrow)$, accuracy on retain set $D_r(\uparrow)$, and Zero-Retrain Forgetting score ZFR ($\uparrow$) which measures the prediction similarity of D_f between unlearned and original models. On TOFU, we use Forget Quality ($\uparrow$) and Model Utility ($\uparrow$) which is a p-value of Kolmogorov-Smirnov test (KS-Test). Additionally, we use ROUGE-L recall score on forget set ($\downarrow$), retain set ($\uparrow$), and real authors ($\uparrow$). Following previous mode connectivity work on language models (Qin et al., 2022), we sample 16 points with equal step size in $[0, 1]$. Appendix 10 provides additional details.

9. 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 parameter 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 Graclsff (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 functions only the losing response with no winning response.

10. 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 aclstion, 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.

On TOFU (Maini et al., 2024), we follow the original paper to use 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. Aclstionally, we also include verbatim evaluation using ROUGE-L recall score on Retain Authors (↑), Forget Authors (↓), Real Authors (↑), and World Knowledge (↑).

11. 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

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

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

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

12. Additional Results

We present detailed results on TOFU in Figure 7–46 and on classification datasets in Figure 47–106.

12.1. 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 7 for details. As the size of forget set increases, indicated by different rows in Figure 7, 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 47), 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.

Unlearning effectiveness does not guarantee smooth mode connectivity SalUn is generally an effective unlearning method. However, it often fails to yield smooth mode connectivity curves. This suggests that strong unlearning efficacy does not necessarily imply the existence of smooth paths in the loss landscape.

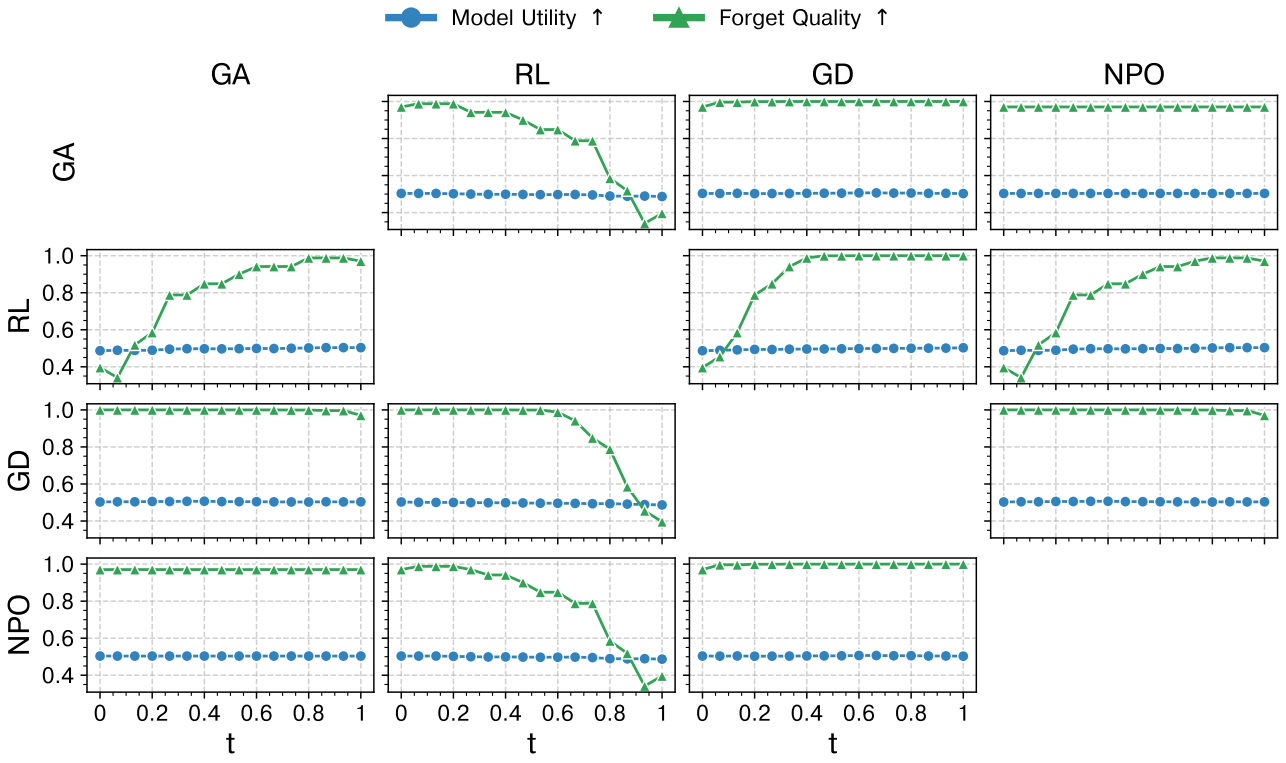


Figure 3: Linear MCU when $|D_f| = 10\%$

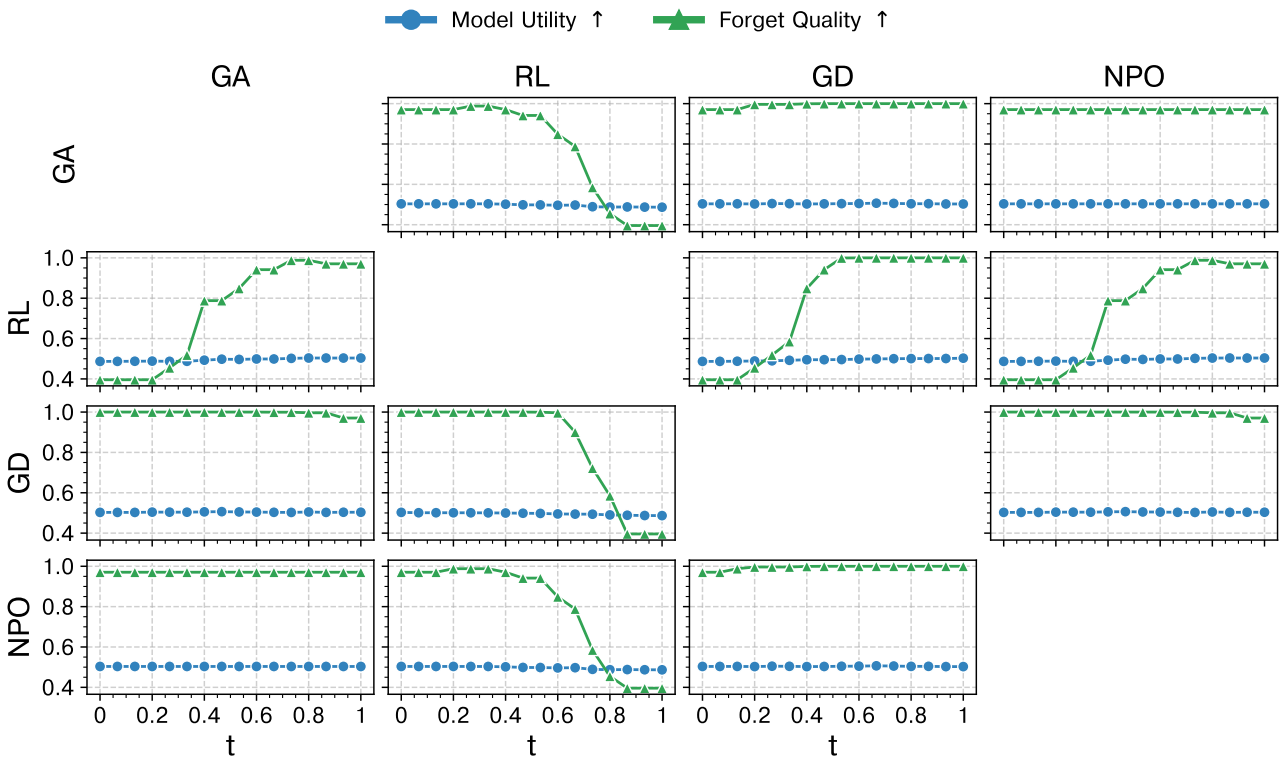


Figure 4: Quadratic MCU when $|D_f| = 10\%$

Figure 5: MCU under **Met** setting on **TOFU dataset**. Methods on rows and columns correspond to θ_1' and θ_2' respectively. In (a), linear MCU is symmetric. In (b), Quadratic MCU is asymmetric as the curve is optimized using methods shown in rows. Additional results are shown in Appendix 12, Figures 18–46 for TOFU and Figures 62–106 for classification tasks.

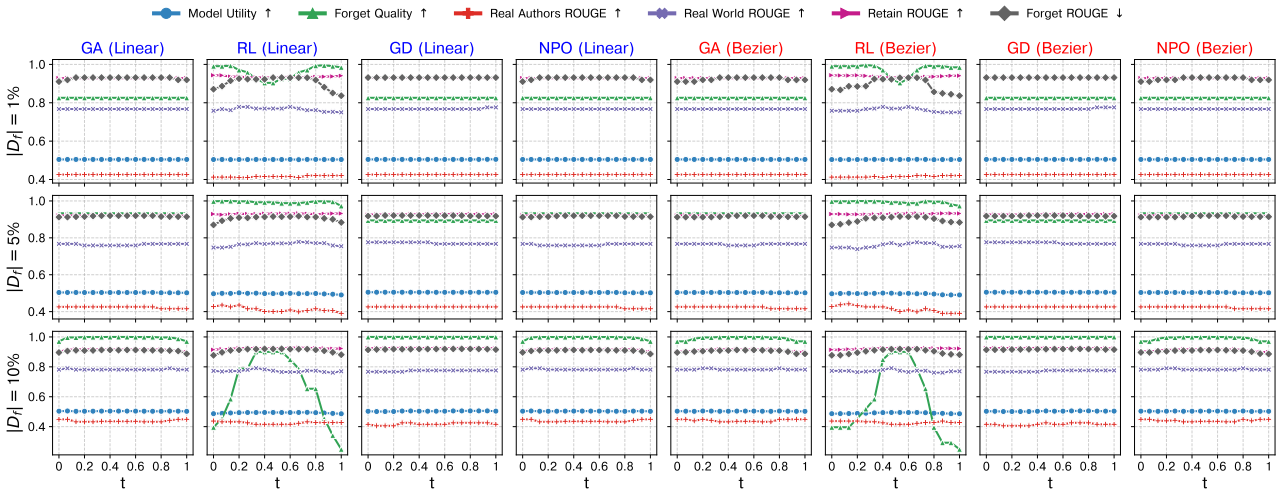


Figure 7: MCU under **Rand** setting on **TOFU dataset**.

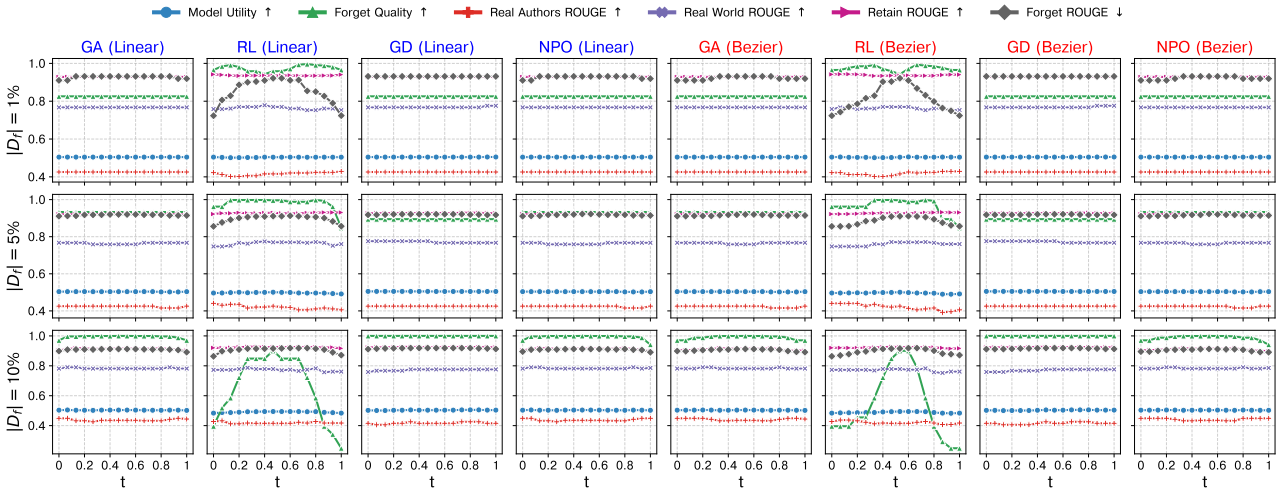


Figure 8: MCU under **Rand-CL** setting on **TOFU dataset**.

