OPTIMIZING OVER MODE CONNECTIVITY FOR MULTI-TASK LEARNING AND MACHINE UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Mode connectivity refers to the phenomenon that independently trained neural networks can be connected by continuous low-loss paths in the parameter space, revealing a family of high-quality candidate models rather than isolated optima. While existing theoretical studies have focused exclusively on the single-task learning setting, the role of mode connectivity in more challenging learning paradigms remains largely unexplored. In this paper, we provide the first theoretical study of *mode connectivity-based multi-task learning and machine unlearning*. We consider overparameterized fully connected multilayer perceptrons (MLPs) trained from a shared initialization in the *Neural Tangent Kernel* (NTK) regime, which yields closed-form expressions for the loss along a path up to a bounded error. We then extend these results to finite-width nonlinear networks. We establish theoretical conditions under which quadratic Bézier curves achieve lower loss than linear interpolation, demonstrating the advantage of optimizing over nonlinear connectivity manifolds when balancing competing learning objectives. Finally, we conduct experiments on both NTK-linearized models and standard finite-width neural networks, and the results closely validate our theoretical predictions.

1 INTRODUCTION

Mode connectivity refers to the phenomenon that neural networks trained from independent runs can be connected by simple continuous paths in the parameter space while maintaining low loss along the path Garipov et al. (2018); Draxler et al. (2019). Mode connectivity has become a useful tool for understanding and manipulating trained models, with applications including ensembling, robustness, continual learning, and model merging. It is particularly relevant to *Multi-task learning* and *machine unlearning*, as both involve understanding how models optimized for different learning objectives are related in parameter space, making the paths between these models a natural object of study. Its central challenge is to achieve sufficient forgetting while preserving performance on the retained data. In both settings, it is therefore natural to study not only individual trained models, but also the continuous parameter-space paths connecting models optimized for different objectives. Indeed, recent empirical studies have observed useful connectivity structure between multi-task and continual-learning solutions Mirzadeh et al. (2020), and have used mode connectivity to analyze and improve machine unlearning Cheng & Amiri (2026); Shi & Wang (2025).

Despite the empirical progress, the theoretical understanding of mode connectivity under *multiple or competing learning objectives* remains limited. Existing theory primarily focuses on explaining why low-loss solutions of a *single learning problem* can be connected. For example, prior work establishes connectivity under dropout- or noise-stability assumptions Kuditipudi et al. (2020), through over-parameterization and properties of SGD solutions Shevchenko & Mondelli (2020), and under assumptions on intermediate feature quality and separability Nguyen et al. (2021). Other studies investigate the role of neuron permutation symmetries and show linear connectivity after appropriate alignment in wide neural networks Entezari et al. (2022); Ferbach et al. (2024). These results provide important explanations for connectivity between solutions that optimize the same objective, but leave open a different question that arises naturally in multi-task learning and unlearning:

When and why can mode connectivity effectively balance different learning objectives, and when can nonlinear connectivity provably outperform linear interpolation?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1.1 Major Contributions

Our theoretical analysis considers over-parameterized multi-layer perceptrons trained from a shared initialization. Following the Neural Tangent Kernel (NTK) framework (Jacot et al., 2018; Cao & Gu, 2019), we analyze the networks through their first-order linearization around initialization. In the sufficiently wide regime, the resulting tangent model provides a linear representation of the network predictions, while the discrepancy from the original neural network is controlled by the linearization error. This formulation retains the geometry of parameter-space interpolation while making it possible to characterize the interaction among task-specific solutions analytically. Our main contributions are summarized as follows.

1. **A theoretical framework for mode connectivity across different learning objectives.** We formulate multi-task mode connectivity in the tangent feature space of an over-parameterized neural network. Under the NTK approximation, task-specific models approximately admit closed-form representations in terms of their tangent features and kernel matrices, allowing the geometry of connectivity to be characterized explicitly. Our framework contains over-parameterized linear regression as a special case and connects the resulting linear theory to sufficiently wide neural networks through a bounded linearization error.
2. **Mode connectivity for multi-task learning.** For two task-specific models, we derive closed-form expressions for both the pointwise and cumulative multi-task losses along linear interpolation and an optimized quadratic Bézier curve. The losses are controlled by *cross-task residuals*, which quantify how well a model trained on one task transfers to the other. We show the optimized Bézier curve achieves a lower cumulative multi-task loss than linear interpolation except in the degenerate case where both endpoint models are optimal for both tasks. We also characterize the region along the path where the Bézier curve provably attains lower pointwise loss.
3. **Mode connectivity for machine unlearning.** We study paths from an original model trained on both tasks to different right endpoints, including a forget-task model and a partially modified model. We evaluate intermediate models through their task-loss discrepancies from a retain-only reference. For the forget-task endpoint, we characterize when the best Bézier model strictly improves on the best linear model. For a general endpoint, we derive a sufficient condition for improving the cumulative squared discrepancy in terms of endpoint and control residuals.

1.2 Related Work

Related Work. On the empirical side, Garipov et al. (2018) proposes learning low-loss paths by minimizing the expected training loss along either a polygonal chain with a single bend or a quadratic Bézier curve. Tatro et al. (2020) improves curve-finding algorithms by incorporating neuron alignment based on similarities in activation distributions. Mode connectivity has also been used as a tool for studying model robustness and loss-landscape geometry Wang et al. (2023; 2024); Kim et al. (2026a;b); Zhao et al. (2020). More closely related to our setting, Zhou et al. (2025) proposes a multi-control-point spline-based path optimization method to trace task trade-offs near the Pareto front. Mirzadeh et al. (2020) empirically demonstrates linear mode connectivity in multi-taskcontinual learningrning settings and leveraged this property to mitigate catastrophic forgetting, while Hendawy et al. (2025) used mode connectivity to connect single-task and multi-task policies for efficient multi-task reinforcement learning. These studies demonstrate the empirical usefulness of connectivity across different learning objectives, whereas our work focuses on providing a theoretical characterization of such paths in multi-task learning and machine unlearning.

On the theoretical side, existing studies establish mode connectivity under various conditions, including stability, over-parameterization, feature separability, and neuron permutation symmetries Kuditipudi et al. (2020); Shevchenko & Mondelli (2020); Nguyen et al. (2021); Entezari et al. (2022); Ferbach et al. (2024). As discussed above these results primarily concern models trained for the same learning objective, whereas we study connectivity between models associated with different and potentially competing objectives.

Paper Outline Section 2 lays out the problem setups and definitions used throughout the paper. Section 3 presents theoretical results for the multi-task learning problem. Section 4 presents theoretical results for the machine unlearning problem. Section 5 presents experimental results which

validates our theories on both the linear regression model and neural network models. Source code for the experiments can be found at `https://anonymous.4open.science/r/E484/`.

2 PRELIMINARIES

The Model We consider an L-layer fully connected neural network with width m and input dimension d. Let $\theta := (W_1, \ldots, W_L)$ denote the collection of all trainable parameters, where W_ℓ is the weight matrix of layer ℓ. For an input $X \in \mathbb{R}^d$, the network output is

$$f_\theta(X) = \sqrt{m}\, W_L \sigma(W_{L-1}\sigma(\cdots \sigma(W_1 X))),$$

where $\sigma(\cdot)$ is the entry-wise activation function. Throughout the paper, we consider ReLU activation $\sigma(z) = \max\{0, z\}$. Suppose there exists a radius $\rho > 0$ such that the relevant parameter values lie in the neighborhood $\mathcal{B}(\theta^{(0)}, \rho) = \{\theta : \|\theta - \theta^{(0)}\|_2 \le \rho\}$, where $\|\cdot\|_2$ denotes the Euclidean norm. We assume all task-specific models are trained from a common random initialization $\theta^{(0)}$.

We adopt the following input assumption, which is standard in the literature Cao & Gu (2019); Du et al. (2019); Allen-Zhu et al. (2019).

Assumption.

$$0 < c_1 \le \|X\|_2 \le c_2 \qquad \text{for all relevant inputs } X,$$

where c_1, c_2 are absolute constants.

The following consequence of (Cao & Gu, 2019, Lemma 4.1) tells us that the neural network $f_\theta(X)$ can be approximated by its linearized version, with a uniform bound on the approximation error.

Lemma 2.1 (Uniform NTK Linearization Bound)**.** *For all relevant inputs X and $\theta \in \mathcal{B}(\theta^{(0)}, \rho)$, we write the linearized model as*

$$f_\theta(X) = f_{\theta^{(0)}}(X) + J_{\theta^{(0)}}(X)(\theta - \theta^{(0)}) + \varepsilon_{\mathrm{NTK}}(X, \theta),$$

where $J_{\theta^{(0)}}(X) = \big(\nabla_\theta f_\theta(X)\big)^\top\Big|_{\theta=\theta^{(0)}}$ denotes the Jacobian of the network output, and $\varepsilon_{\mathrm{NTK}}(X, \theta)$ denotes the linearization error.

Suppose $\rho \le \kappa L^{-6}(\log m)^{-3/2}$ for a sufficiently small absolute constant $\kappa > 0$. Then there exists a constant $C > 0$, depending only on c_2, such that, if the network is sufficiently wide relative to ρ, L, and the number of relevant inputs, with high probability over the random initialization, simultaneously for all relevant inputs X and $\theta \in \mathcal{B}(\theta^{(0)}, \rho)$,

$$|\varepsilon_{\mathrm{NTK}}(X, \theta)| \le C\rho^{4/3} L^3 \sqrt{m \log m}. \tag{1}$$

Linear and Quadratic Paths Let θ_L denote the left model, θ_R the right model, and $P(\alpha) : \alpha \in [0, 1]$ a parametrized path connecting the two endpoints. For any fixed $\alpha_0 \in [0, 1]$, $P(\alpha_0)$ can be viewed as a merged model between θ_L and θ_R, with the merging rule defined by P. In this paper, we study two types of paths that connect θ_L and θ_R. We use the terms *mode* and *model* interchangeably to refer to a local optimum obtained through training.

The *linear interpolation path* is given by

$$I(\alpha) = \{\theta_I(\alpha) = (1 - \alpha)\theta_L + \alpha\theta_R : \alpha \in [0, 1]\}, \tag{2}$$

The *quadratic Bézier curve* is given by

$$B(\alpha) = \Big\{\theta_B(\alpha) = (1 - \alpha)^2\theta_L + 2\alpha(1 - \alpha)\theta + \alpha^2\theta_R : \alpha \in [0, 1]\Big\}, \tag{3}$$

where θ is a control model that controls the bending of the curve.

Loss Functions We consider the standard squared loss for model training. Given a dataset $(X, Y) = (x_j, y_j)_{j=1}^n$ and the model parameter θ,

$$\mathcal{L}(\theta) := \|f_\theta(X) - Y\|^2 = \sum_{j=1}^{n} |f_\theta(x_j) - y_j|^2. \tag{4}$$

3 MULTI-TASK LEARNING

In this section, we study the *Mode Connectivity-Based Multi-Task Learning*: suppose θ_L and θ_R are trained on two distinct tasks, how well do models on the path that connects them perform on both tasks? We analyze both the linear interpolation path (equation 2) and the quadratic Bèzier curve (equation 3) by characterizing the *multi-task loss* $\mathcal{L}_+$(equation 5) along these paths in the NTK linearized model and then transferring the results to finite-width neural networks through a uniform linearization bound. The key quantities governing the loss are the cross-task residuals, which measure how well each endpoint model performs on the other task. We show that optimizing the control point of a quadratic Bézier curve induces partial cancellation of these residuals. As a result, the Bézier path achieves lower pointwise multi-task loss over a central portion of the path and lower cumulative loss when averaged over the full path. Proof details of results in this section can be found in Appendix A.