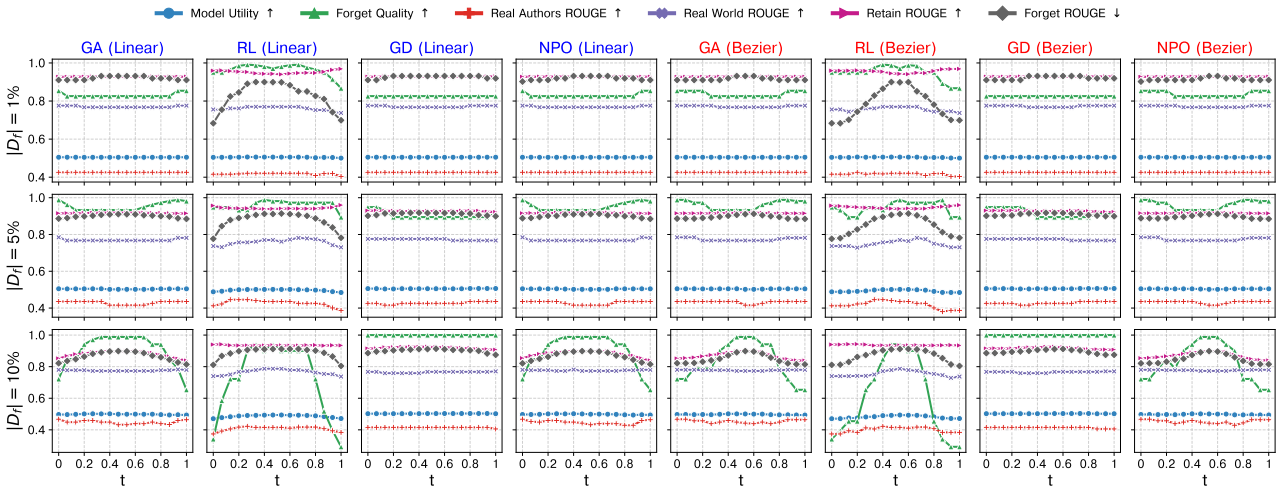


Figure 9: MCU under **Rand-SO** setting on **TOFU dataset**.

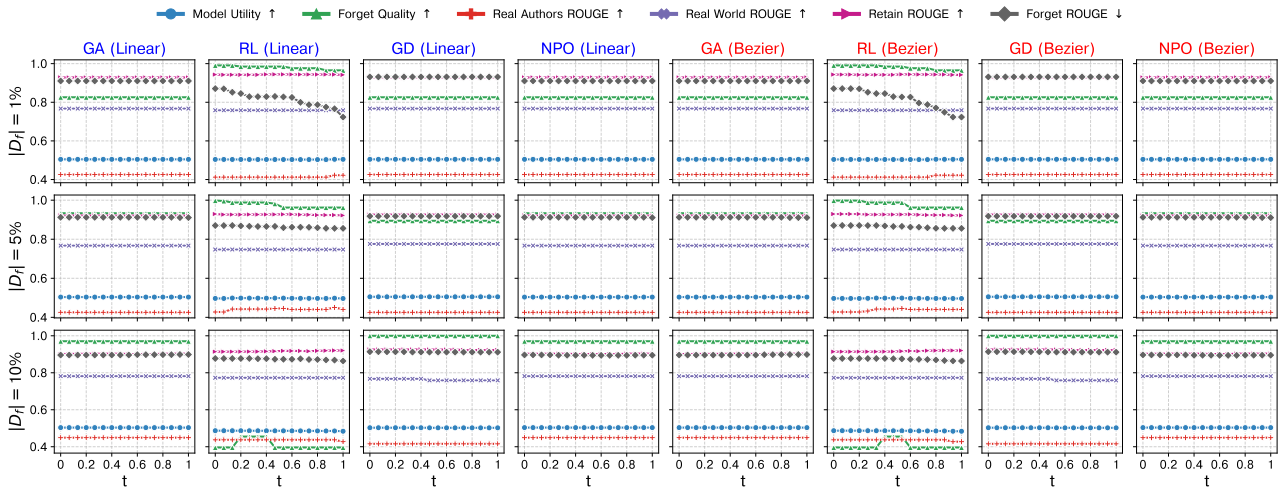


Figure 10: MCU under **CL-Non-CL** setting on **TOFU dataset**.

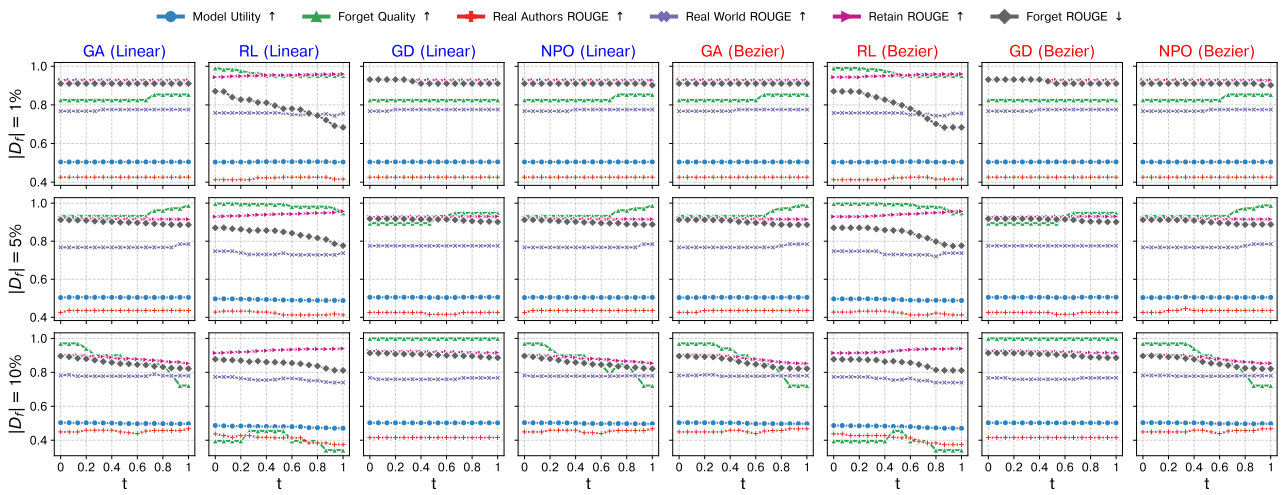


Figure 11: MCU under **FO-SO** setting on **TOFU dataset**.

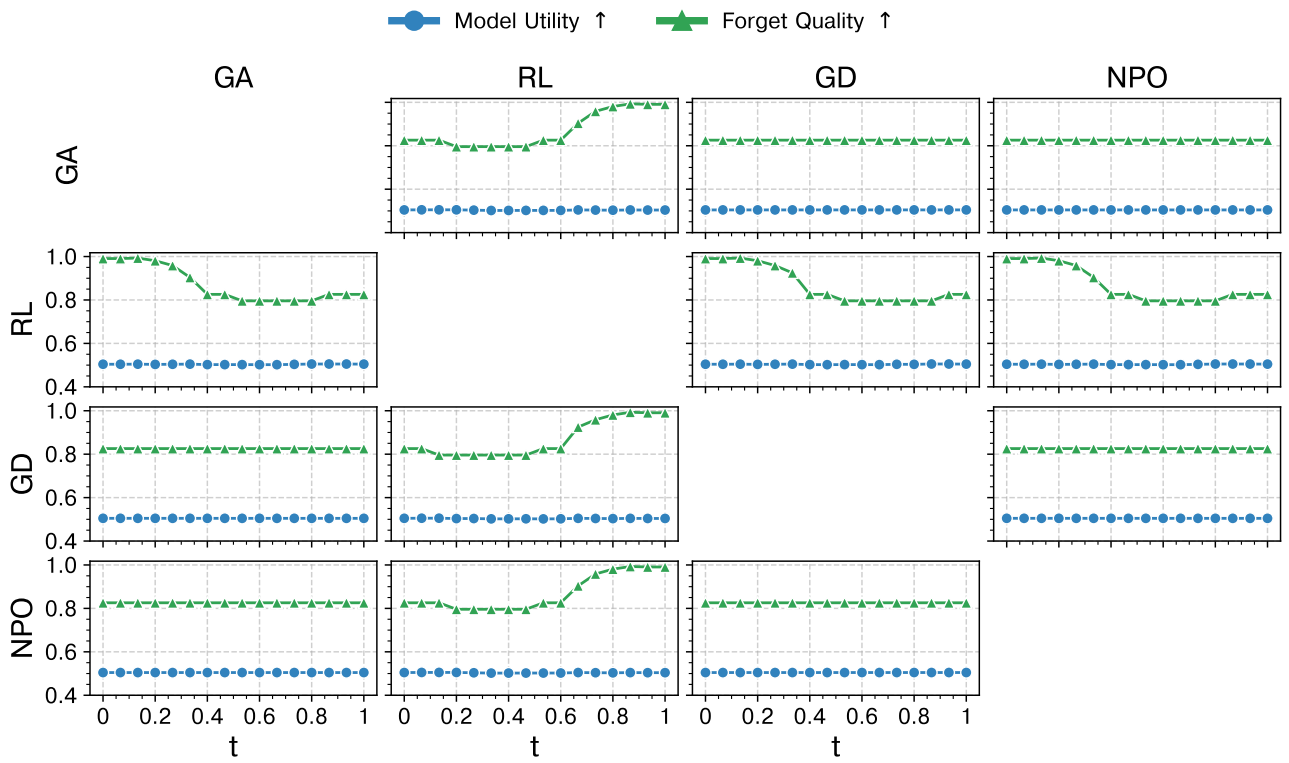


Figure 12: Linear MCU when $|D_f| = 1\%$

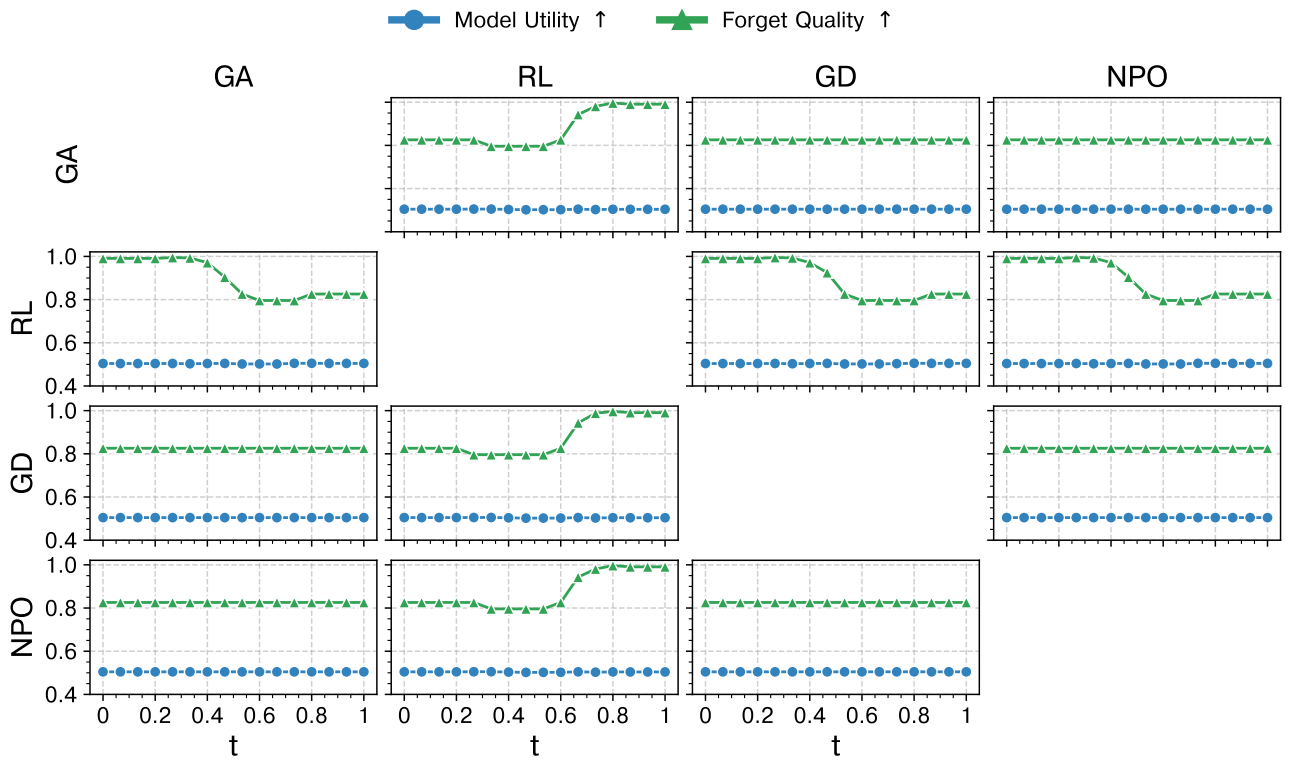
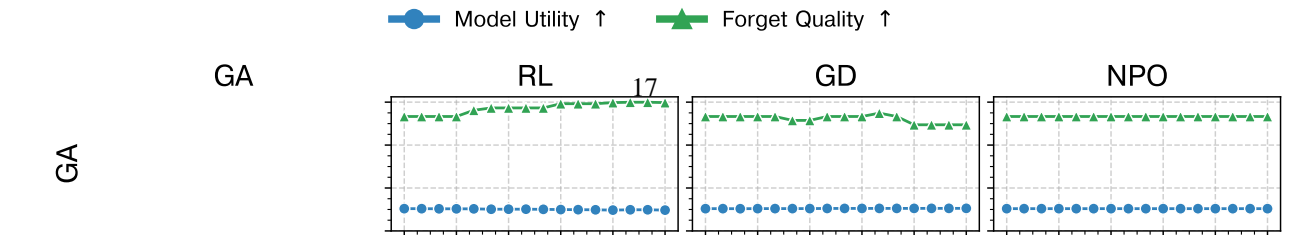