Let (X_1, Y_1) and (X_2, Y_2) be the training dataset of size n for tasks $\mathcal{T}_1$ and $\mathcal{T}_2$, respectively, and let θ_1 and θ_2 be the corresponding models trained from the common initialization $\theta^{(0)}$. Along the linear interpolation path and the quadratic Bézier curve, we evaluate the performance of a merged model using the multi-task loss $\mathcal{L}_+$, defined as the sum of its losses on the left and right endpoint tasks.

$$\mathcal{L}_+(\theta) := \mathcal{L}_1(\theta) + \mathcal{L}_2(\theta) = |f_\theta(X_1) - Y_1|^2 + |f_\theta(X_2) - Y_2|^2. \tag{5}$$

For $i \in \{1, 2\}$, let $\Phi_i := J_{\theta^{(0)}}(X_i), \widetilde{Y}_i := Y_i - F_{\theta^{(0)}}(X_i), u_i := \theta_i - \theta^{(0)}$. Then the NTK-linearized model on task $\mathcal{T}_i$ is $f_\theta^{\mathrm{NTK}}(X_i) := F_{\theta^{(0)}}(X_i) + \Phi_i(\theta - \theta^{(0)})$, for $i = 1, 2$, with residual $f_\theta^{\mathrm{NTK}}(X_i) - Y_i = \Phi_i(\theta - \theta^{(0)}) - \widetilde{Y}_i$. We assume that the endpoint models interpolate their respective tasks in the linearized model, i.e., $\Phi_i u_i = \widetilde{Y}_i$ for $i = 1, 2$, and that the stacked tangent feature matrix $\Phi = \begin{bmatrix} \Phi_1 \\ \Phi_2 \end{bmatrix}$ has full row rank.

Following existing works Garipov et al. (2018) and Ren et al. (2025), for the quadratic Bézier curve we choose the control model θ^* by minimizing the cumulative multi-task loss in the linearized model with respect to a uniform distribution over $\alpha \in [0, 1]$. Define the cross-task tangent residuals $R_{21}^{\mathrm{NTK}} := \Phi_1 u_2 - \widetilde{Y}_1$ and $R_{12}^{\mathrm{NTK}} := \Phi_2 u_1 - \widetilde{Y}_2$. Under the full-row-rank assumption, we can find the optimal control model satisfies

$$\Phi_1(\theta^* - \theta^{(0)}) - \widetilde{Y}_1 = -\frac{3}{4} R_{21}^{\mathrm{NTK}}, \qquad \Phi_2(\theta^* - \theta^{(0)}) - \widetilde{Y}_2 = -\frac{3}{4} R_{12}^{\mathrm{NTK}}. \tag{6}$$

Theorem 3.1 (Pointwise Multi-Task Loss)**.** *For any $\alpha \in [0, 1]$, the pointwise multi-task losses along the linear interpolation path and the quadratic Bézier curve satisfy*

$$\mathcal{L}_+(\theta_I(\alpha)) = (1-\alpha)^2 \|R_{12}^{\mathrm{NTK}}\|^2 + \alpha^2 \|R_{21}^{\mathrm{NTK}}\|^2 + \Delta_I(\alpha), \tag{7}$$

and

$$\mathcal{L}_+(\theta_B(\alpha)) = (1-\alpha)^2 \left(1 - \frac{5}{2}\alpha\right)^2 \|R_{12}^{\mathrm{NTK}}\|^2 + \alpha^2 \left(\frac{5\alpha - 3}{2}\right)^2 \|R_{21}^{\mathrm{NTK}}\|^2 + \Delta_B(\alpha). \tag{8}$$

And for all $\alpha \in [\frac{1}{5}, \frac{4}{5}]$,

$$\mathcal{L}_+\Big(\theta_B(\alpha)\Big) \le \mathcal{L}_+\Big(\theta_I(\alpha)\Big) \qquad \text{as } m \to \infty.$$

In Theorem 3.1, $\Delta_I(\alpha)$ and $\Delta_B(\alpha)$ denote the discrepancies between the neural-network losses and their corresponding NTK linearized approximations. Under the standard scaling $\rho = O(m^{-1/2})$, with L and the dataset size fixed and the tangent residuals bounded, equation 1 implies $\Delta_I(\alpha)$ and $\Delta_B(\alpha)$ both vanish as $m \to \infty$. Therefore, for finite-width networks, the same loss ordering holds whenever the Bézier curve's NTK-predicted advantage exceeds the combined approximation errors. Outside the interval $\alpha \in [\frac{1}{5}, \frac{4}{5}]$, there is no general ordering of the pointwise losses along the two paths. The following corollary gives the cumulative multi-task losses over the two paths.

Corollary 3.2 (Cumulative Multi-Task Loss)**.** *Along the linear interpolation path,*

$$\int_0^1 \mathcal{L}_+(\theta_I(\alpha))\, d\alpha = \frac{1}{3}\|R_{12}^{\mathrm{NTK}}\|^2 + \frac{1}{3}\|R_{21}^{\mathrm{NTK}}\|^2 + \overline{\Delta}_I. \tag{9}$$

Along the quadratic Bézier curve,

$$\int_0^1 \mathcal{L}_+(\theta_B(\alpha))\, d\alpha = \frac{1}{8}\|R_{12}^{\mathrm{NTK}}\|^2 + \frac{1}{8}\|R_{21}^{\mathrm{NTK}}\|^2 + \overline{\Delta}_B. \tag{10}$$

Consequently, under the same scaling conditions as in Theorem 3.1, as $m \to \infty$, the cumulative losses satisfy

$$\int_0^1 \mathcal{L}_+(\theta_B(\alpha))\, d\alpha \le \int_0^1 \mathcal{L}_+(\theta_I(\alpha))\, d\alpha,$$

with equality if and only if $R_{12}^{\mathrm{NTK}} = R_{21}^{\mathrm{NTK}} = 0$.

Proof. The result follows by integrating the expressions in Theorem 3.1 and using $\int_0^1 (1-\alpha)^2\, d\alpha = \int_0^1 \alpha^2\, d\alpha = \frac{1}{3}$ and $\int_0^1 (1-\alpha)^2 \left(1 - \frac{5}{2}\alpha\right)^2\, d\alpha = \int_0^1 \alpha^2 \left(\frac{5\alpha-3}{2}\right)^2\, d\alpha = \frac{1}{8}$. □

In Corollary 3.2, $\overline{\Delta}_I = \int_0^1 \Delta_I(\alpha)\, d\alpha$ and $\overline{\Delta}_B = \int_0^1 \Delta_B(\alpha)\, d\alpha$ denote the accumulated approximation errors. For finite-width networks, the same loss ordering holds whenever the NTK-predicted cumulative loss gap exceeds $|\overline{\Delta}_I| + |\overline{\Delta}_B|$.

4 Machine Unlearning

In this section, we study *mode connectivity-based machine unlearning*. Suppose there are two tasks $\mathcal{T}_1$ and $\mathcal{T}_2$. Let θ_i denote the model trained on $\mathcal{T}_i$ with training datasets (X_i, Y_i) for $i = 1, 2$. Let $\theta_{1,2}$ denote the model trained on the combined dataset $(X_1 \cup X_2, , Y_1 \cup Y_2)$. We consider both the linear interpolation path (equation 2) and the quadratic Bézier curve (equation 3) that originates from the left endpoint $\theta_{1,2}$. We designate $\mathcal{T}_1$ as the *retain task* and $\mathcal{T}_2$ as the *forget task*. We seek a model that forgets $\mathcal{T}_2$ while preserving good performance on $\mathcal{T}_1$.

Our key finding is that the quadratic Bézier curve offers a provable advantage over linear interpolation: its optimal model achieves a loss no greater than that of any model on the linear path. This advantage extends beyond the optimal model. We show that in a typical parameter regime, the Bézier curve also achieves lower cumulative loss over the entire path. We derive closed-form conditions under which these performance guarantees hold. As in Section 3, we state our results in the NTK-linearized model and then provide corresponding results for finite-width neural networks. Detailed proofs for the results in this section can be found in Appendix B.

In this section, we consider two different pre-unlearning models as choices for the right endpoint mode: ❶ θ_2, a model obtained by training only on the forget task ($\mathcal{T}_2$). Here θ_2 serves as a rudimentary pre-unlearning model, which represents an unfavorable scenario in the machine unlearning setting: the model learns the forget task without any training on the retain task. ❷ $\bar{\theta}_{1',2'}$, a model obtained by training on partial data from both the forget and retain tasks. We view $\bar{\theta}_{1',2'}$ as a simplified proxy for a pre-unlearning model produced by common unlearning methods such as gradient ascent Thudi et al. (2022); Kurmanji et al. (2023). This setting is intended to represent a model whose performance on the forget task has significantly diminished compared to the original model $\theta_{1,2}$, while its performance on the retain task is also slightly impaired.

We define two loss functions for this section. The *unlearning loss* is defined as

$$\mathcal{L}_-(\theta) := \mathcal{L}_1(\theta) - \mathcal{L}_2(\theta). \tag{11}$$

Minimizing $\mathcal{L}_-$ favors a higher loss on the forget task ($\mathcal{T}_2$) and a lower loss on the retain task ($\mathcal{T}_1$). On the quadratic Bézier curve, we find the control model θ^* by minimizing a regularized version of $\mathcal{L}_-$, so that the choice of θ^* is aligned with this direct unlearning objective.

To evaluate a model's unlearning performance, we use a slightly different objective, the *retain-task similarity loss*. The end goal after machine unlearning is to obtain a model that performs as if it

had never been trained on the forget task, and the retain-only model θ_1 serves as such a reference. We therefore use $\mathcal{L}_{\simeq 1}$ as a similarity measure for assessing how closely a model approximates the behavior of θ_1. Specifically, for a model θ, $\mathcal{L}_{\simeq 1}(\theta)$ measures the discrepancy between its task losses and those of θ_1:

$$\mathcal{L}_{\simeq 1}(\theta) := \Big|\mathcal{L}_1(\theta) - \mathcal{L}_1(\theta_1)\Big| + \Big|\mathcal{L}_2(\theta) - \mathcal{L}_2(\theta_1)\Big|. \tag{12}$$

We emphasize that $\mathcal{L}_{\simeq 1}$ is not intended as a standalone machine unlearning benchmark that separately evaluates retain utility and forgetting efficacy. Rather, it serves as a path-wise similarity measure, consistent with the goal of recovering the behavior of the retain-only model.

4.1 Setting 1: Right Endpoint Model is θ_2.

In this setting, the Bézier curve control point θ^* is obtained by minimizing the regularized cumulative unlearning loss, $\int_0^1 \big(\lambda\mathcal{L}_1(\theta_B(\alpha)) - \mathcal{L}_2(\theta_B(\alpha))\big)\, d\alpha + \|\theta - \theta^{(0)}\|^2$. We assume that this regularization is sufficient to make the objective strictly convex, ensuring a unique minimizer.

Theorem 4.1 compares the optimal merged models along the two paths.

Theorem 4.1 (Optimal Unlearning Model Comparison)**.** *In the NTK-linearized model, the retain-task similarity loss along the linear interpolation path is minimized at $\alpha = 0$, giving*

$$\theta_I^{\mathrm{OPT}} = \theta_{1,2}, \qquad \mathcal{L}_{\simeq 1}(\theta_I^{\mathrm{OPT}}) = \|R_{12}^{\mathrm{NTK}}\|^2.$$

An optimal model on the quadratic Bézier curve appears at

$$\alpha_B^{\mathrm{OPT}} \in \arg\min_{\alpha\in[0,1]} \big\{ \|2\alpha(1-\alpha)R_1^* + \alpha^2 R_{21}^{\mathrm{NTK}}\|^2 + \big|4\alpha^2(1-\alpha)^2\|R_2^*\|^2 - \|R_{12}^{\mathrm{NTK}}\|^2\big| \big\},$$

which implies θ_B^{OPT} is either the left endpoint or an interior point.

Consequently,

$$\mathcal{L}_{\simeq 1}(\theta_B^{\mathrm{OPT}}) \le \mathcal{L}_{\simeq 1}(\theta_I^{\mathrm{OPT}}) = \|R_{12}^{\mathrm{NTK}}\|^2.$$

Strict improvement occurs if and only if some $\alpha \in (0,1)$ yields a retain-task similarity loss below $\|R_{12}^{\mathrm{NTK}}\|^2$.

In this theorem, $R_i^* := \Phi_i(\theta^* - \theta^{(0)}) - \widetilde{Y}_i$ denotes the NTK-linearized residual of the Bézier control model θ^* on task $\mathcal{T}_i$, for $i = 1, 2$. Corollary 4.2 gives the condition under which θ_B^{OPT} achieves a strictly smaller retain-task similarity loss than θ_I^{OPT}.

Corollary 4.2. *In the infinite-width NTK limit, $\mathcal{L}_{\simeq 1}\left(\theta_B^{\mathrm{OPT}}\right) < \mathcal{L}_{\simeq 1}\left(\theta_I^{\mathrm{OPT}}\right)$ if and only if there exists some $\alpha \in [0,1]$ such that*

$$\mathcal{L}_1^{\mathrm{NTK}}\big(\theta_B(\alpha)\big) < \mathcal{L}_2^{\mathrm{NTK}}\big(\theta_B(\alpha)\big) < 2\mathcal{L}_2^{\mathrm{NTK}}(\theta_1) - \mathcal{L}_1^{\mathrm{NTK}}\big(\theta_B(\alpha)\big). \tag{13}$$

Condition equation 13 requires the loss on the forget task to be larger than that on the retain task, while remaining sufficiently close to the forget-task loss of the retain-only model θ_1. Thus, Bézier curve improves over the optimal linear interpolation model when it contains a model that sufficiently forgets $\mathcal{T}_2$ without deviating too far from the behavior of θ_1. For a finite-width network, the same comparison holds when the NTK improvement margin dominates the linearization errors.

4.2 Setting 2: Right Endpoint Model is $\bar{\theta}_{1',2'}$.

In this setting, we take the right endpoint to be a model trained on partial data from both the retain and forget tasks. This way, the right endpoint retains information from both tasks, making it a natural starting point for further unlearning through model merging. To give a theoretical characterization of such a pre-unlearning model $\bar{\theta}_{1',2'}$ is not trivial. We first introduce the *random-subset model* $\theta_{1',2'}$, then use the same sampling fractions to construct a deterministic right endpoint $\bar{\theta}_{1',2'}$ through a regularized unlearning objective.

The random-subset model $\theta_{1',2'}$. For $i = 1, 2$, let $S_i \subseteq [n]$ be subsets of size $|S_i| = \lfloor \lambda_i n \rfloor$ sampled uniformly at random, and $\lambda_i \in [0, 1]$ denotes the fractions of sampled data. Our regime of interest is $\lambda_1 \simeq 1, \lambda_2 \simeq 0$, corresponding to retaining most of the retain-task data while reducing the contribution of the forget-task data. Let $M_i \in \mathbb{R}^{n \times |S_i|}$ be the column-selection matrices corresponding to the subsets S_i. In the NTK linearized model, the selected tangent features and shifted labels are $M_i^\top \Phi_i$ and $M_i^\top \widetilde{Y}_i$. Then $\theta_{1',2'} = \theta^{(0)} + \left[M_1^\top \Phi_1 \; M_2^\top \Phi_2\right]^\dagger \left[M_1^\top \widetilde{Y}_1 \; M_2^\top \widetilde{Y}_2\right]$. Notice that $\theta_{1',2'}$ depends on the particular random subsets S_1 and S_2, and carrying the random selection matrices M_1, M_2 explicitly makes the subsequent analysis intractable. Motivated by the same sampling fractions, we introduce a deterministic, subset-weighted unlearning model $\bar{\theta}_{1',2'}$, defined through a regularized unlearning objective rather than subset retraining.

Right endpoint model $\bar{\theta}_{1',2'}$ We obtain a deterministic right endpoint $\bar{\theta}_{1',2'}$ by minimizing the expected regularized unlearning objective $\mathcal{L}_-^{\mathrm{Reg}}(\theta) = \mathcal{L}_1(\theta) - \mathcal{L}_2(\theta) + \|\theta\|^2$. We assume $\lambda_1 \Phi_1^\top \Phi_1 - \lambda_2 \Phi_2^\top \Phi_2 + I \succ 0$ under such regularization.

The closed-form expression for $\bar{\theta}_{1',2'}$ can be solved to be

$$\bar{\theta}_{1',2'} = \theta^{(0)} + \left(\lambda_1 \Phi_1^\top \Phi_1 - \lambda_2 \Phi_2^\top \Phi_2 + I\right)^{-1} \left(\lambda_1 \Phi_1^\top \widetilde{Y}_1 - \lambda_2 \Phi_2^\top \widetilde{Y}_2\right). \tag{14}$$

Let $R_i := \Phi_i \left(\bar{\theta}_{1',2'} - \theta^{(0)}\right) - \widetilde{Y}_i, i = 1, 2$ be the residual of $\bar{\theta}_{1',2'}$ on the two tasks. Theorem 4.3 provides a sufficient condition for comparing a squared surrogate of the cumulative retain-task similarity loss: $\widetilde{\mathcal{L}}_{\simeq 1}(\theta) := \left(\mathcal{L}_1(\theta) - \mathcal{L}_1(\theta_1)\right)^2 + \left(\mathcal{L}_2(\theta) - \mathcal{L}_2(\theta_1)\right)^2$. This surrogate is introduced because it admits a closed-form integral, whereas the original absolute-value objective is less tractable.

Theorem 4.3 (Cumulative Squared Retain-Task Similarity Loss Comparison)**.** *The quadratic Bézier curve achieves a strictly smaller cumulative squared retain-task similarity loss than the linear interpolation path such that $\int_0^1 \widetilde{\mathcal{L}}_{\simeq 1}(\theta_B(\alpha))\, d\alpha < \int_0^1 \widetilde{\mathcal{L}}_{\simeq 1}(\theta_I(\alpha))\, d\alpha$, provided that*

$$\begin{aligned} &\|R_{12}^{\mathrm{NTK}}\|^2 \Big(7\|R_2^*\|^2 - \|R_2\|^2 - 3\|R_2^* - R_2\|^2\Big) \\ &\qquad > \frac{1}{84} \sum_{i=1}^{2} \Big(73\|R_i\|^4 + 187\|R_i^*\|^4\Big) + 15\Big(|\overline{\Delta}_I| + |\overline{\Delta}_B|\Big). \end{aligned} \tag{15}$$

$\overline{\Delta}_I$ and $\overline{\Delta}_B$ are the accumulated finite-width approximation errors along the linear and quadratic Bézier paths, respectively. Under the same scaling assumptions as in Theorem 3.1, the approximation errors vanish as $m \to \infty$.

As a simple representative regime, suppose $R_1 \approx R_1^* \approx 0$, and $R_2^* \approx R_2$. Then condition equation 15 reduces approximately to $65\|R_2\|^4 - 126\|R_{12}^{\mathrm{NTK}}\|^2\|R_2\|^2 < 0$, which holds whenever $0 < \|R_2\|^2 < \frac{126}{65}\|R_{12}^{\mathrm{NTK}}\|^2$. Thus, when the retain-task residuals are small and the control point and right endpoint behave similarly on the forget task, the quadratic Bézier curve can outperform linear interpolation under a mild condition on the forget-task residual.

5 Experiments

5.1 Linear Models

We conduct experiments on over-parameterized linear regression models as a computationally simple baseline for the linearized setting underlying our infinite-width NTK analysis. These experiments allow us to examine the theoretical predictions directly, without finite-width approximation errors.

We consider two tasks with synthetic Gaussian data under different levels of feature and task alignment. For multi-task learning, we compare the cumulative multi-task losses along the linear interpolation path and the quadratic Bézier curve. For machine unlearning, we compare the optimal models and cumulative retain-task similarity losses in both right-endpoint settings. Detailed experimental setups and all figures are provided in Appendix C.

The results agree closely with our theoretical predictions. For multi-task learning, the cumulative losses closely follow the theoretical coefficients of $1/3$ for linear interpolation and $1/8$ for the quadratic Bézier curve. Both losses are essentially zero when the features and tasks are perfectly aligned. For machine unlearning in Setting 1, the minimum along the linear path is attained at the left endpoint, while the optimal Bézier model lies either at the left endpoint or in the interior, with strict improvement whenever the theoretical condition holds. In Setting 2, the Bézier curve achieves a smaller cumulative retain-task similarity loss in every trial, and the sufficient condition for the squared surrogate is also satisfied. Together, these results support our theoretical comparisons of the two merging paths.

5.2 Neural Network Models

Multi-Task Learning We further evaluate the multi-task behavior of mode connectivity using PreResNet110 on CIFAR-10 and a pretrained ViT-Tiny on ImageNet-100. For both datasets, we consider three class-partition settings, 9:1, 8:2, and 5:5, where the two endpoint models are trained on disjoint subsets of classes according to the specified ratio. Each endpoint therefore specializes in a different task, while the quadratic Bézier curve is optimized over the union of the two subsets to integrate the knowledge learned by both endpoints. As shown in Fig. 1, the optimized Bézier paths substantially outperform direct linear interpolation across all partition settings and both model–dataset combinations.

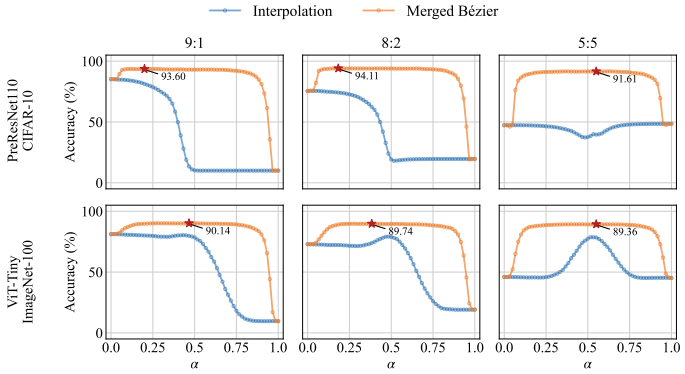

Figure 1: Multi-task learning results under the 9:1, 8:2, and 5:5 class-partition settings. The top row shows results for PreResNet110 on CIFAR-10, while the bottom row shows results for pretrained ViT-Tiny on ImageNet-100. Each panel compares the test accuracy along the optimized quadratic Bézier path with direct linear interpolation between the two endpoint models. The red star marks the best-performing point on the Bézier path.

On CIFAR-10, the Bézier paths achieve peak accuracies of 93.60%, 94.11%, and 91.61% for the 9:1, 8:2, and 5:5 settings, respectively, compared with 85.21%, 75.62%, and 48.53% for linear interpolation. The same behavior is observed on ImageNet-100, where the corresponding Bézier peak accuracies are 90.14%, 89.74%, and 89.36%, compared with 81.24%, 79.18%, and 78.68% for linear interpolation. More importantly, the improvement is sustained over broad regions of the trajectory rather than being confined to a single interpolation point, indicating that the learned control point effectively integrates knowledge from the two task-specific endpoints.