Figure 13: Quadratic MCU when $|D_f| = 1\%$



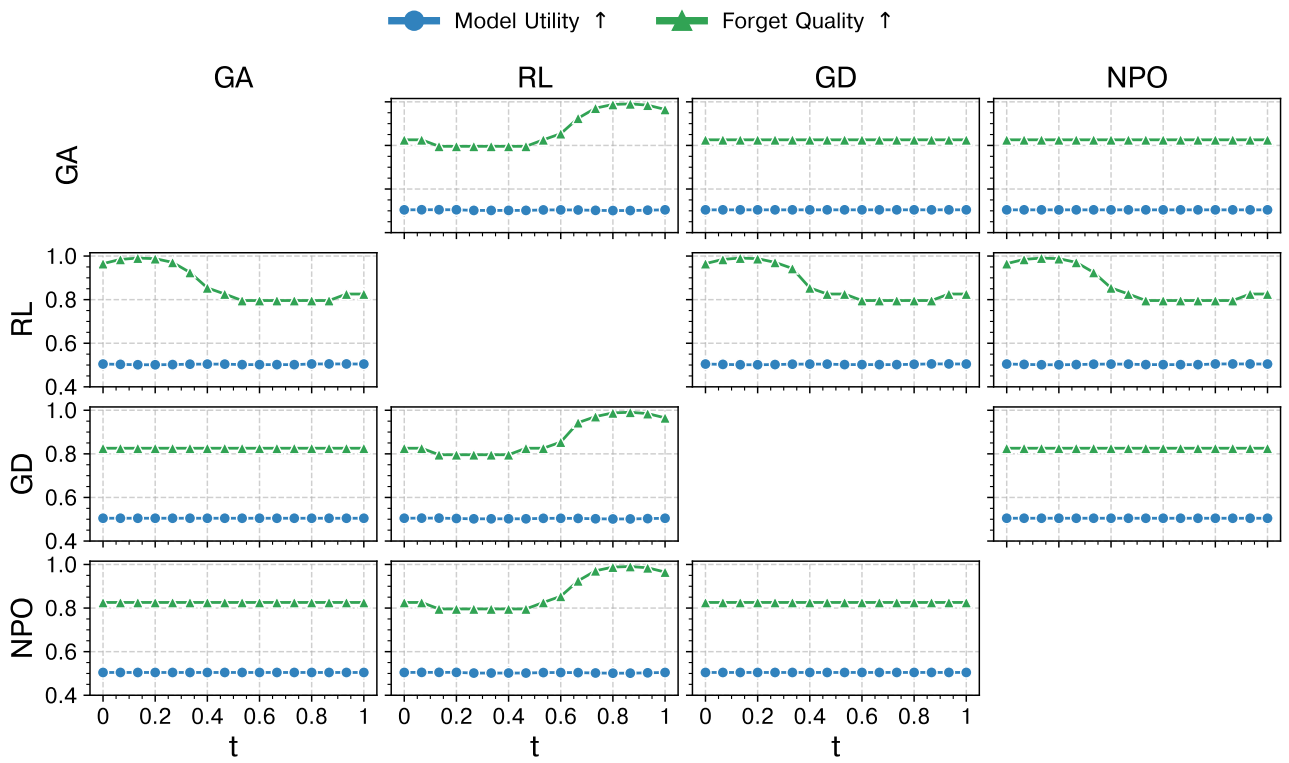


Figure 19: Linear MCU when $|D_f| = 1\%$

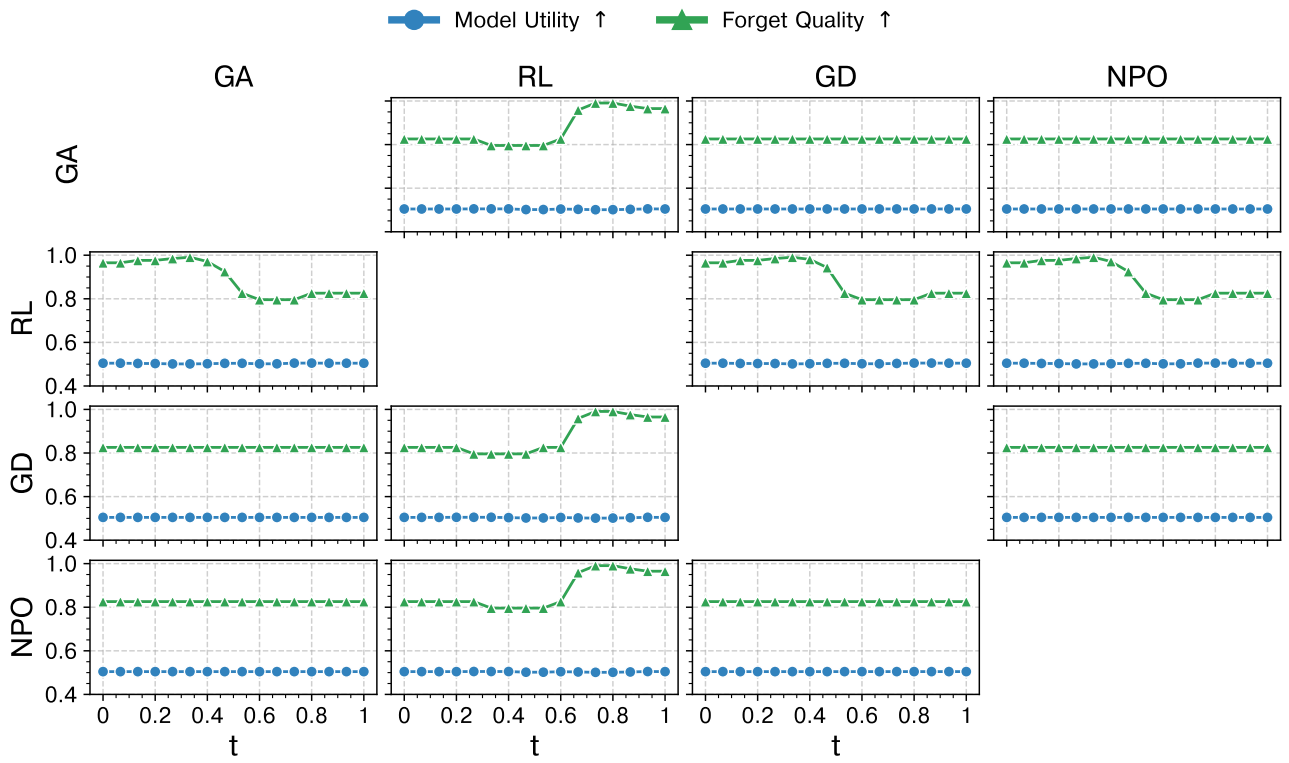
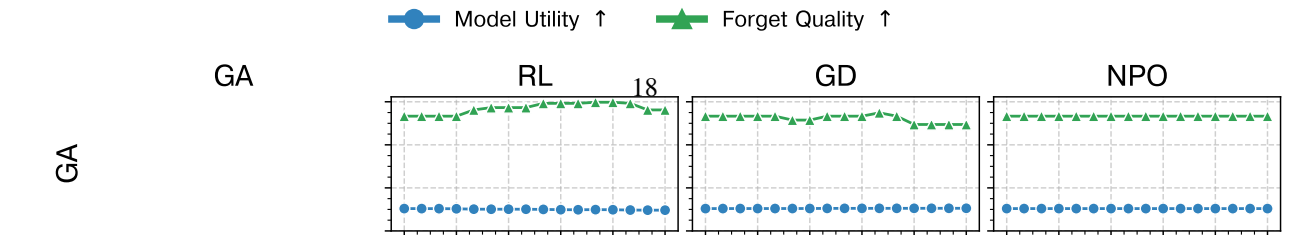


Figure 20: Bezier MCU when $|D_f| = 1\%$



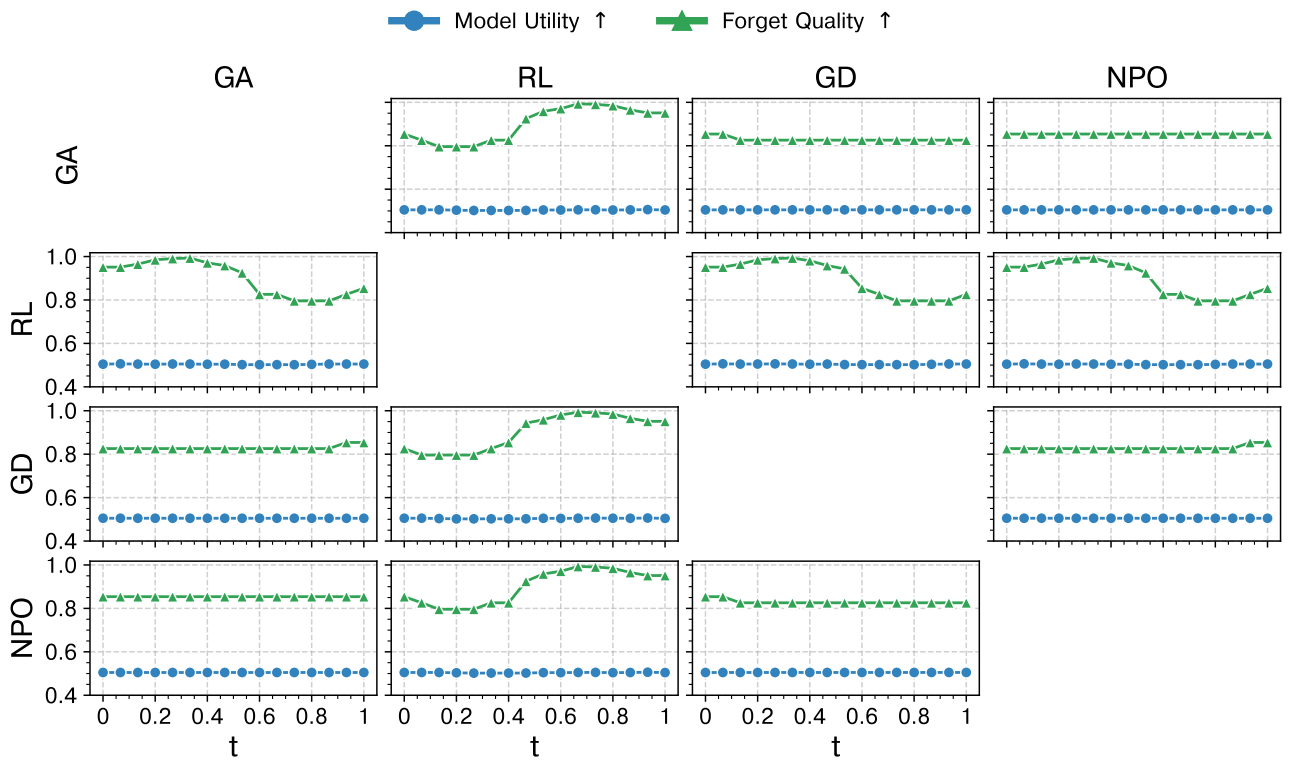


Figure 26: Linear MCU when $|D_f| = 1\%$

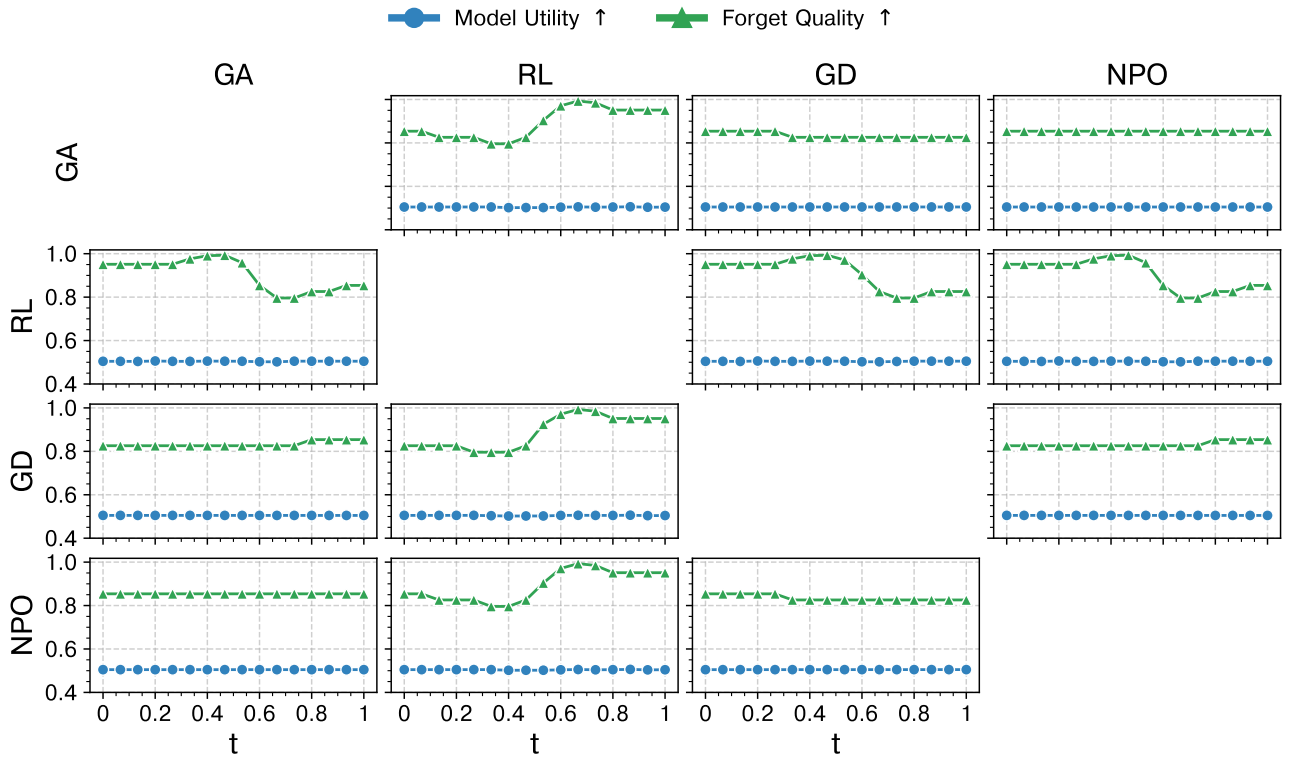
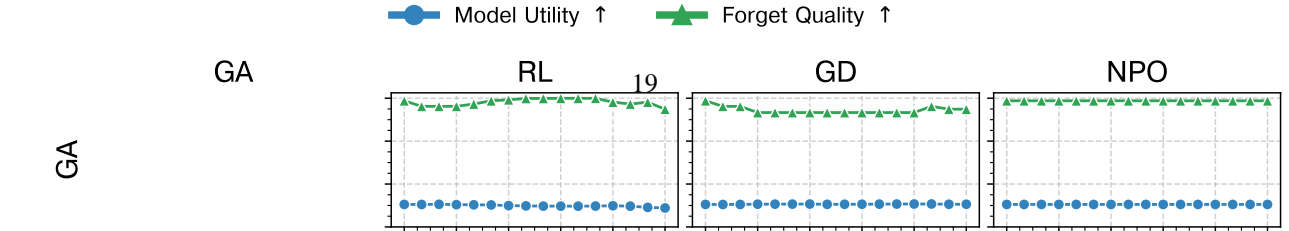