These empirical results are consistent with the theoretical analysis in Section 3. Theorem 3.1 establishes that the quadratic Bézier curve has no greater pointwise multi-task loss than linear interpolation for $\alpha \in [1/5, 4/5]$. Consistent with this prediction, the best-performing points on the learned Bézier curves all lie within this theoretically favorable interval, as indicated by the red stars in Fig. 1. Moreover, Corollary 3.2 predicts a lower cumulative multi-task loss for the quadratic Bézier curve over the entire trajectory. This trajectory-level advantage is also reflected empirically: the integrated accuracies of the Bézier paths on CIFAR-10 are 82.48, 83.72, and 80.58 for the 9:1, 8:2, and 5:5 settings, substantially exceeding 38.65, 42.63, and 45.42 for linear interpolation. The same trend holds on ImageNet-100, where the corresponding integrated accuracies are 79.64, 79.09, and 79.17 for the Bézier paths versus 55.92, 55.79, and 54.16 for linear interpolation. Together, these results show that the empirical behavior of deep neural networks is consistent with both the pointwise and cumulative advantages predicted by our theoretical analysis.

Machine Unlearning The unlearning experimental settings on the neural network are identical to the setup described for multi-task learning. The results on CIFAR-10 with PreResNet110 and

Table 1: The accuracy results of $\mathcal{T}_1$ and $\mathcal{T}_2$ at control points under different class partition settings for the unlearning task.

Class Partition	1:9		2:8		5:5	
Tasks	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_1$	$\mathcal{T}_2$
CIFAR-10	32.44%	0.00%	25.38%	0.00%	27.43%	0.00%
ImageNet-100	45.39%	0.00%	33.71%	0.00%	27.77%	0.00%

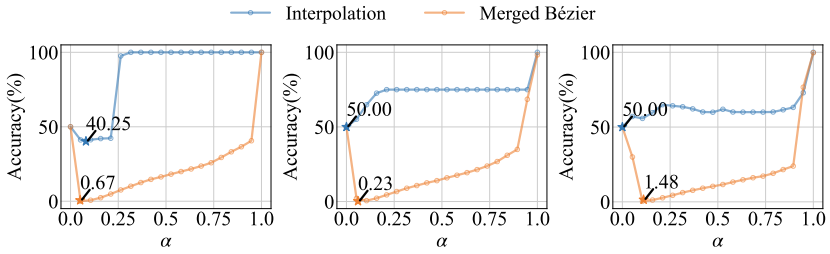

Figure 2: Unlearning results under the 9:1, 8:2 and 5:5 class partition setting on CIFAR-10. The star ($\star$) denotes the optimal unlearning models $\theta^{\mathrm{OPT}}_{\mathrm{interp}}$ or $\theta^{\mathrm{OPT}}_{\mathrm{merged}}$ on the Bézier curve, and the annotated values represent the average accuracy gap of $\mathcal{T}_1$ and $\mathcal{T}_2$ between this optimal model and the baseline model θ_1.

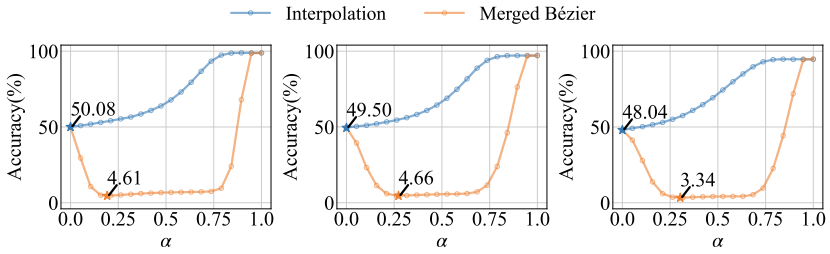

Figure 3: Unlearning results under the 9:1, 8:2 and 5:5 class partition setting on ImageNet-100. The figures adopt the same format as Figure 2.

ImageNet-100 with ViT can be found in Figure 2 and 3, respectively, where the lines represent the average accuracy gap between the retain and forget accuracy. Experimental results across forgetting classes 1, 2, and 5 consistently demonstrate that the Bézier curve achieves a smaller average accuracy gap than linear interpolation, thereby yielding superior unlearning performance. Furthermore, the accuracy results of $\mathcal{T}_1$ and $\mathcal{T}_2$ at the control point can be found in Table 1. The accuracy of $\mathcal{T}_1$ at the control point consistently outperforms that of $\mathcal{T}_2$, which aligns perfectly with our theoretical conclusions presented in Section 4.1.

6 Conclusion

We analyzed linear interpolation and quadratic Bézier paths for multi-task learning and machine unlearning in a shared tangent feature space. For multi-task learning, optimizing the control point yields explicit residual cancellation, a pointwise loss comparison on a central interval, and an exact reduction in cumulative tangent-model loss. For machine unlearning, we characterized conditions under which curvature improves agreement with a retain-only reference, measured through task-loss discrepancies. These comparisons transfer to finite-width networks when the improvement exceeds the linearization error. The nonlinear network experiments show the practical usefulness of optimized Bézier paths, while direct validation of the tangent-model predictions remains an important experimental requirement.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs and providing feedback on our writing. We have not used generative AI tools for generating synthetic data sets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, designing research methodology or experiments, proposing or refining hypotheses, or interpreting results, and implementing methods, assisting with translation, cleaning and reformatting dataset, supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for sourcing/searching for information, and editing to improve readability. We have proofread LLM-generated suggestions on edits and checked LLM-generated research ideas for potential plagiarism through a manual literature survey. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization, 2019. URL https://arxiv.org/abs/1811.03962.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Jiali Cheng and Hadi Amiri. Understanding machine unlearning through the lens of mode connectivity. In *Third Conference on Language Modeling*, 2026. URL https://openreview.net/forum?id=wIeRDYrSeg.

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred A. Hamprecht. Essentially no barriers in neural network energy landscape, 2019. URL https://arxiv.org/abs/1803.00885.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks, 2019. URL https://arxiv.org/abs/1810.02054.

Rahim Entezari, Hanie Sedghi, Olga Saukh, and Behnam Neyshabur. The role of permutation invariance in linear mode connectivity of neural networks, 2022. URL https://arxiv.org/abs/2110.06296.

Damien Ferbach, Baptiste Goujaud, Gauthier Gidel, and Aymeric Dieuleveut. Proving linear mode connectivity of neural networks via optimal transport, 2024. URL https://arxiv.org/abs/2310.19103.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Ahmed Hendawy, Henrik Metternich, Jan Peters, Gabriele Tiboni, and Carlo D'Eramo. It is all connected: Multi-task reinforcement learning via mode connectivity. In *Eighteenth European Workshop on Reinforcement Learning*, 2025. URL https://openreview.net/forum?id=WKfhk8wizf.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Hyo Seo Kim, Gang Luo, Can Chen, Binghui Wang, Yue Duan, and Ren Wang. Geometry-aware crossover for effective and efficient evolutionary attacks. In *ICLR 2026 Workshop on Principled Design for Trustworthy AI - Interpretability, Robustness, and Safety across Modalities*, 2026a. URL https://openreview.net/forum?id=XUgVnSMdjp.

Hyo Seo Kim, Gang Luo, Can Chen, Binghui Wang, Yue Duan, and Ren Wang. Moco-ea: Exploiting adversarial mode connectivity for efficient evolutionary attacks, 2026b. URL https://arxiv.org/abs/2605.18919.

Rohith Kuditipudi, Xiang Wang, Holden Lee, Yi Zhang, Zhiyuan Li, Wei Hu, Sanjeev Arora, and Rong Ge. Explaining landscape connectivity of low-cost solutions for multilayer nets, 2020. URL https://arxiv.org/abs/1906.06247.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning, 2023. URL https://arxiv.org/abs/2302.09880.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. Linear mode connectivity in multitask and continual learning, 2020. URL https://arxiv.org/abs/2010.04495.

Quynh N Nguyen, Pierre Bréchet, and Marco Mondelli. When are solutions connected in deep networks? In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 20956–20969. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/af5baf594e9197b43c9f26f17b205e5b-Paper.pdf.

Jie Ren, Pin-Yu Chen, and Ren Wang. Revisiting mode connectivity in neural networks with bezier surface. In *International Conference on Learning Representations*, 2025.

Alexander Shevchenko and Marco Mondelli. Landscape connectivity and dropout stability of sgd solutions for over-parameterized neural networks, 2020. URL https://arxiv.org/abs/1912.10095.

Yingdan Shi and Ren Wang. Mcu: Improving machine unlearning through mode connectivity. *arXiv e-prints*, pp. arXiv–2505, 2025.

N. Joseph Tatro, Pin-Yu Chen, Payel Das, Igor Melnyk, Prasanna Sattigeri, and Rongjie Lai. Optimizing mode connectivity via neuron alignment, 2020. URL https://arxiv.org/abs/2009.02439.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning, 2022. URL https://arxiv.org/abs/2109.13398.

Ren Wang, Yuxuan Li, and Sijia Liu. Exploring diversified adversarial robustness in neural networks via robust mode connectivity. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2346–2352, 2023. doi: 10.1109/CVPRW59228.2023.00231.

Ren Wang, Yuxuan Li, and Alfred Hero. Deep adversarial defense against multilevel-lp attacks, 2024. URL https://arxiv.org/abs/2407.09251.

Pu Zhao, Pin-Yu Chen, Payel Das, Karthikeyan Natesan Ramamurthy, and Xue Lin. Bridging mode connectivity in loss landscapes and adversarial robustness, 2020. URL https://arxiv.org/abs/2005.00060.

Zhipeng Zhou, Ziqiao Meng, Pengcheng Wu, Peilin Zhao, and Chunyan Miao. Exploring tradeoffs through mode connectivity for multi-task learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 81268–81285. Curran Associates, Inc., 2025. doi: 10.52202/085713-2715. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/74b8ccb5ad21871fd436867df19c35ac-Paper-Conference.pdf.

A Section 3 Proofs

A.1 Derivation of equation 6

For a candidate control model θ, the NTK-linearized $\mathcal{T}_1$ residual at the control model is $\Phi_1(\theta - \theta^{(0)}) - \widetilde{Y}_1$. The $\mathcal{T}_1$ residuals at the left and right endpoints are $\Phi_1 u_1 - \widetilde{Y}_1 = 0$ and $\Phi_1 u_2 - \widetilde{Y}_1 =$

R_{21}^{NTK}, respectively. Since the linearized residual is affine in the parameters and the Bézier weights sum to one, the $\mathcal{T}_1$ residual along the curve is $2\alpha(1-\alpha)\big(\Phi_1(\theta-\theta^{(0)})-\widetilde{Y}_1\big)+\alpha^2 R_{21}^{\mathrm{NTK}}$.

Integrating the squared norm of this $\mathcal{T}_1$ residual over $\alpha\in[0,1]$ and completing the square gives the cumulative $\mathcal{T}_1$ loss $\frac{2}{15}\big\|\Phi_1(\theta-\theta^{(0)})-\widetilde{Y}_1+\frac{3}{4}R_{21}^{\mathrm{NTK}}\big\|^2+\frac{1}{8}\|R_{21}^{\mathrm{NTK}}\|^2$. Similarly, the cumulative $\mathcal{T}_2$ loss is $\frac{2}{15}\big\|\Phi_2(\theta-\theta^{(0)})-\widetilde{Y}_2+\frac{3}{4}R_{12}^{\mathrm{NTK}}\big\|^2+\frac{1}{8}\|R_{12}^{\mathrm{NTK}}\|^2$.