Figure 27: Bezier MCU when $|D_f| = 1\%$



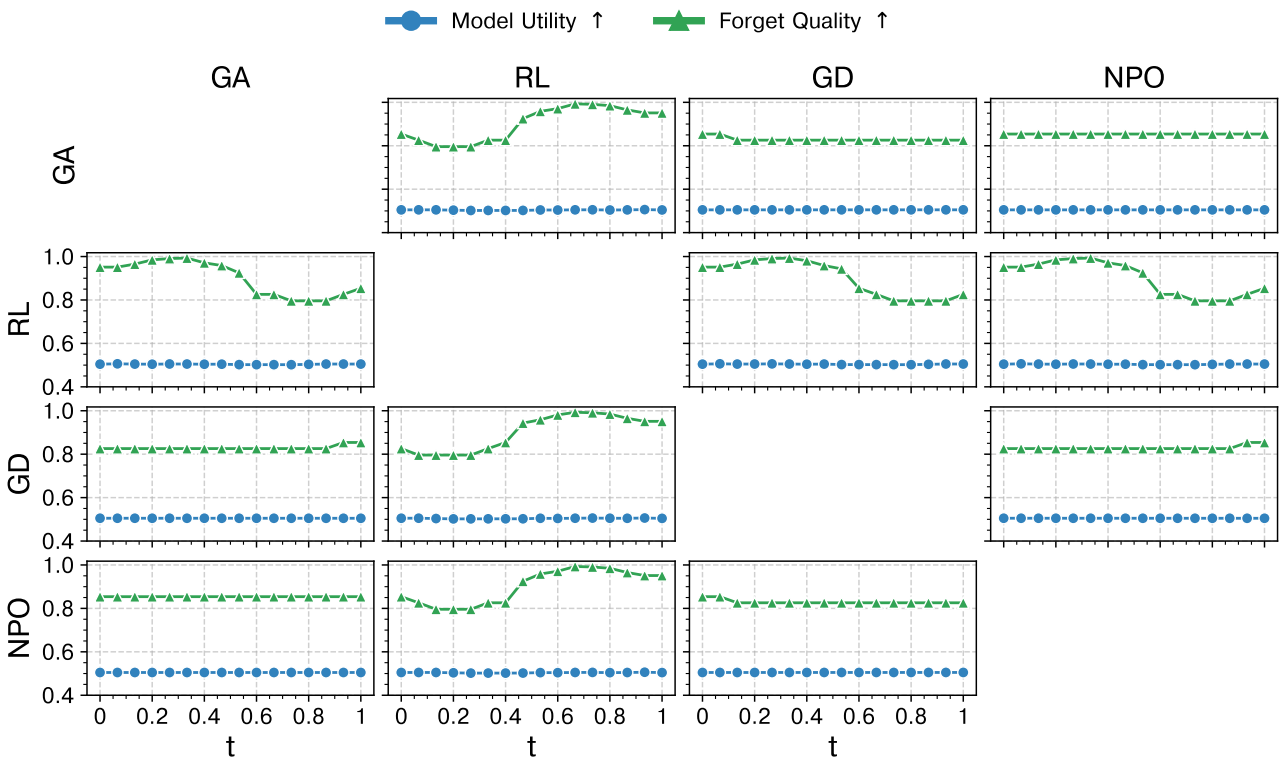


Figure 33: Linear MCU when $|D_f| = 1\%$

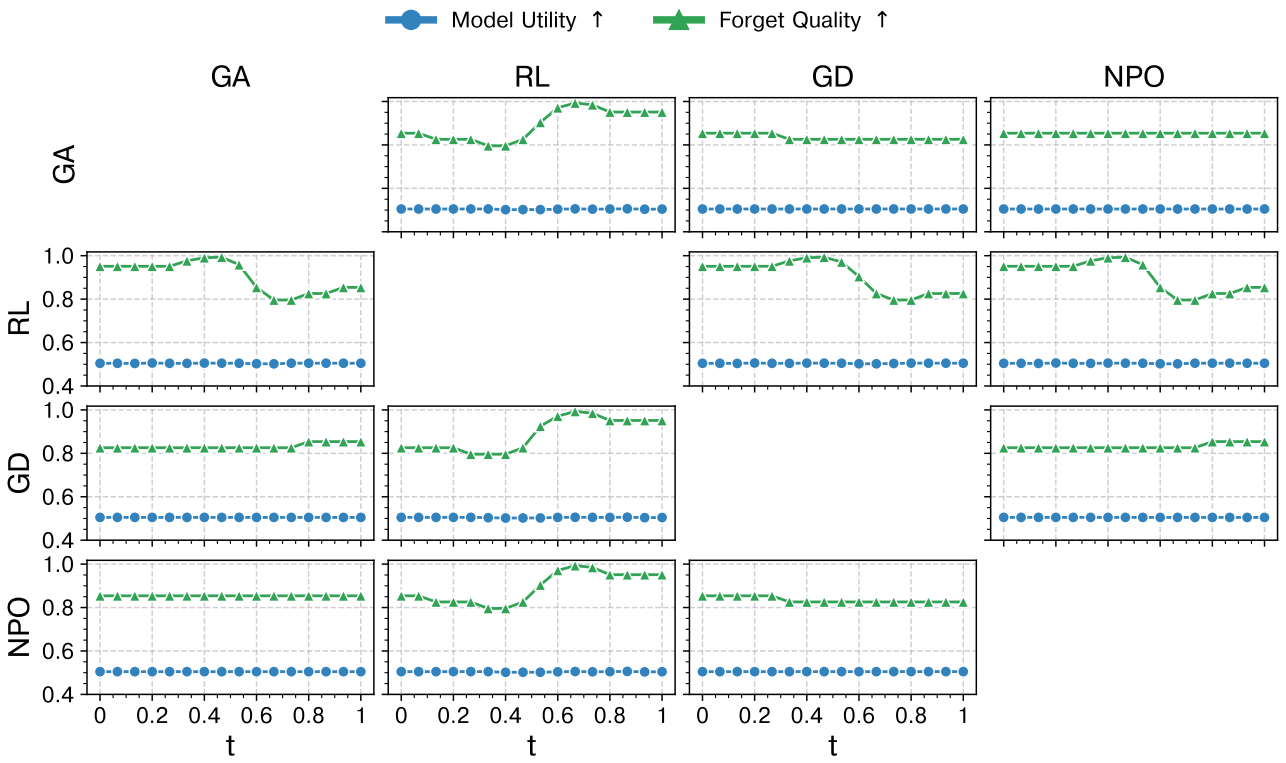
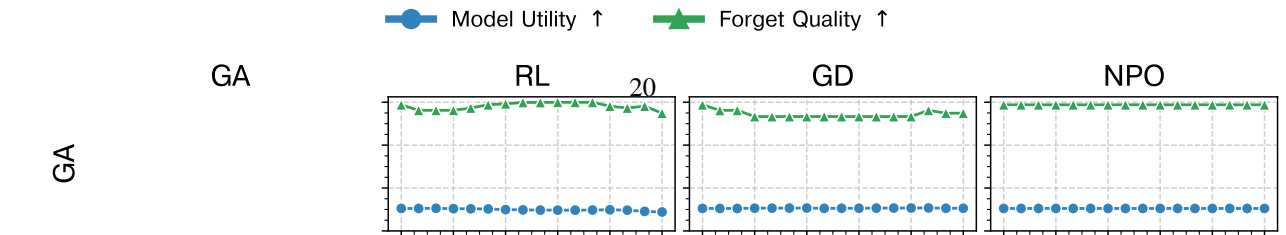