The cumulative multi-task loss is the sum of the cumulative $\mathcal{T}_1$ and $\mathcal{T}_2$ losses. Since the stacked feature matrix Φ has full row rank, the control-model residuals on the two tasks can be prescribed independently. Thus, both control-dependent squared terms can vanish simultaneously, while the remaining terms depend only on the fixed endpoint residuals. Consequently, a control model θ^* minimizes the cumulative multi-task loss if and only if $\Phi_1(\theta^*-\theta^{(0)})-\widetilde{Y}_1=-\frac{3}{4}R_{21}^{\mathrm{NTK}}$ and $\Phi_2(\theta^*-\theta^{(0)})-\widetilde{Y}_2=-\frac{3}{4}R_{12}^{\mathrm{NTK}}$, as claimed.

A.2 Proof of Theorem 3.1

Proof. By assumption, we have $\Phi_1 u_1-\widetilde{Y}_1=0$ and $\Phi_2 u_2-\widetilde{Y}_2=0$. Since the NTK-linearized residual is affine in the parameters, along the linear path, the $\mathcal{T}_1$ residual is $\Phi_1(\theta_I(\alpha)-\theta^{(0)})-\widetilde{Y}_1=\alpha R_{21}^{\mathrm{NTK}}$, and the $\mathcal{T}_2$ residual is $\Phi_2(\theta_I(\alpha)-\theta^{(0)})-\widetilde{Y}_2=(1-\alpha)R_{12}^{\mathrm{NTK}}$.

For the quadratic Bézier curve, the optimal control-point conditions in equation 6 give the $\mathcal{T}_1$ residual $-\frac{3}{2}\alpha(1-\alpha)R_{21}^{\mathrm{NTK}}+\alpha^2 R_{21}^{\mathrm{NTK}}=\alpha\frac{5\alpha-3}{2}R_{21}^{\mathrm{NTK}}$. Similarly, the $\mathcal{T}_2$ residual along the Bézier curve is $(1-\alpha)^2R_{12}^{\mathrm{NTK}}-\frac{3}{2}\alpha(1-\alpha)R_{12}^{\mathrm{NTK}}=(1-\alpha)(1-\frac{5}{2}\alpha)R_{12}^{\mathrm{NTK}}$. Summing the squared norms of the two task residuals along each path gives the expressions in equation 7 and equation 8 without the correction terms.

To obtain the nonlinear-network losses, let $\Delta_I(\alpha)$ and $\Delta_B(\alpha)$ denote the differences between the network losses and the corresponding linearized losses. The assumed inclusion of $\theta_1,\theta_2,\theta^*$ in $\mathcal{B}(\theta^{(0)},\rho)$ ensures that both paths remain in this neighborhood, since the path parameters are convex combinations of these models. By Lemma 2.1, the prediction-error vector on each task has norm at most $C\rho^{4/3}L^3\sqrt{nm\log m}$. Under $\rho=O(m^{-1/2})$ with L,n fixed, this bound is $O(m^{-1/6}\sqrt{\log m})$. Expanding the squared losses and applying Cauchy–Schwarz shows that $\Delta_I(\alpha)$ and $\Delta_B(\alpha)$ vanish uniformly over $\alpha\in[0,1]$.

Finally, for $\alpha\in[1/5,4/5]$,observe that both $|1-\frac{5}{2}\alpha|\le 1$ and $|\frac{5\alpha-3}{2}|\le 1$. Thus, each task's squared residual along the Bézier curve is no greater than the corresponding squared residual along the linear path in the NTK-linearized model. Consequently, $\mathcal{L}_+(\theta_B(\alpha))-\mathcal{L}_+(\theta_I(\alpha))\le|\Delta_B(\alpha)|+|\Delta_I(\alpha)|$. Letting $m\to\infty$ gives $\limsup_{m\to\infty}\big(\mathcal{L}_+(\theta_B(\alpha))-\mathcal{L}_+(\theta_I(\alpha))\big)\le 0$, which establishes the claimed asymptotic comparison. □

A.3 Proof of Corollary 3.2

Proof. Integrating the pointwise loss expressions in Theorem 3.1, we use $\int_0^1\alpha^2\,d\alpha=\frac{1}{3}$ and $\int_0^1\alpha^2(\frac{5\alpha-3}{2})^2\,d\alpha=\frac{1}{8}$. The corresponding coefficients of $\|R_{12}^{\mathrm{NTK}}\|^2$ have the same integrals by the substitution $\alpha\mapsto 1-\alpha$. This gives equation 9 and equation 10, with $\overline{\Delta}_I=\int_0^1\Delta_I(\alpha)\,d\alpha$ and $\overline{\Delta}_B=\int_0^1\Delta_B(\alpha)\,d\alpha$.

Subtracting the two cumulative losses yields

$$\int_0^1\big(\mathcal{L}_+(\theta_I(\alpha))-\mathcal{L}_+(\theta_B(\alpha))\big)\,d\alpha = \frac{5}{24}\big(\|R_{12}^{\mathrm{NTK}}\|^2+\|R_{21}^{\mathrm{NTK}}\|^2\big)+\overline{\Delta}_I-\overline{\Delta}_B.$$

In the NTK-linearized model, the correction terms are zero. Thus, the cumulative loss along the Bézier curve is no greater than that along the linear path, with equality if and only if $R_{12}^{\mathrm{NTK}}=R_{21}^{\mathrm{NTK}}=0$.

For the nonlinear networks, the uniform error bound established in the proof of Theorem 3.1 implies that both accumulated correction terms vanish under the stated scaling conditions. The same loss-gap identity therefore establishes the asymptotic comparison. □

B Section 4 Proofs

B.1 Proof of Theorem 4.1

Proof. We work in the NTK-linearized model. θ_1 therefore satisfies $\mathcal{L}_1(\theta_1) = 0$ and $\mathcal{L}_2(\theta_1) = \|R_{12}^{\mathrm{NTK}}\|^2$.

We first consider the linear path from $\theta_{1,2}$ to θ_2. Since the linearized residual is affine in the parameters, the residual along this path on $\mathcal{T}_1$ is $\alpha R_{21}^{\mathrm{NTK}}$. The residual on $\mathcal{T}_2$ is zero, since both endpoints interpolate $\mathcal{T}_2$. Substituting these residuals into equation 12 gives $\mathcal{L}_{\simeq 1}(\theta_I(\alpha)) = \alpha^2\|R_{21}^{\mathrm{NTK}}\|^2 + \|R_{12}^{\mathrm{NTK}}\|^2$. Thus, $\alpha = 0$ is a minimizer, giving $\theta_I^{\mathrm{OPT}} = \theta_{1,2}$ with loss $\|R_{12}^{\mathrm{NTK}}\|^2$.

For the quadratic Bézier curve, the control-model residuals R_1^* and R_2^* give

$$\Phi_1(\theta_B(\alpha) - \theta^{(0)}) - \widetilde{Y}_1 = 2\alpha(1-\alpha)R_1^* + \alpha^2 R_{21}^{\mathrm{NTK}},$$
$$\Phi_2(\theta_B(\alpha) - \theta^{(0)}) - \widetilde{Y}_2 = 2\alpha(1-\alpha)R_2^*.$$

Consequently, the retain-task similarity loss is

$$\mathcal{L}_{\simeq 1}(\theta_B(\alpha)) = \left\|2\alpha(1-\alpha)R_1^* + \alpha^2 R_{21}^{\mathrm{NTK}}\right\|^2 + \left|4\alpha^2(1-\alpha)^2\|R_2^*\|^2 - \|R_{12}^{\mathrm{NTK}}\|^2\right|.$$

This loss is continuous on $[0,1]$, so a minimizer exists. Minimizing the displayed expression gives the stated characterization of α_B^{OPT}.

Since $\theta_B(0) = \theta_{1,2} = \theta_I^{\mathrm{OPT}}$, we have $\mathcal{L}_{\simeq 1}(\theta_B^{\mathrm{OPT}}) \le \mathcal{L}_{\simeq 1}(\theta_I^{\mathrm{OPT}}) = \|R_{12}^{\mathrm{NTK}}\|^2$. At the right endpoint, $\mathcal{L}_{\simeq 1}(\theta_2) = \|R_{21}^{\mathrm{NTK}}\|^2 + \|R_{12}^{\mathrm{NTK}}\|^2$, which is no smaller than the loss at the left endpoint. Hence, an optimal Bézier model can always be chosen at the left endpoint or in the interior. Since neither endpoint gives strict improvement over θ_I^{OPT}, strict improvement occurs if and only if some $\alpha \in (0,1)$ yields a loss below $\|R_{12}^{\mathrm{NTK}}\|^2$. □

B.2 Proof of Corollary 4.2

Proof. In the NTK-linearized model, by Theorem 4.1, the optimal linear model has retain-task similarity loss $\|R_{12}^{\mathrm{NTK}}\|^2$. Thus, the optimal Bézier model achieves strict improvement if and only if some $\alpha \in [0,1]$ satisfies

$$\mathcal{L}_1^{\mathrm{NTK}}(\theta_B(\alpha)) + \left|\mathcal{L}_2^{\mathrm{NTK}}(\theta_B(\alpha)) - \mathcal{L}_2^{\mathrm{NTK}}(\theta_1)\right| < \mathcal{L}_2^{\mathrm{NTK}}(\theta_1).$$

Moving the loss on $\mathcal{T}_1$ to the right and expanding the absolute-value inequality into its lower and upper bounds gives the equivalent condition

$$\mathcal{L}_1^{\mathrm{NTK}}(\theta_B(\alpha)) < \mathcal{L}_2^{\mathrm{NTK}}(\theta_B(\alpha)) < 2\mathcal{L}_2^{\mathrm{NTK}}(\theta_1) - \mathcal{L}_1^{\mathrm{NTK}}(\theta_B(\alpha)).$$

This is precisely equation 13, establishing both necessity and sufficiency. □

B.3 Derivation of Equation (14)

We work in the NTK-linearized model. Since each subset S_i is sampled uniformly with size $\lambda_i n$, the selection matrix satisfies $\mathbb{E}[M_i M_i^\top] = \lambda_i I$. Thus, the expected squared loss on the sampled data from $\mathcal{T}_i$ is $\lambda_i\|\Phi_i(\theta - \theta^{(0)}) - \widetilde{Y}_i\|^2$. Consequently, the expected regularized unlearning objective is

$$\lambda_1\|\Phi_1(\theta - \theta^{(0)}) - \widetilde{Y}_1\|^2 - \lambda_2\|\Phi_2(\theta - \theta^{(0)}) - \widetilde{Y}_2\|^2 + \|\theta - \theta^{(0)}\|^2.$$

Differentiating this objective with respect to θ and setting the gradient to zero gives

$$\left(\lambda_1\Phi_1^\top\Phi_1 - \lambda_2\Phi_2^\top\Phi_2 + I\right)\left(\bar{\theta}_{1',2'} - \theta^{(0)}\right) = \lambda_1\Phi_1^\top\widetilde{Y}_1 - \lambda_2\Phi_2^\top\widetilde{Y}_2.$$

By the assumed positive definiteness of $\lambda_1\Phi_1^\top\Phi_1 - \lambda_2\Phi_2^\top\Phi_2 + I$, the objective is strictly convex and the coefficient matrix is invertible. Multiplying by its inverse gives equation 14 and establishes that $\bar{\theta}_{1',2'}$ is the unique minimizer.