Figure 34: Bezier MCU when $|D_f| = 1\%$



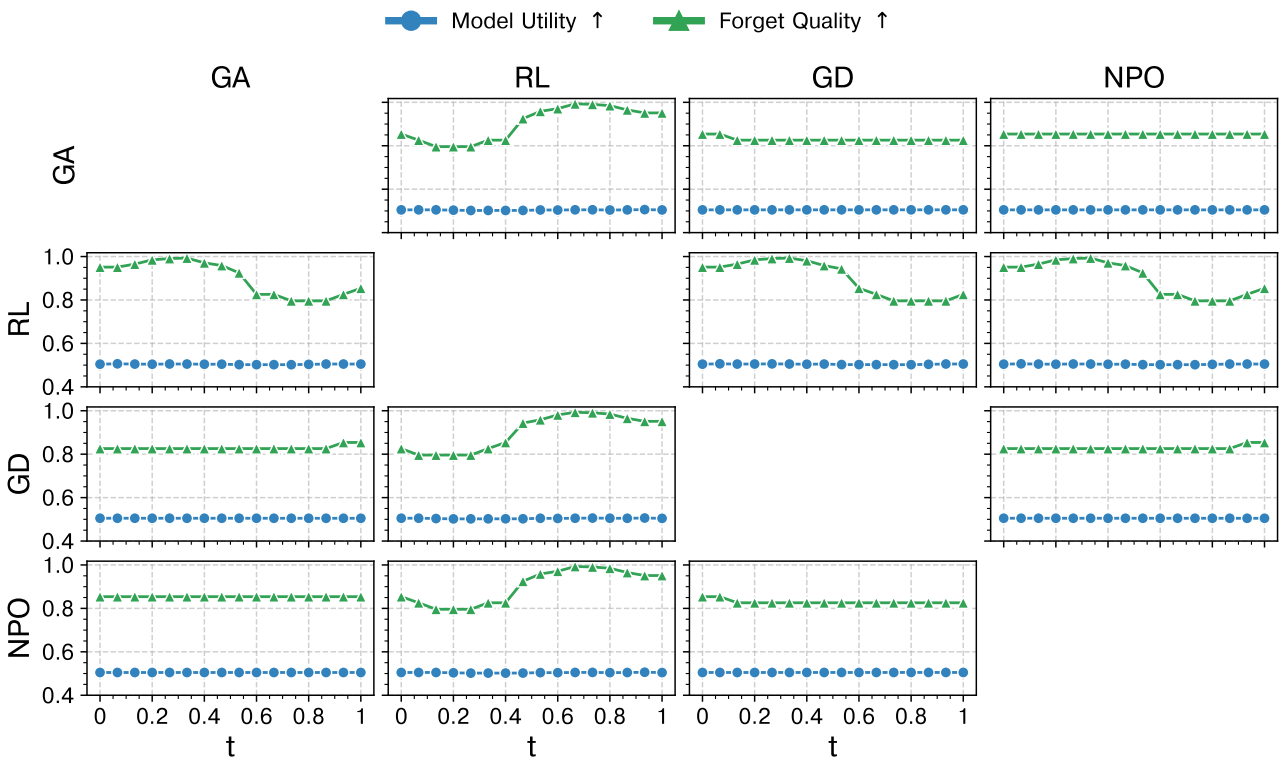


Figure 40: Linear MCU when $|D_f| = 1\%$

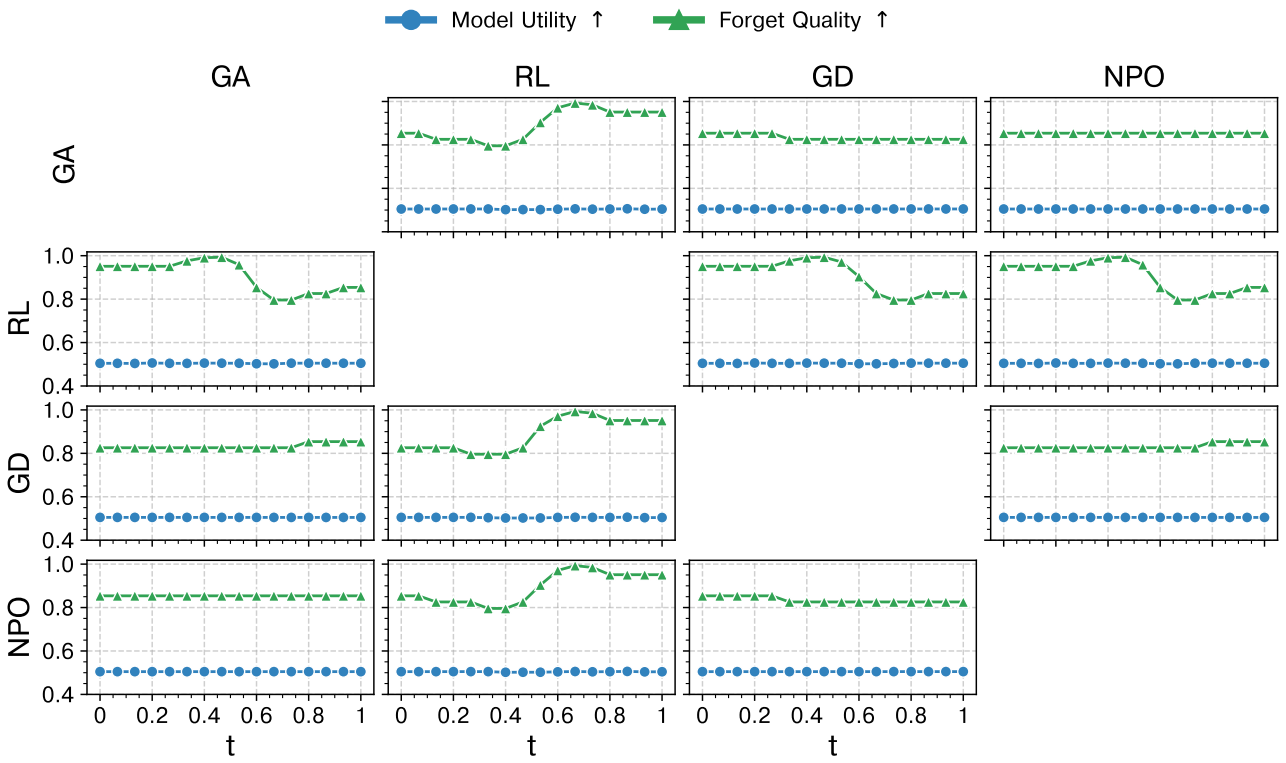
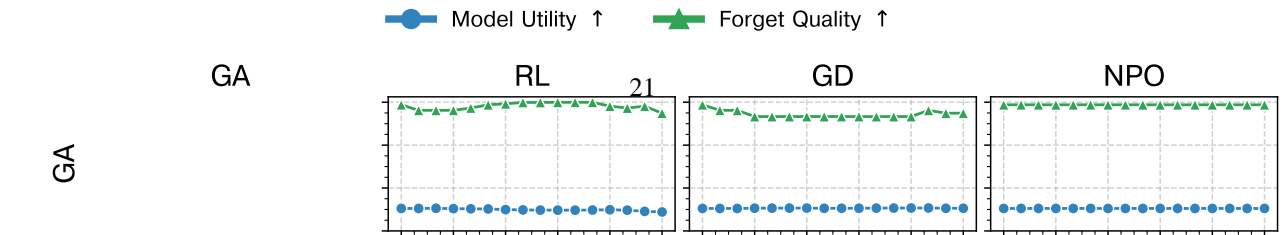


Figure 41: Bezier MCU when $|D_f| = 1\%$



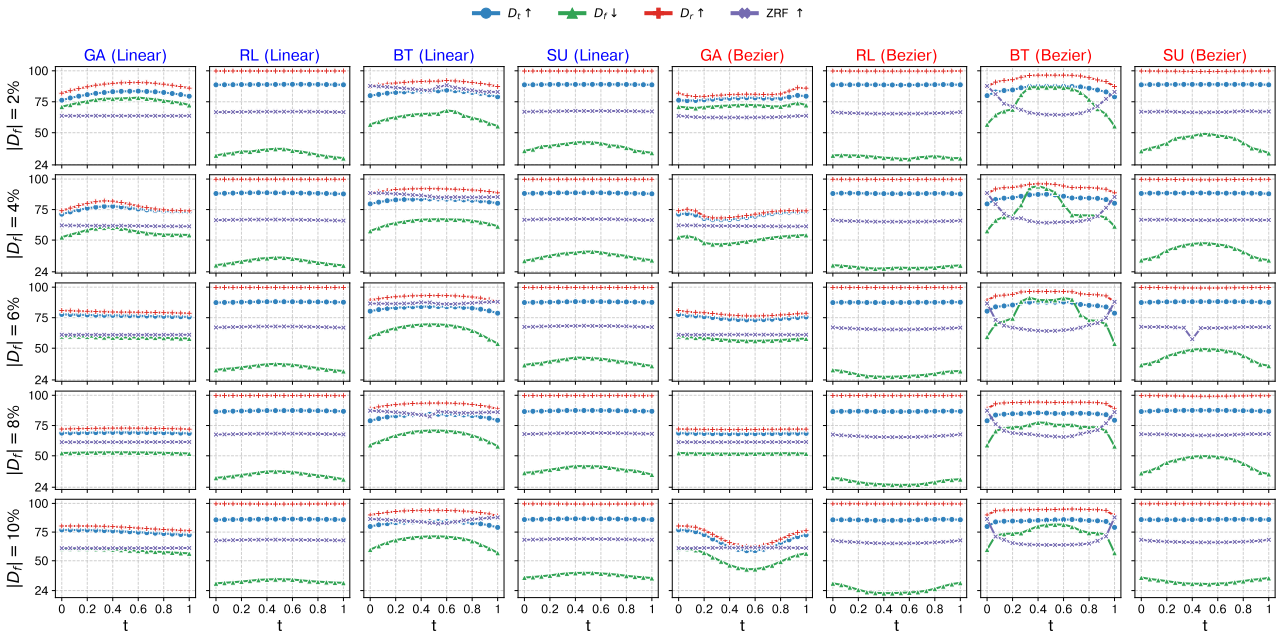


Figure 47: MCU under **Rand** setting on **classification dataset**.

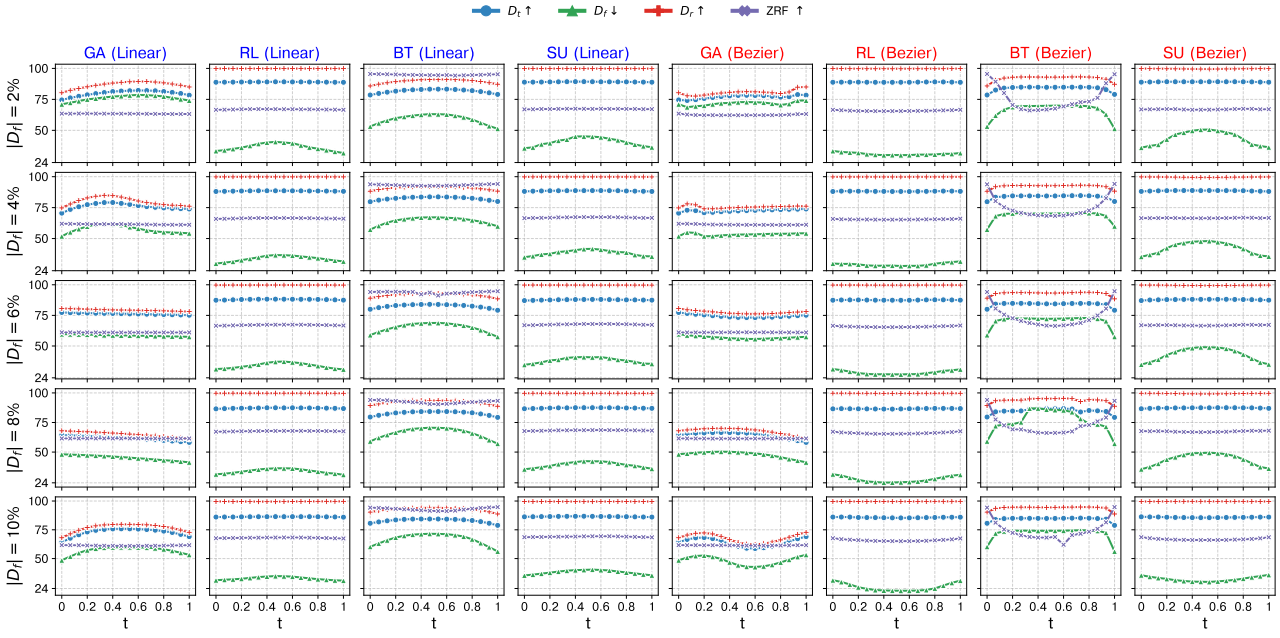


Figure 48: MCU under **Rand-CL** setting on **classification dataset**.

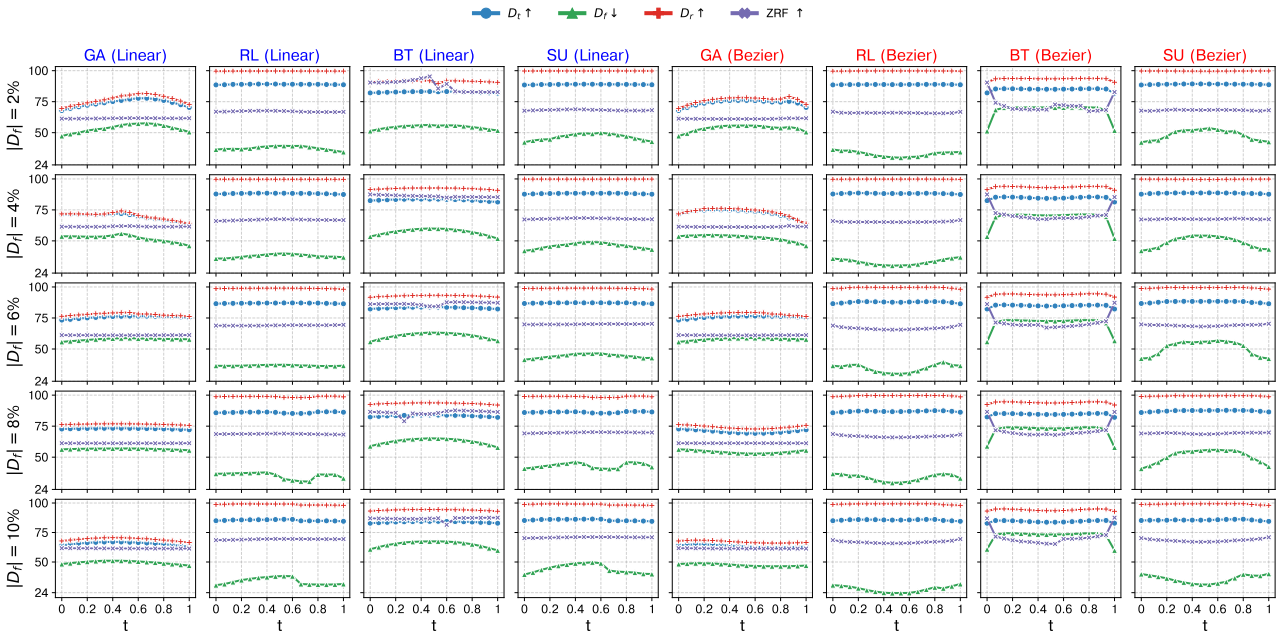


Figure 49: MCU under **Rand-SO** setting on **classification dataset**.

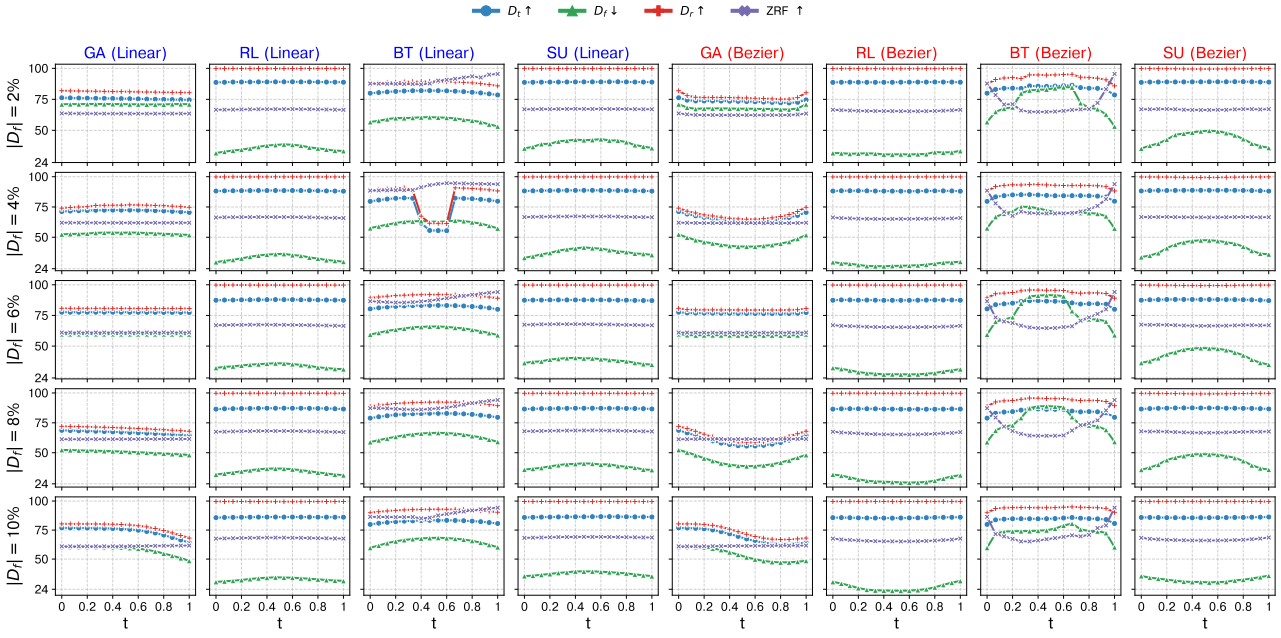


Figure 50: MCU under **CL-Non-CL** setting on **classification dataset**.

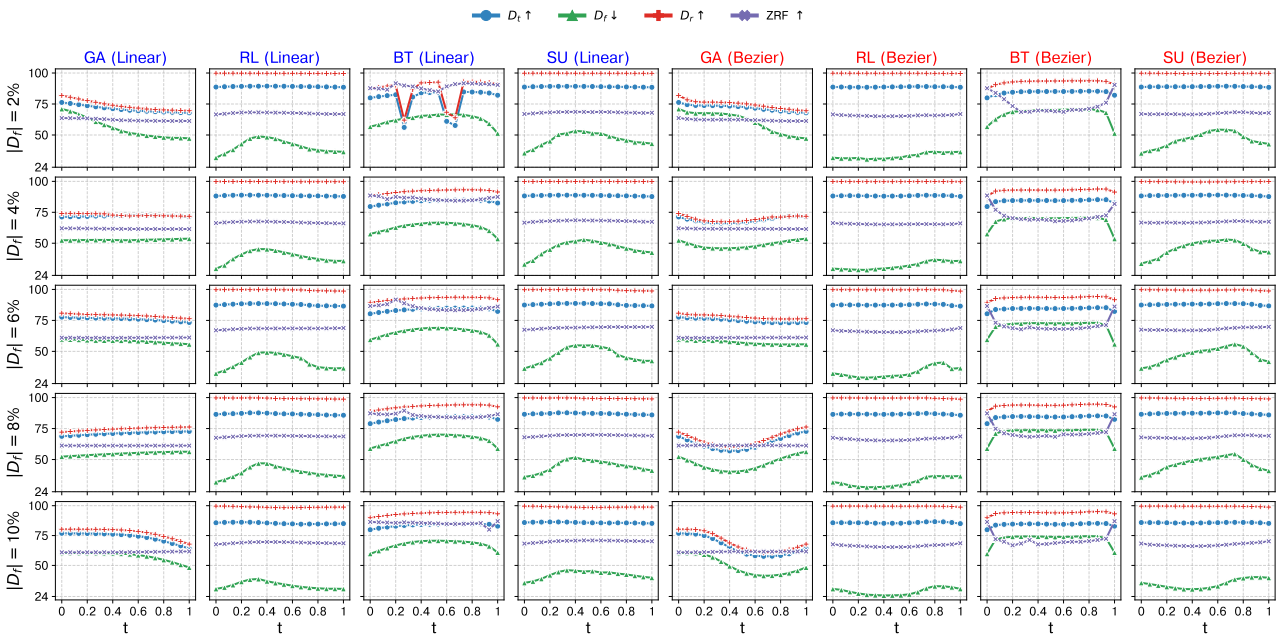


Figure 51: MCU under **FO-SO** setting on **classification dataset**.

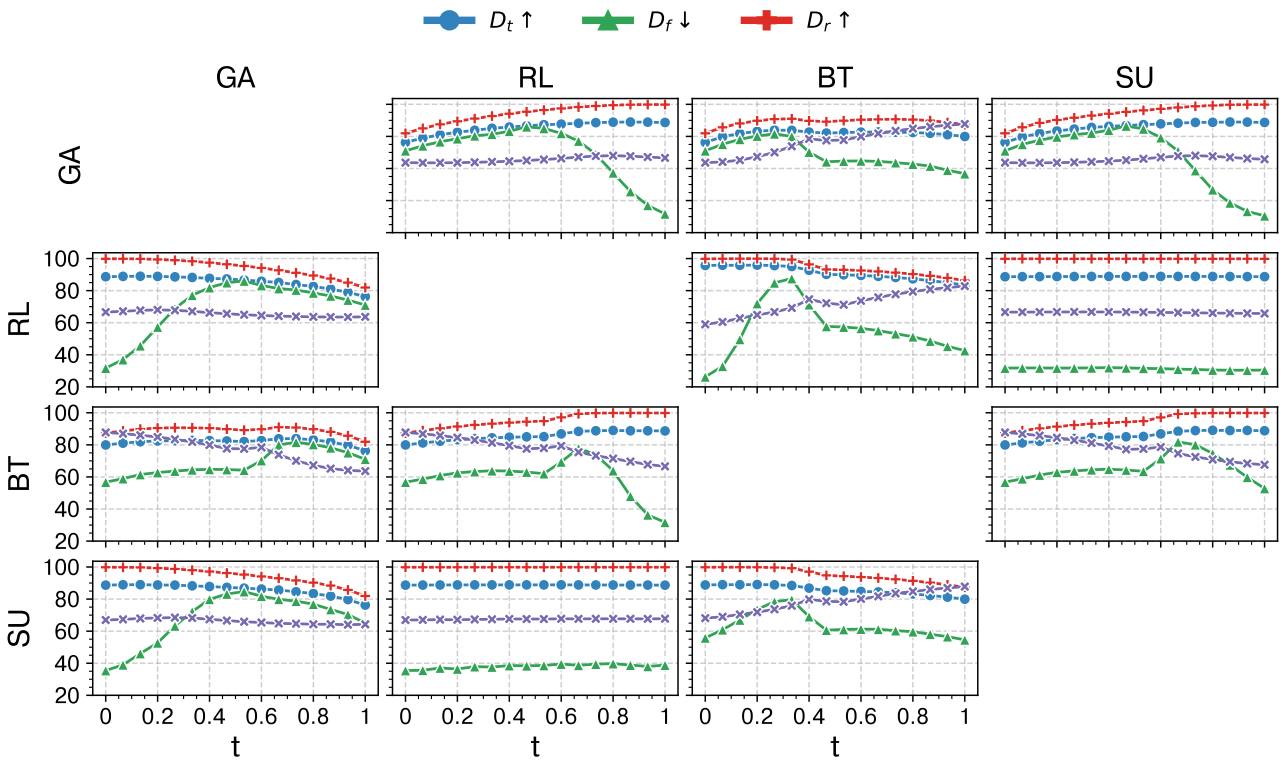


Figure 52: Linear MCU when $|D_f| = 2.0\%$

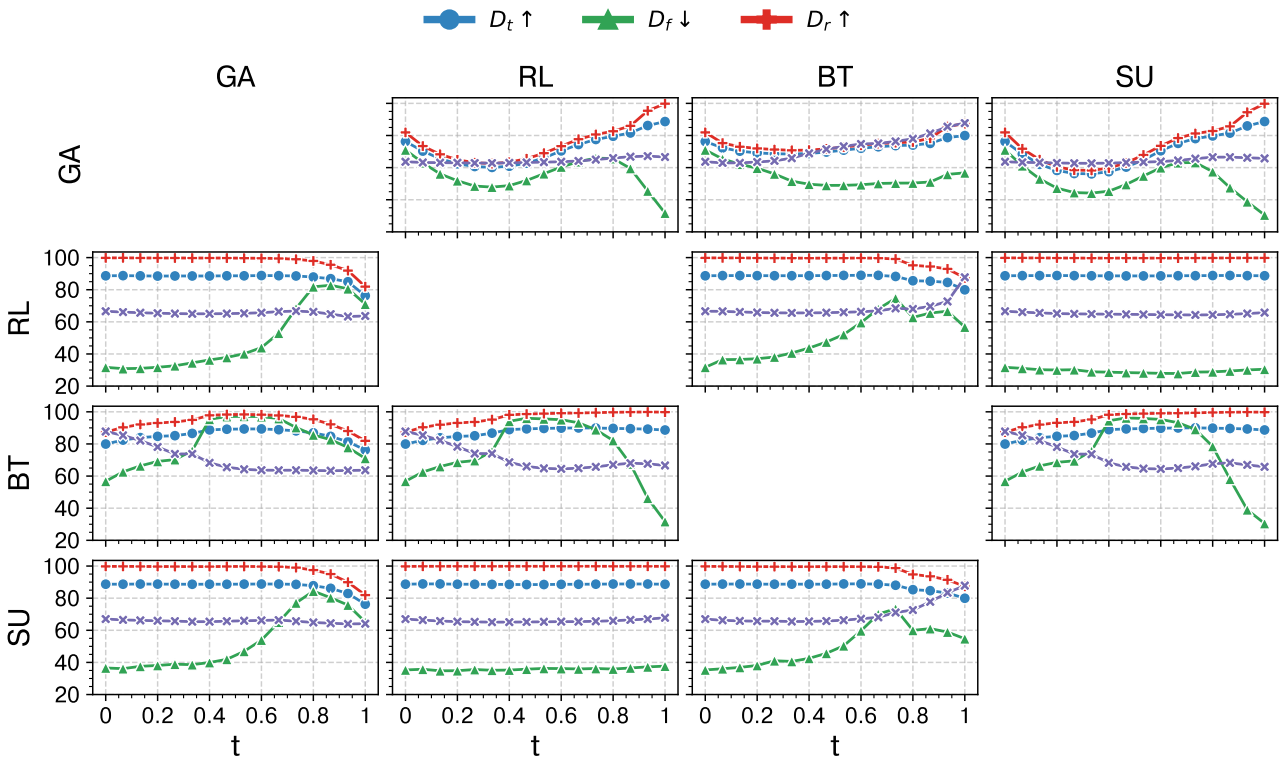
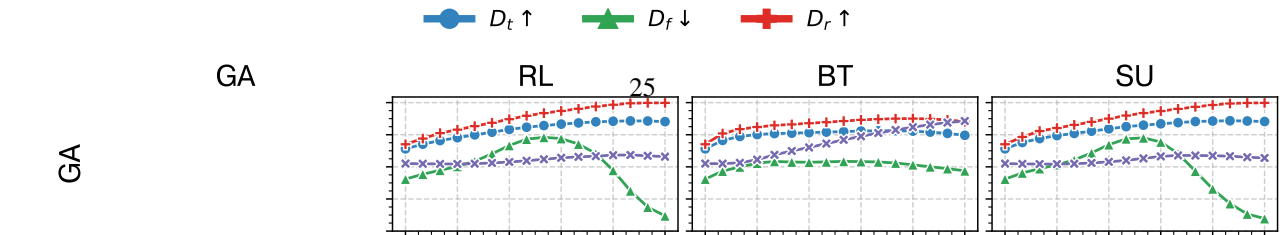