B.4 Proof of Theorem 4.3

Proof. We first work in the NTK-linearized model. Since $\theta_{1,2}$ interpolates both tasks, the residuals on $\mathcal{T}_i$ along the linear and Bézier paths are αR_i and $2\alpha(1-\alpha)R_i^* + \alpha^2 R_i$, respectively. We know θ_1 satisfies $\mathcal{L}_1^{\mathrm{NTK}}(\theta_1) = 0$ and $\mathcal{L}_2^{\mathrm{NTK}}(\theta_1) = \|R_{12}^{\mathrm{NTK}}\|^2$. Thus, in the NTK-linearized model,

$$\begin{aligned}\widetilde{\mathcal{L}}_{\simeq 1}(\theta) = \sum_{i=1}^{2} \left\|\Phi_i(\theta - \theta^{(0)}) - \widetilde{Y}_i\right\|^4 \\ - 2\|R_{12}^{\mathrm{NTK}}\|^2 \left\|\Phi_2(\theta - \theta^{(0)}) - \widetilde{Y}_2\right\|^2 + \|R_{12}^{\mathrm{NTK}}\|^4.\end{aligned}$$

We first compare the fourth-power terms. Since $2\alpha(1-\alpha) + \alpha^2 = \alpha(2-\alpha)$, convexity of $\|\cdot\|^4$ gives, for $i = 1, 2$,

$$\begin{aligned}&\int_0^1 \left\|2\alpha(1-\alpha)R_i^* + \alpha^2 R_i\right\|^4 \, d\alpha \\ &\quad\le \int_0^1 [\alpha(2-\alpha)]^3 \left(2\alpha(1-\alpha)\|R_i^*\|^4 + \alpha^2\|R_i\|^4\right) d\alpha \\ &\quad= \frac{187}{1260}\|R_i^*\|^4 + \frac{65}{252}\|R_i\|^4.\end{aligned}$$

The corresponding integral along the linear path is $\int_0^1 \alpha^4\|R_i\|^4 \, d\alpha = \frac{1}{5}\|R_i\|^4$. Subtracting this integral leaves an upper bound of $\frac{1}{1260}\left(73\|R_i\|^4 + 187\|R_i^*\|^4\right)$ for the difference between the fourth-power terms.

For the squared residual norm on $\mathcal{T}_2$, direct integration gives

$$\begin{aligned}&\int_0^1 \left(\left\|2\alpha(1-\alpha)R_2^* + \alpha^2 R_2\right\|^2 - \alpha^2\|R_2\|^2\right) d\alpha \\ &\quad= \frac{2}{15}\|R_2^*\|^2 - \frac{2}{15}\|R_2\|^2 + \frac{1}{5}\langle R_2^*, R_2\rangle \\ &\quad= \frac{1}{30}\left(7\|R_2^*\|^2 - \|R_2\|^2 - 3\|R_2^* - R_2\|^2\right).\end{aligned}$$

To account for the nonlinear networks, let $\overline{\Delta}_I$ and $\overline{\Delta}_B$ denote the differences between the nonlinear and linearized cumulative squared similarity losses along the respective paths. These correction terms include the approximation errors in the reference losses at θ_1. Combining the preceding calculations and using $\overline{\Delta}_B - \overline{\Delta}_I \le |\overline{\Delta}_B| + |\overline{\Delta}_I|$ yields

$$\begin{aligned}&\int_0^1 \left(\widetilde{\mathcal{L}}_{\simeq 1}(\theta_B(\alpha)) - \widetilde{\mathcal{L}}_{\simeq 1}(\theta_I(\alpha))\right) d\alpha \\ &\quad\le \frac{1}{1260}\sum_{i=1}^{2}\left(73\|R_i\|^4 + 187\|R_i^*\|^4\right) \\ &\qquad - \frac{\|R_{12}^{\mathrm{NTK}}\|^2}{15}\left(7\|R_2^*\|^2 - \|R_2\|^2 - 3\|R_2^* - R_2\|^2\right) + |\overline{\Delta}_I| + |\overline{\Delta}_B|.\end{aligned}$$

By simple algebraic manipulation, we can observe that the upper bound is strictly negative, proving the claimed comparison. □

C Linear Model Experiments

In the subsequent experiments, we consider the linear model

$$Y = \theta X,$$

where $X \in \mathbb{R}^{d\times n}, Y \in \mathbb{R}^{1\times n}$, $\theta \in \mathbb{R}^{1\times d}$, and $d > n$. We take

$$n = 10, \qquad d = 100,$$

so the model is over-parameterized with $d \gg n$.

Suppose there are two tasks $\mathcal{T}_1$ and $\mathcal{T}_2$, each with training data (X_1, Y_1), (X_2, Y_2), and ground truth models θ_1^{true}, θ_2^{true}. We assume the pooled data matrix has full column rank:

$$\operatorname{rank}[X_1 \quad X_2] = 2n.$$

Let $\theta_{1,2}^{\text{true}}$ denote the ground truth model to the pooled problem, so that

$$Y_i = \theta_{1,2}^{\text{true}} X_i \quad \text{for } i = 1, 2.$$

We consider the linear interpolation path and the quadratic Bézier curve that connect the two endpoints θ_L and θ_R.

C.1 Multi-Task Learning

Let $\theta_L = \theta_1^{\text{true}}$ and $\theta_R = \theta_2^{\text{true}}$. For both the linear and quadratic paths, we compute the accumulated multitask loss

$$\int_0^1 \mathcal{L}_+(\theta(\alpha))\, d\alpha.$$

We repeat this experiment over 50 random trials. For each trial, we compute the cumulative multitask loss along both the linear interpolation path and the quadratic Bézier curve, and plot the two losses side by side.

Training Data For $\mathcal{T}_1$, we sample

$$X_1 \in \mathbb{R}^{d \times n}, \qquad (X_1)_{jk} \sim \text{Gaussian}\left(0, \frac{1}{d}\right).$$

For $\mathcal{T}_2$, we sample

$$(X_2)_{jk} = \rho\, X_1 + \sqrt{1-\rho^2}\; Z,$$

where

$$Z \sim \text{Gaussian}\left(0, \frac{1}{d} I_{d \times n}\right)$$

is with respect to X_1. With this setup, the correlation between X_1 and X_2 is

$$\begin{aligned}\operatorname{Corr}(X_1, X_2) &= \frac{\operatorname{Cov}(X_1, X_2)}{\sqrt{\operatorname{Var}(X_1)\operatorname{Var}(X_2)}} \\ &= \frac{\rho/d}{\sqrt{(1/d)(1/d)}} = \rho.\end{aligned}$$

Ground Truth Model We sample task-specific ground-truth parameters as follows. For $\mathcal{T}_1$, we sample

$$\theta_1^{\text{true}} \in \mathbb{R}^{1 \times d}, \qquad (\theta_1^{\text{true}})_j \sim \text{Gaussian}\left(0, \frac{1}{d}\right).$$

For $\mathcal{T}_2$, we sample

$$\theta_2^{\text{true}} = \tau\, \theta_1^{\text{true}} + \sqrt{1-\tau^2}\; W,$$

where

$$W \sim \text{Gaussian}\left(0, \frac{1}{d} I_d\right)$$

is independent with respect to θ_1^{true}. With this setup, the correlation between θ_1^{true} and $\theta_2 \text{true}$ is

$$\operatorname{Corr}(\theta_1^{\text{true}}, \theta_2^{\text{true}}) = \tau.$$

We define the labels by

$$Y_i = \theta_i^{\text{true}} X_i,$$

so that each task is exactly realizable.

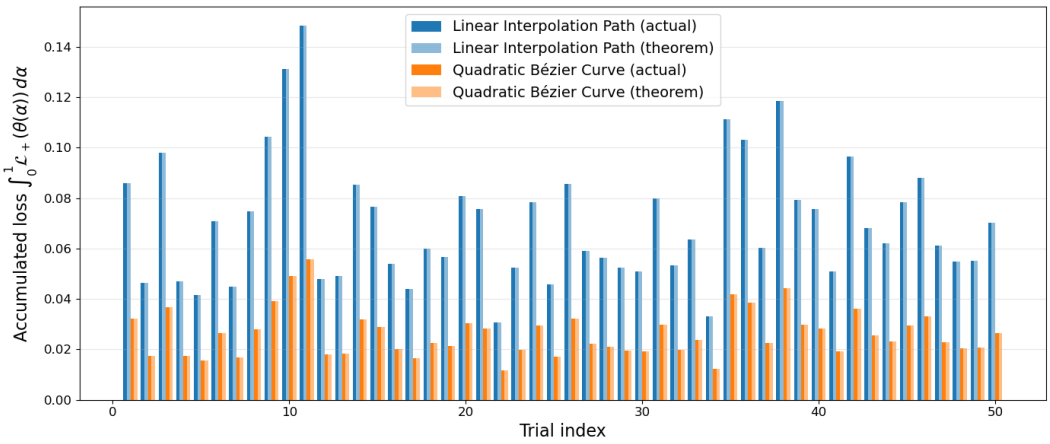

Figure 4: Multi-task learning. Feature correlation $\rho = 0$, task correlation $\tau = 0$.

Discussion The following figures present the results of our experiments under different levels of feature and task alignment.

1. Figure 4 ($\rho = 0, \tau = 0$): the feature matrices X_1 and X_2 are independent, and the ground-truth models θ_1^{true} and θ_2^{true} are independent.
2. Figure 5 ($\rho = 0$, $\tau = 1$): the feature matrices X_1 and X_2 are independent, while the ground-truth models are identical, i.e., $\theta_1^{\text{true}} = \theta_2^{\text{true}}$.
3. Figure 6 ($\rho = \tau = 1$): the feature matrices are identical, i.e., $X_1 = X_2$, and the ground-truth models are identical, i.e., $\theta_1^{\text{true}} = \theta_2^{\text{true}}$.

Comparing the three figures, we can observe that the results align with our theoretical findings in Corollary 3.2 perfectly. In particular:

1. When both the feature matrices and the underlying tasks are poorly aligned (Figure 4), both the linear interpolation path and the quadratic Bézier curve incur a nontrivial cumulative multi-task loss. Moreover, the cumulative losses closely follow the theoretical coefficients of $1/3$ for the linear interpolation path and $1/8$ for the quadratic Bézier curve predicted by Corollary 3.2.
2. When both the feature matrices and the underlying tasks are perfectly aligned (Figure 6), the cumulative multi-task loss along both paths is essentially zero, exactly as predicted by Corollary 3.2.
3. When $\rho = 0$ and $\tau = 1$ (Figure 5), the underlying regression model is shared between the two tasks, but the feature matrices are independent. Consequently, the cross-task residuals remain nonzero, resulting in a nonzero cumulative multi-task loss along both paths. The loss coefficients follow $1/3$ for the linear interpolation path and $1/8$ for the quadratic Bézier curve. This behavior is again consistent with Corollary 3.2.

C.2 Machine Unlearning

The left endpoint is $\theta_L = \theta_{1,2}$. For both the linear and quadratic paths, we compute the retain-task similarity loss

$$\mathcal{L}_{\simeq 1}(\theta) := \Big|\mathcal{L}_1(\theta) - \mathcal{L}_1(\theta_1)\Big| + \Big|\mathcal{L}_2(\theta) - \mathcal{L}_2(\theta_1)\Big|.$$

We repeat this experiment over 9 random trials. For each trial, we compute the retain-task similarity loss along both the linear interpolation path and the quadratic Bézier curve over a uniform grid of 10001 values of $\alpha \in [0, 1]$. We record the optimal value of α, the minimum $\mathcal{L}_{\simeq 1}$ achieved, and the area under the curve for both the interpolation and Bézier merged paths.