Figure 53: Quadratic MCU when $|D_f| = 2.0\%$



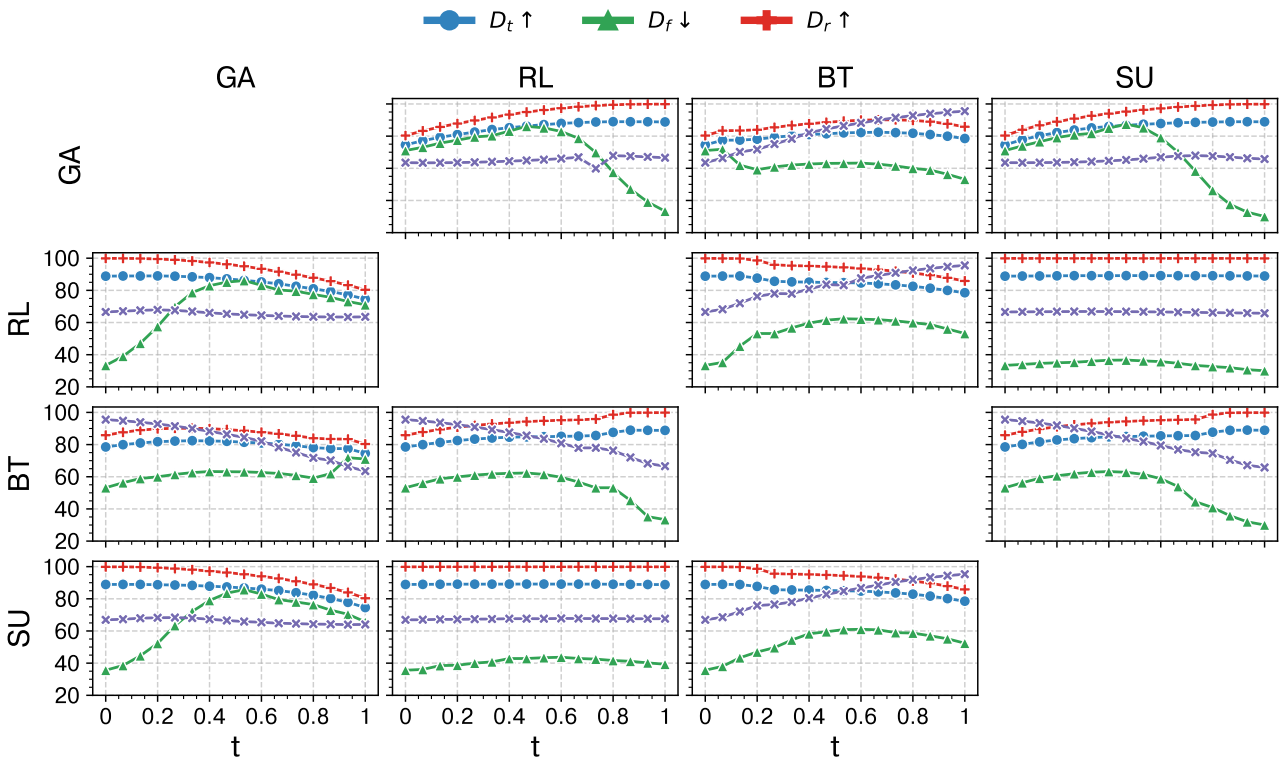


Figure 63: Linear MCU when $|D_f| = 2.0\%$

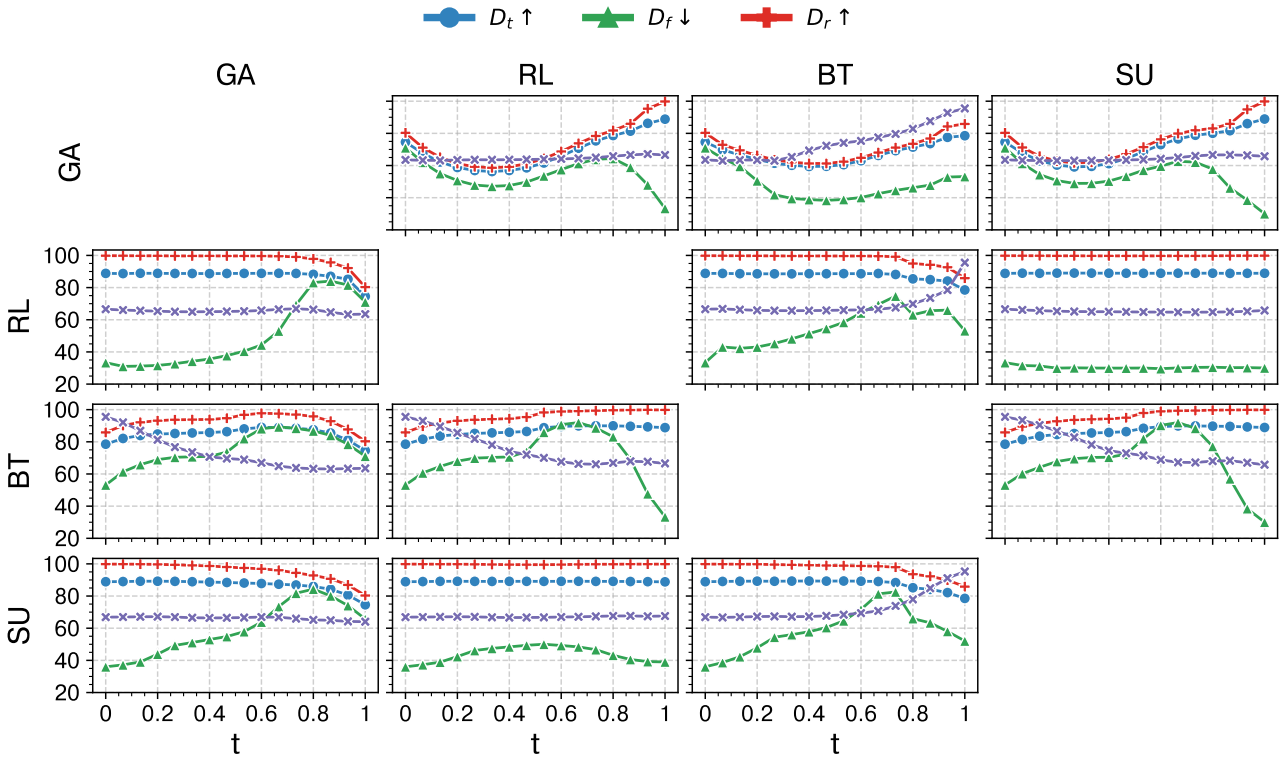
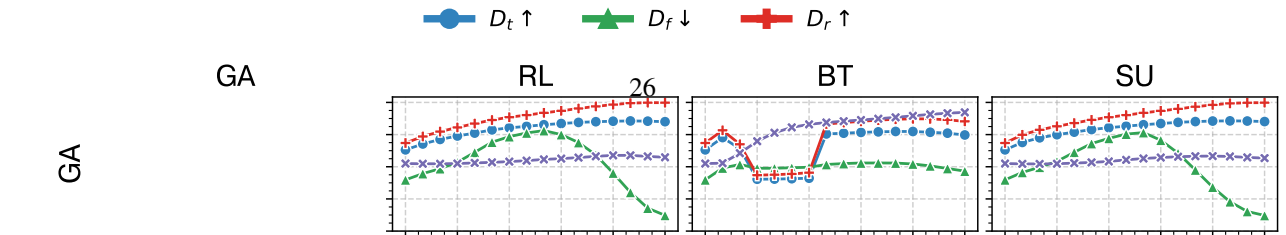


Figure 64: Quadratic MCU when $|D_f| = 2.0\%$



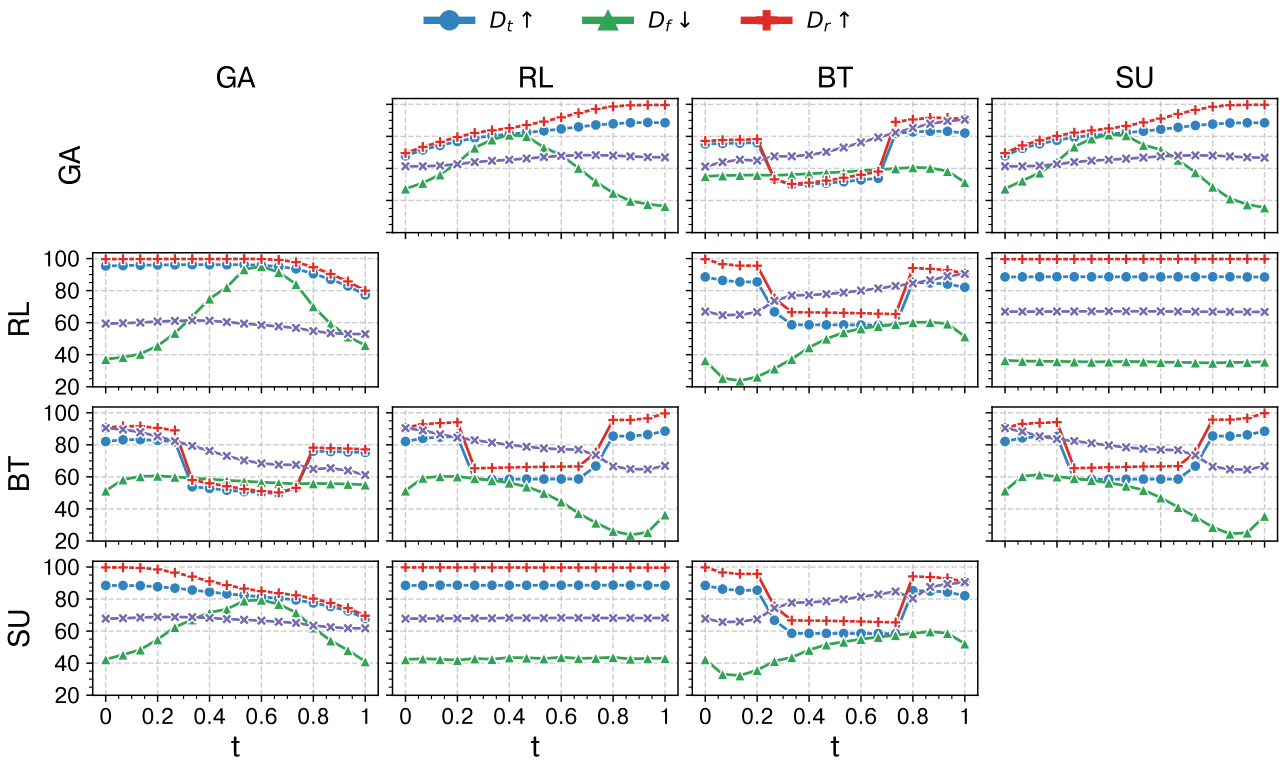


Figure 74: Linear MCU when $|D_f| = 2.0\%$

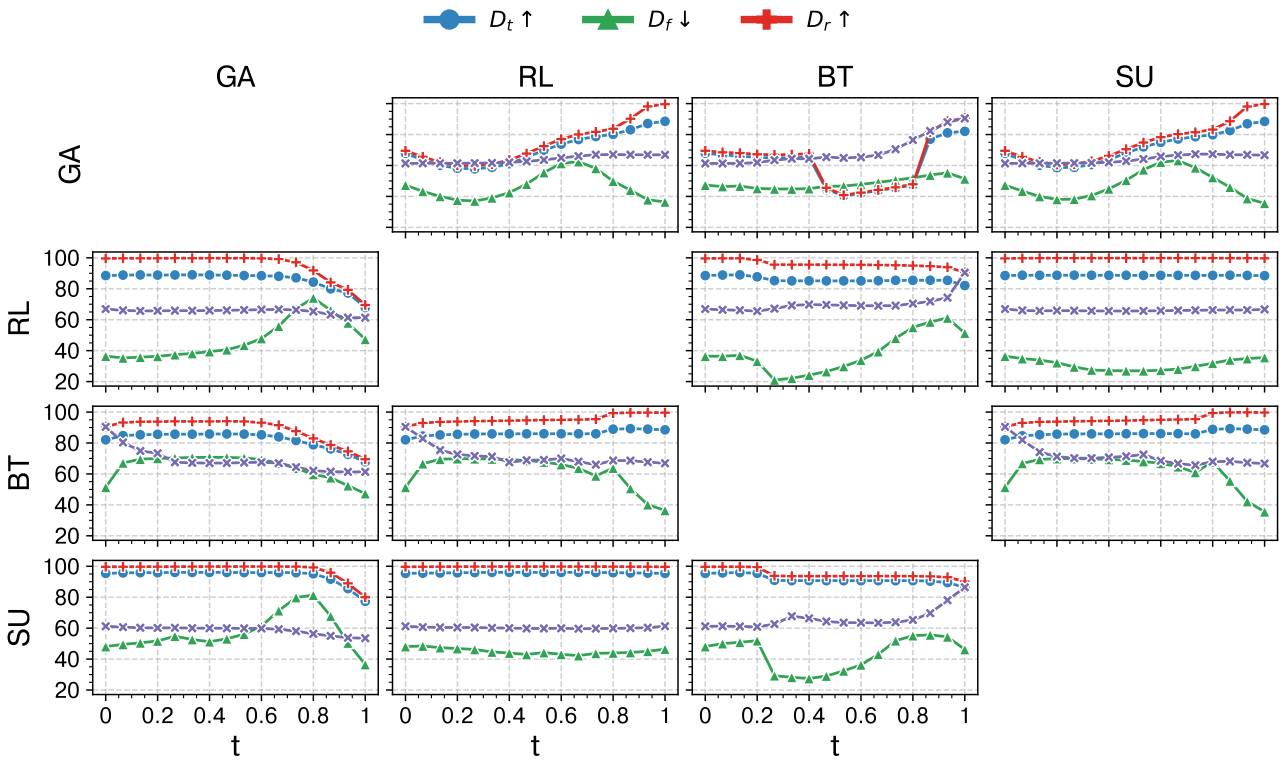
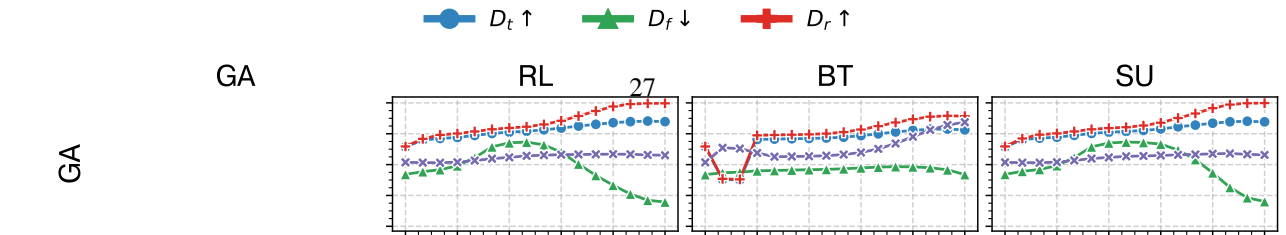


Figure 75: Quadratic MCU when $|D_f| = 2.0\%$



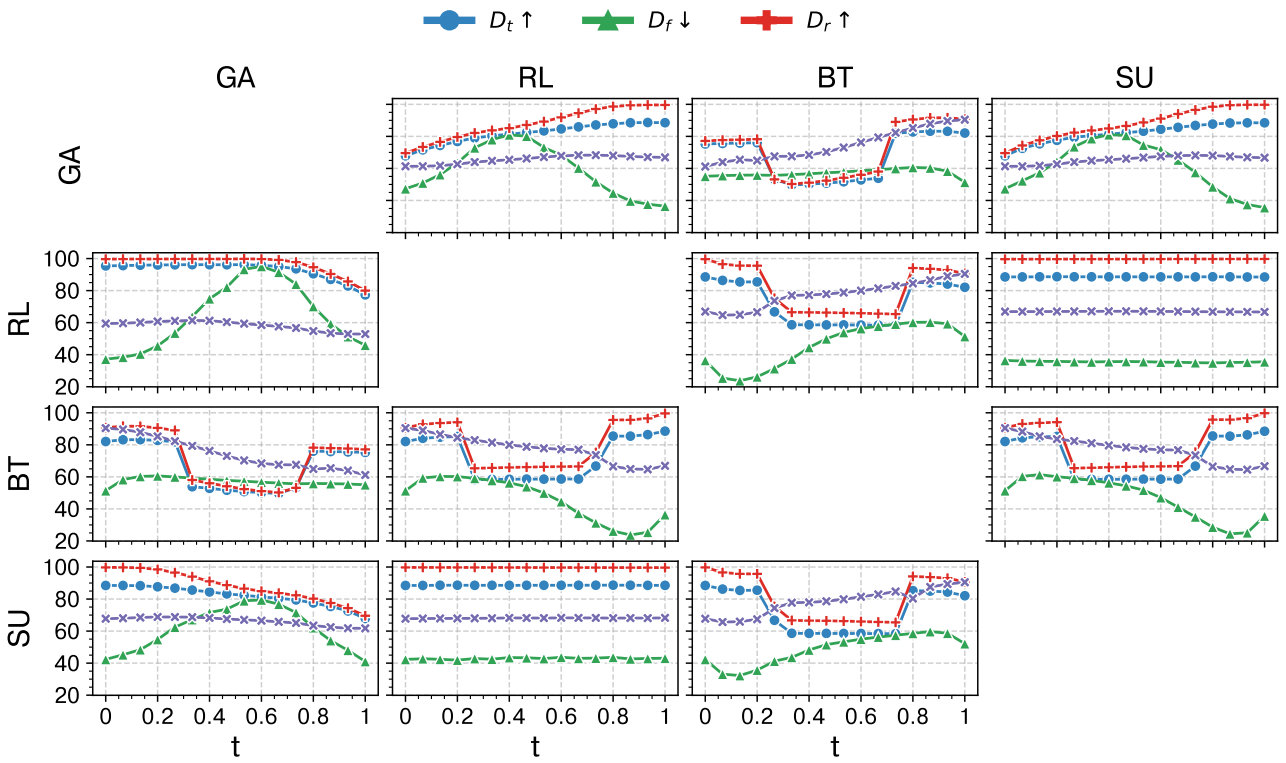


Figure 85: Linear MCU when $|D_f| = 2.0\%$

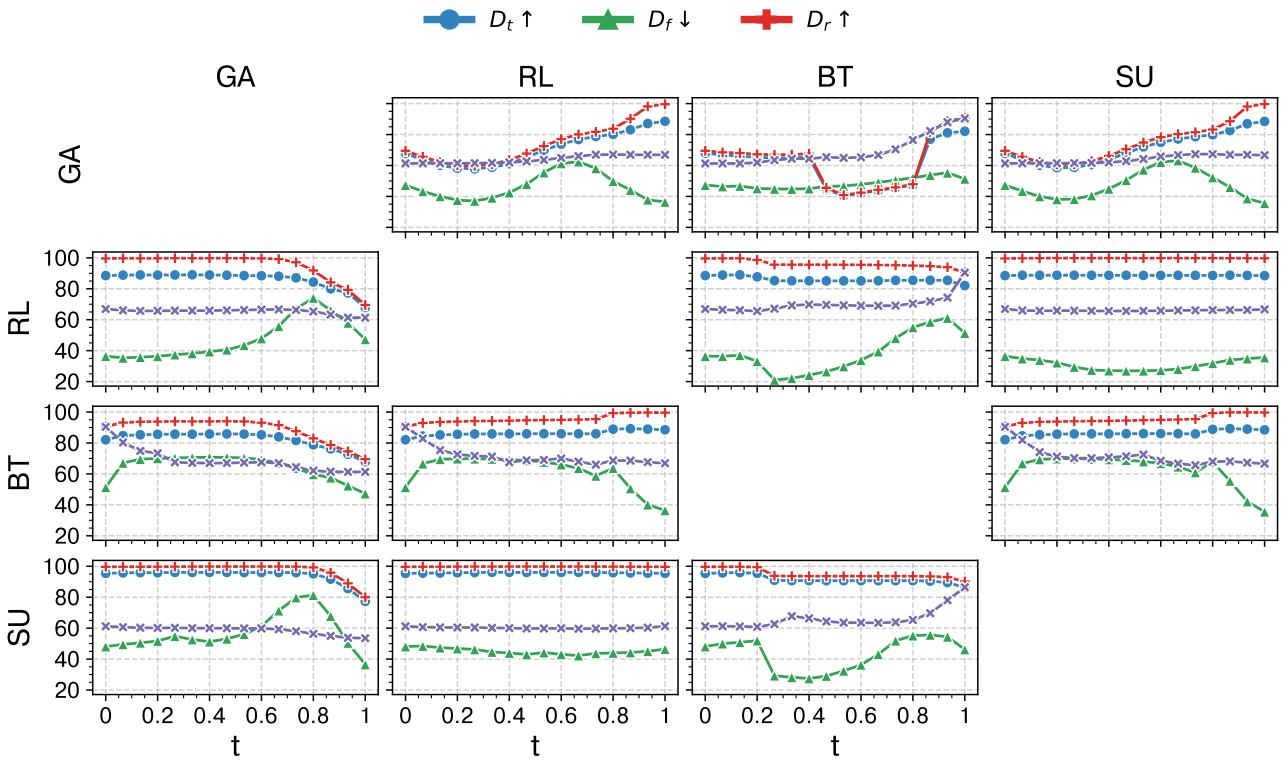
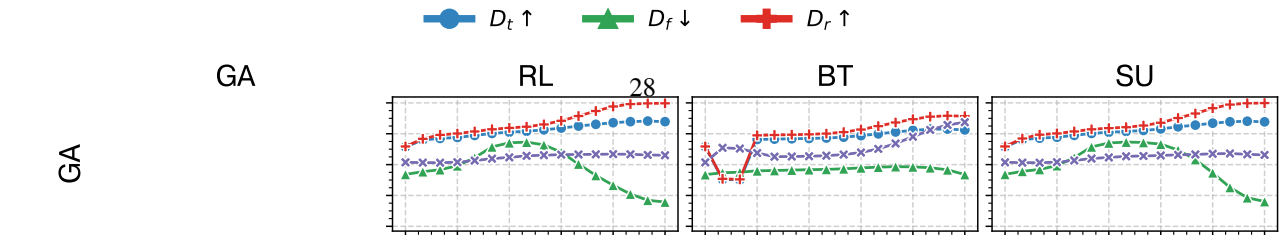


Figure 86: Quadratic MCU when $|D_f| = 2.0\%$



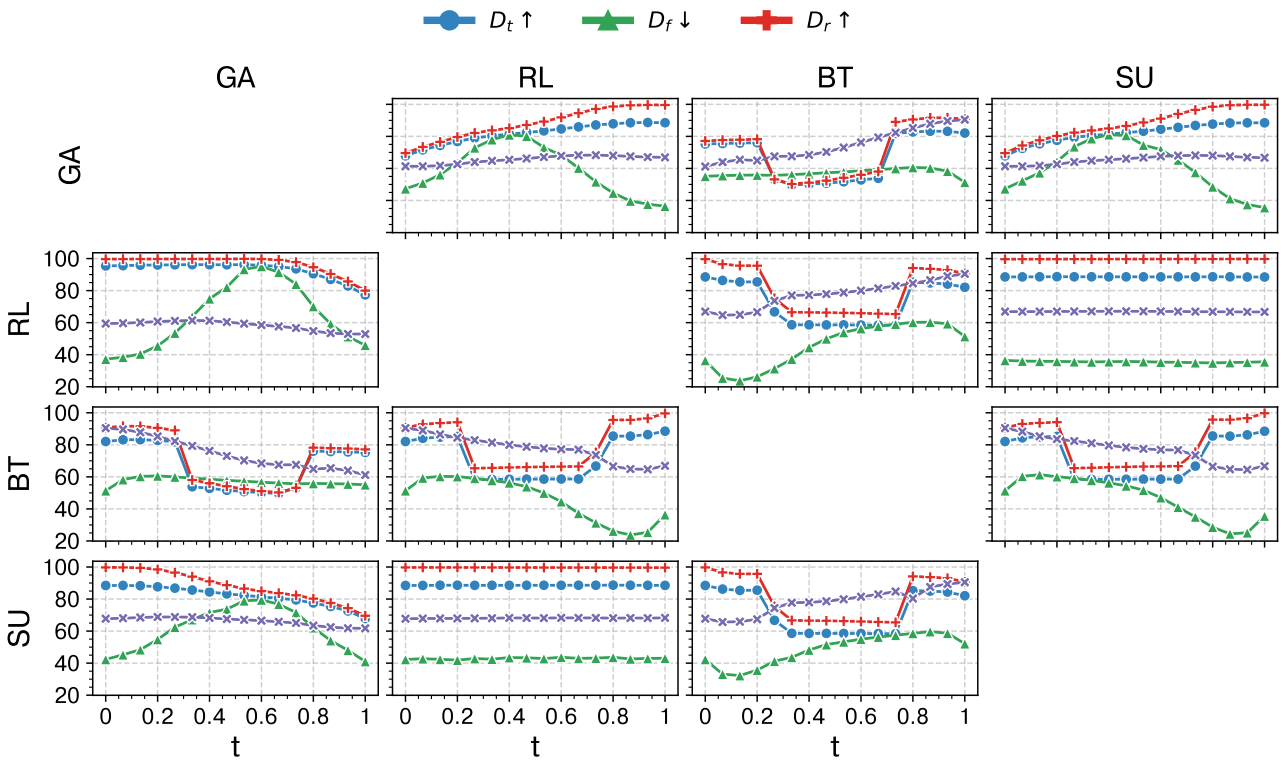


Figure 96: Linear MCU when $|D_f| = 2.0\%$

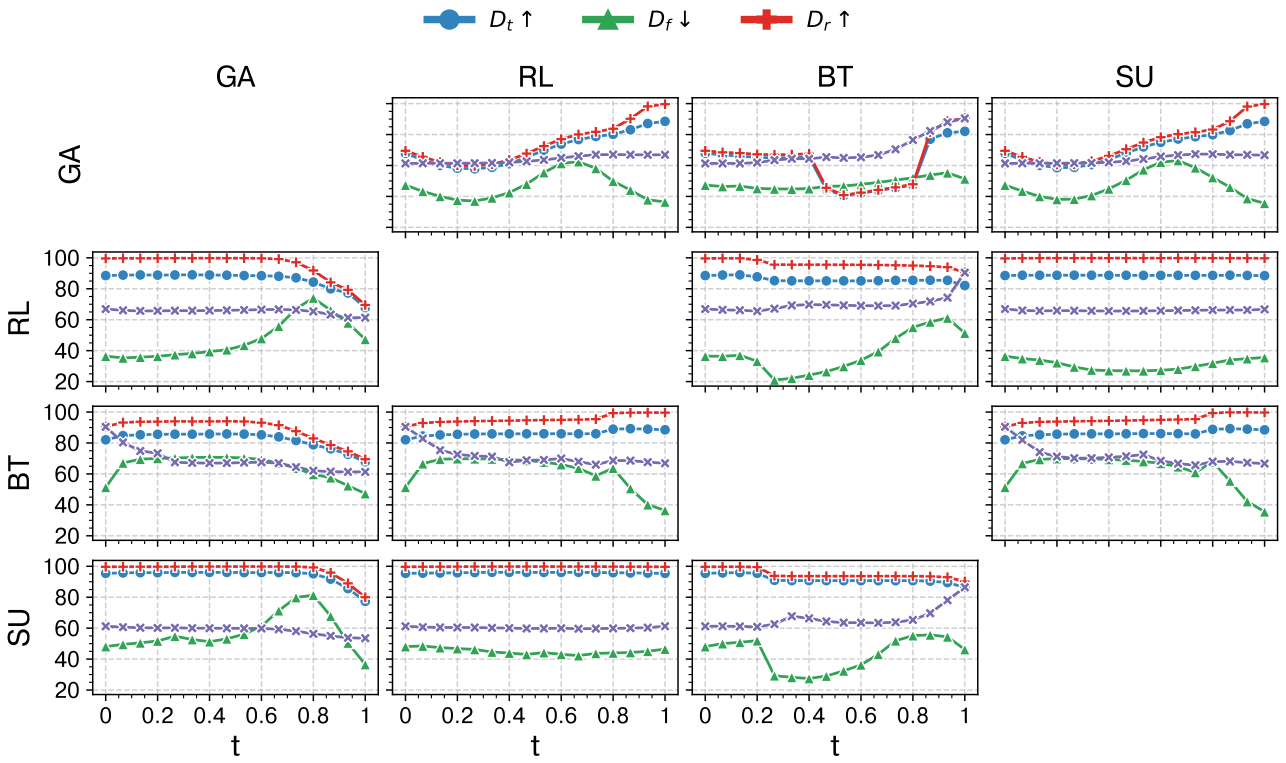


Figure 97: Quadratic MCU when $|D_f| = 2.0\%$