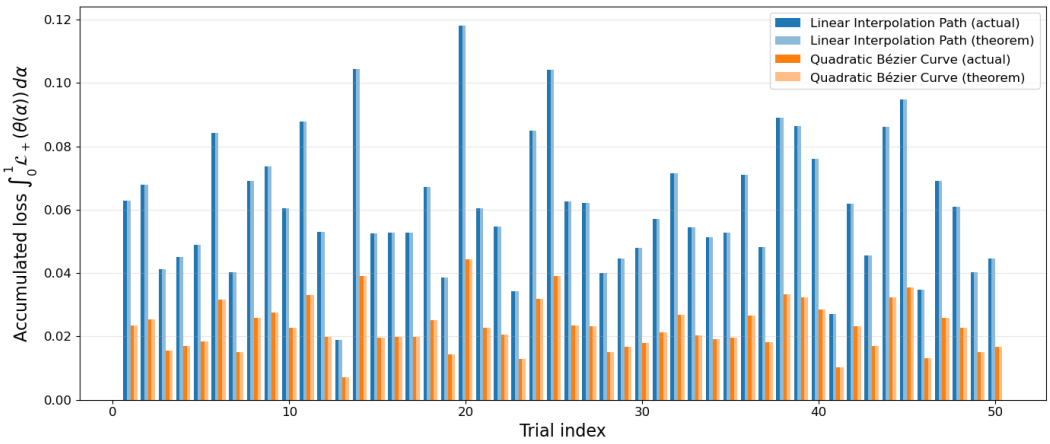

Figure 5: Multi-task learning. Feature correlation $\rho = 0$, task correlation $\tau = 1$.

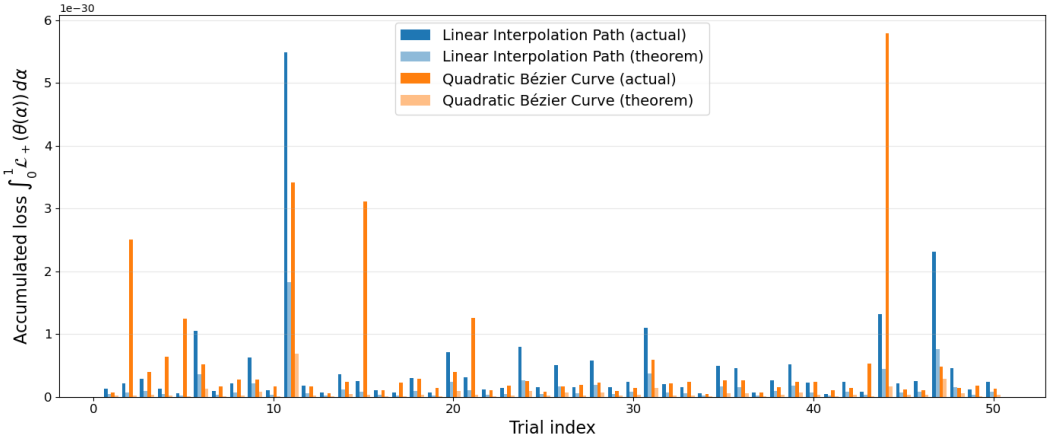

Figure 6: Multi-task learning. Feature correlation $\rho = 1$, task correlation $\tau = 1$.

In definition of the quadratic Bézier curve, we find the control point θ^* by

$$\begin{aligned}\theta^* &= \arg\min_{\theta\in\mathbb{R}^{1\times d}} \int_0^1 \mathcal{L}_-^{\text{reg}}(\theta)\,d\alpha + \|\theta\|^2 \\ &= \arg\min_{\theta\in\mathbb{R}^{1\times d}} \int_0^1 \lambda\mathcal{L}_1(\theta) - \mathcal{L}_2(\theta)\,d\alpha + \|\theta\|^2.\end{aligned}$$

In our experiments, we set the unlearning penalty coefficient to $\lambda = 0.9$.

The feature matrices X_1, X_2, ground-truth models θ_1, θ_2, and labels Y_1, Y_2 are generated in the same manner as in Section C.1. Hence, we have

$$\text{Corr}(X_1, X_2) = \rho \quad \text{and} \quad \text{Corr}(\theta_1^{\text{true}}, \theta_2^{\text{true}}) = \tau.$$

Setting 1: Right Endpoint $\theta_R = \theta_2$ Figure 7 – Figure 10 present the results of our experiments under different levels of feature and task alignment.

1. Figure 7 ($\rho = 0, \tau = 0$): the feature matrices X_1 and X_2 are independent, and the ground-truth models θ_1^{true} and θ_2^{true} are independent.
2. Figure 8 ($\rho = 0$, $\tau = 1$): the feature matrices X_1 and X_2 are independent, while the ground-truth models are identical, i.e., $\theta_1^{\text{true}} = \theta_2^{\text{true}}$.

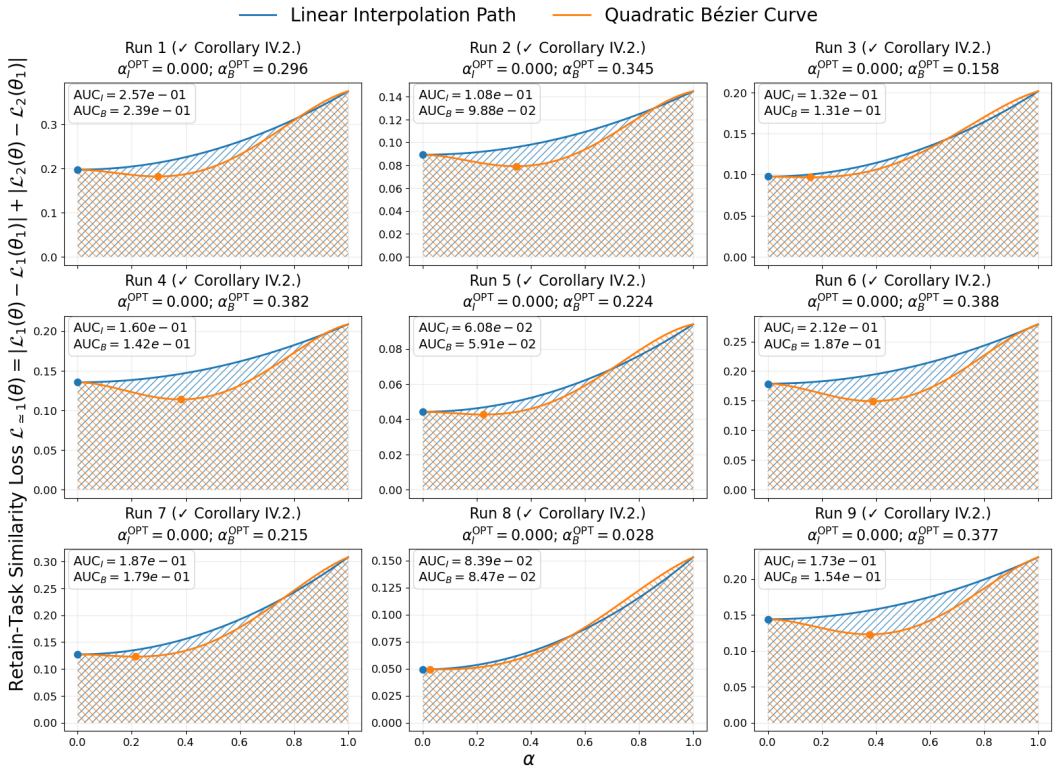

Figure 7: Machine unlearning setting 1: $\theta_R = \theta_2$. Feature correlation $\rho = 0$, task correlation $\tau = 0$.

3. Figure 9 ($\rho = 1, \tau = 0$): the feature matrices are identical, i.e., $X_1 = X_2$, while the ground-truth models θ_1^{true} and θ_2^{true} are independent.
4. Figure 10 ($\rho = \tau = 1$): both the feature matrices and the ground-truth models are identical, i.e., $X_1 = X_2$ and $\theta_1^{\text{true}} = \theta_2^{\text{true}}$.

Comparing these figures, we observe strong agreement with the theoretical predictions of Theorem 4.1 and Corollary 4.2. In particular:

1. Across Figure 7 – Figure 9, the optimal point on the linear interpolation path is always attained at $\alpha = 0$, corresponding to the pooled model $\theta_{1,2}$. In contrast, the optimal point on the quadratic Bézier curve is attained either at $\alpha = 0$ or at an interior point with $\alpha > 0$, exactly as predicted by Theorem 4.1.
2. The observed strict improvements agree with Corollary 4.2: the optimal Bézier model achieves a strictly smaller retain-task similarity loss than the optimal linear interpolation model if and only if the Bézier path contains a model satisfying equation 13.
3. Figure 10 corresponds to the degenerate case in which both the feature matrices and the underlying tasks are perfectly aligned. Consequently, the endpoints $\theta_{1,2}$ and θ_2 coincide and incur essentially zero loss on both tasks. Consequently, the retain-task similarity loss is nearly zero along both paths. The small fluctuations visible in Figure 10 are entirely due to floating-point discrepancies in the numerical optimization procedures, and should not be interpreted as contradicting the theoretical results.

Setting 2: Right Endpoint $\theta_R = \bar{\theta}_{1',2'}$ In this setting, the right endpoint $\bar{\theta}_{1',2'}$ is chosen to minimize the regularized unlearning loss in expectation, as defined in equation 14. We set the random sampling coefficients to $\lambda_1 = 0.9$ and $\lambda_2 = 0.1$, corresponding to uniformly sampling 90% of the data from the retain task ($\mathcal{T}_1$) and 10% of the data from the forget task ($\mathcal{T}_2$).

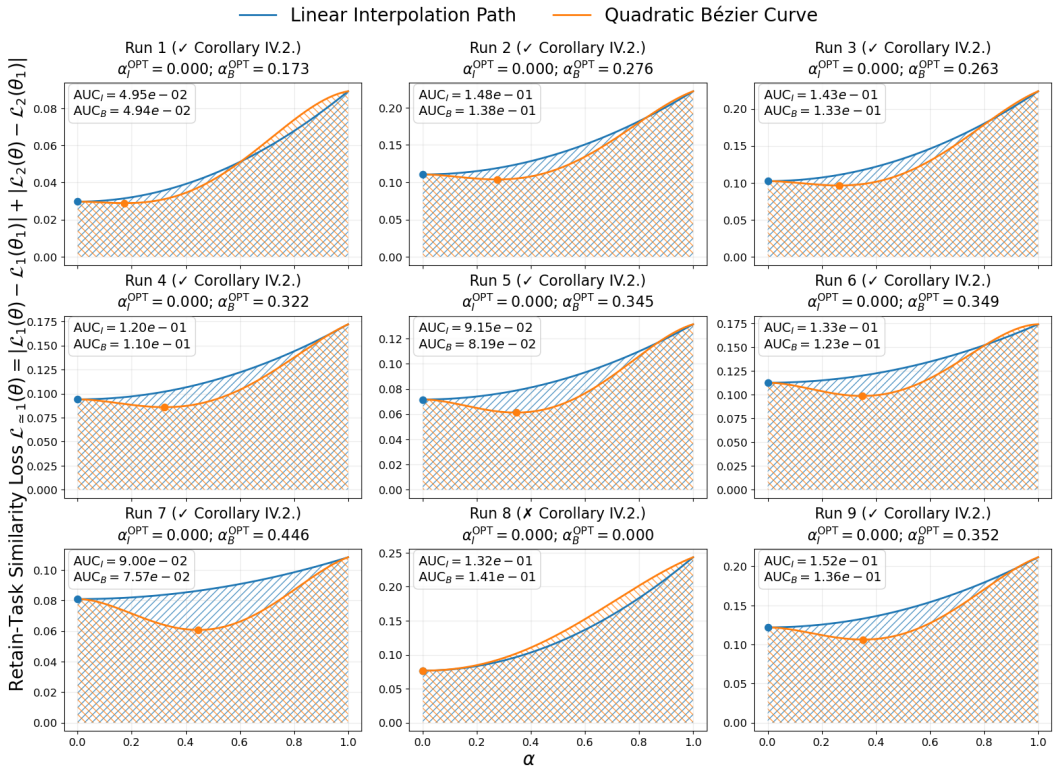

Figure 8: Machine unlearning setting 1: $\theta_R = \theta_2$. Feature correlation $\rho = 0$, task correlation $\tau = 1$.

Figure 11 – Figure 14 present the results of our experiments under different levels of feature and task alignment. We note that the curves in these figures are generally non-convex and non-smooth, since the evaluation objective $\mathcal{L}_{\simeq 1}$ is defined using absolute values.

1. Figure 11 ($\rho = 0$, $\tau = 0$): the feature matrices X_1 and X_2 are independent, and the ground-truth models θ_1^{true} and θ_2^{true} are independent.
2. Figure 12 ($\rho = 0$, $\tau = 1$): the feature matrices X_1 and X_2 are independent, while the ground-truth models are identical, i.e., $\theta_1^{\text{true}} = \theta_2^{\text{true}}$.
3. Figure 13 ($\rho = 1, \tau = 0$): the feature matrices are identical, i.e., $X_1 = X_2$, while the ground-truth models θ_1^{true} and θ_2^{true} are independent.
4. Figure 14 ($\rho = \tau = 1$): both the feature matrices and the ground-truth models are identical, i.e., $X_1 = X_2$ and $\theta_1^{\text{true}} = \theta_2^{\text{true}}$.

Comparing these figures, we observe strong agreement with the theoretical predictions of Theorem 4.3. In particular:

1. Across Figure 11 – Figure 13, the retain and forget tasks are not perfectly aligned. As shown in the figures, the cumulative retain-task similarity loss of the quadratic Bézier curve is consistently smaller than that of the linear interpolation path across all trials in these three settings. Moreover, the sufficient condition in Theorem 4.3, which is derived using the squared surrogate retain-task similarity loss $\widetilde{\mathcal{L}}_{\simeq 1}$, is satisfied for all trials. These experimental results validate Theorem 4.3, and suggest that the quadratic Bézier curve provides a more effective mode connectivity path for machine unlearning than the linear interpolation path.
2. In Figure 14, the retain and forget tasks are perfectly aligned. Consequently, the pooled model $\theta_{1,2}$ is essentially identical to the retain-task model θ_1. Therefore, it is expected that

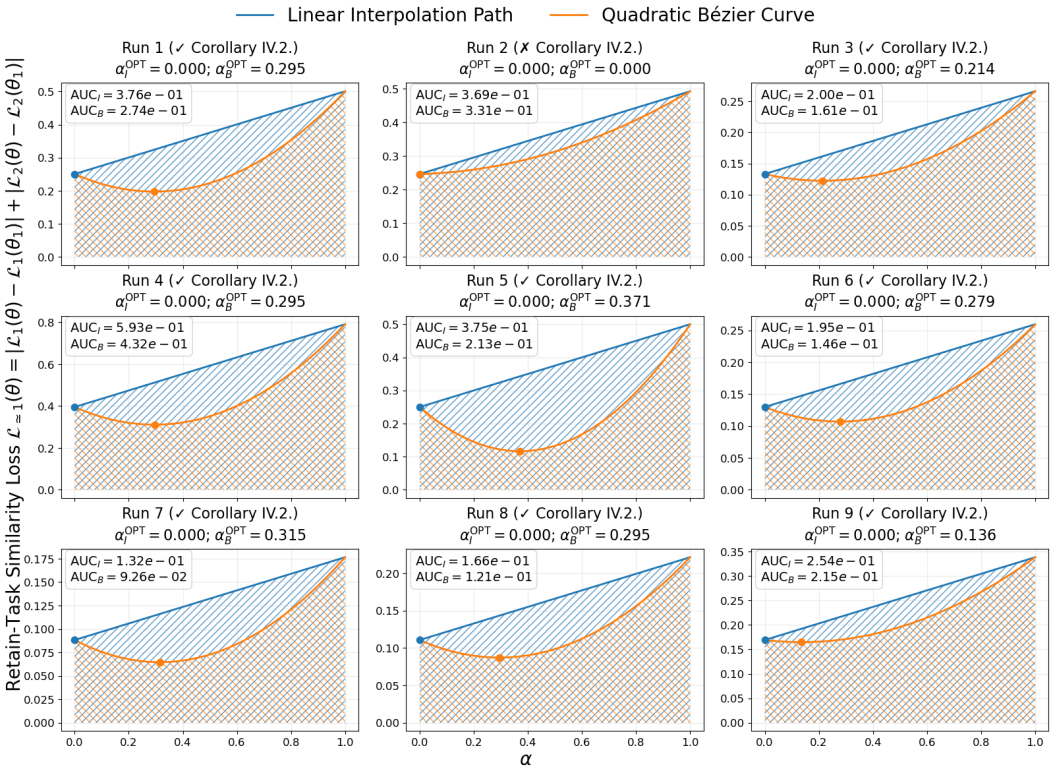

Figure 9: Machine unlearning setting 1: $\theta_R = \theta_2$. Feature correlation $\rho = 1$, task correlation $\tau = 0$.

the optimal points on both paths coincide at the left endpoint,

$$\theta_I^{\mathrm{OPT}} = \theta_B^{\mathrm{OPT}} = \theta_{1,2}.$$

Notice that even in this degenerate setting, the cumulative retain-task similarity loss of the quadratic Bézier curve remains smaller than that of the linear interpolation path. Although the sufficient condition in Theorem 4.3 is not satisfied because the cross-task residual vanishes, the empirical advantage persists beyond the regime directly covered by the theorem. This suggests that the benefit of the quadratic Bézier path may extend to a broader range of settings than those captured by our current sufficient condition.

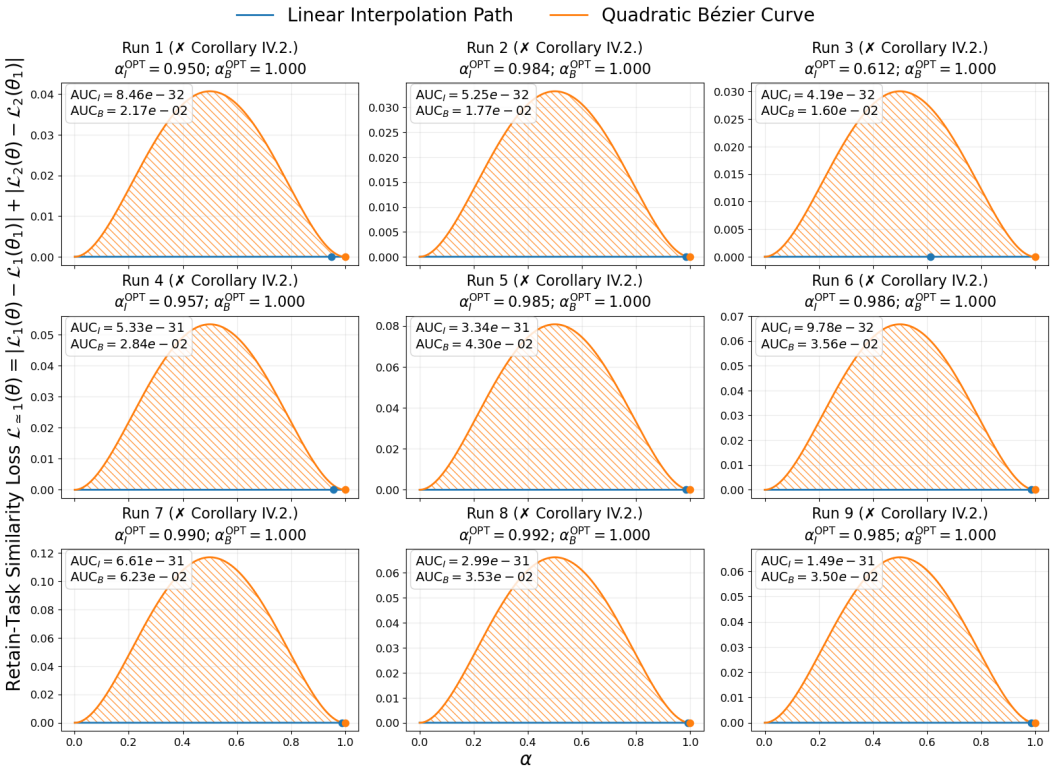

Figure 10: Machine unlearning setting 1: $\theta_R = \theta_2$. Feature correlation $\rho = 1$, task correlation $\tau = 1$.

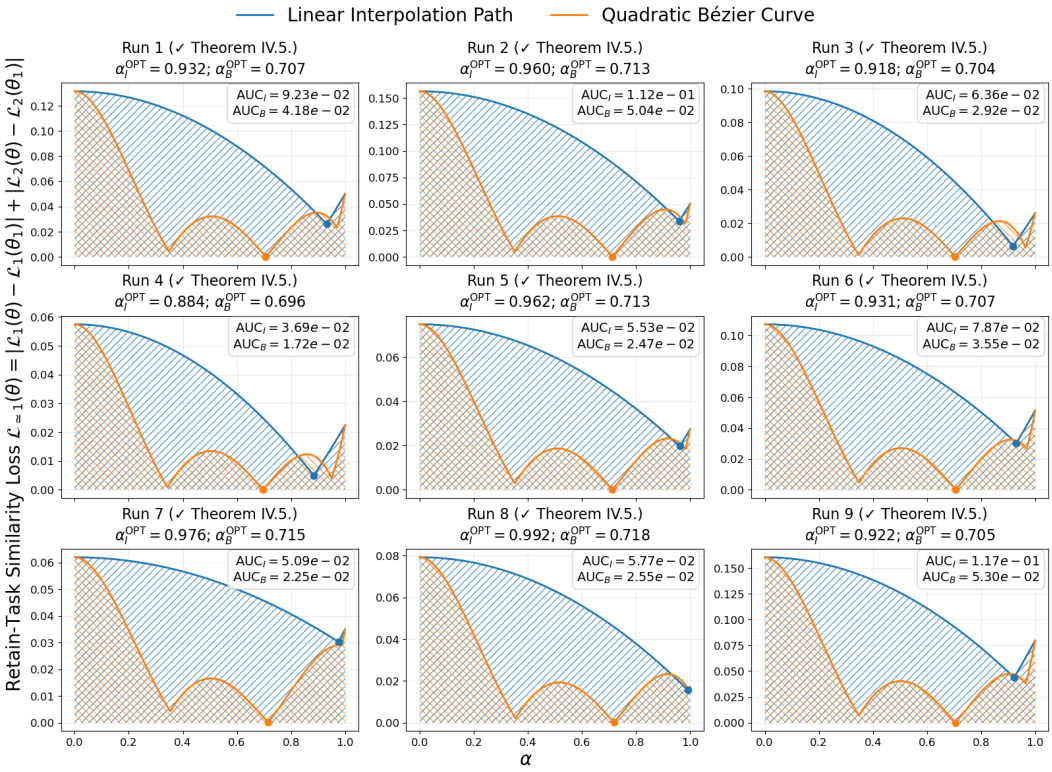

Figure 11: Machine unlearning setting 2: $\theta_R = \bar{\theta}_{1',2'}$. Feature correlation $\rho = 0$, task correlation $\tau = 0$.

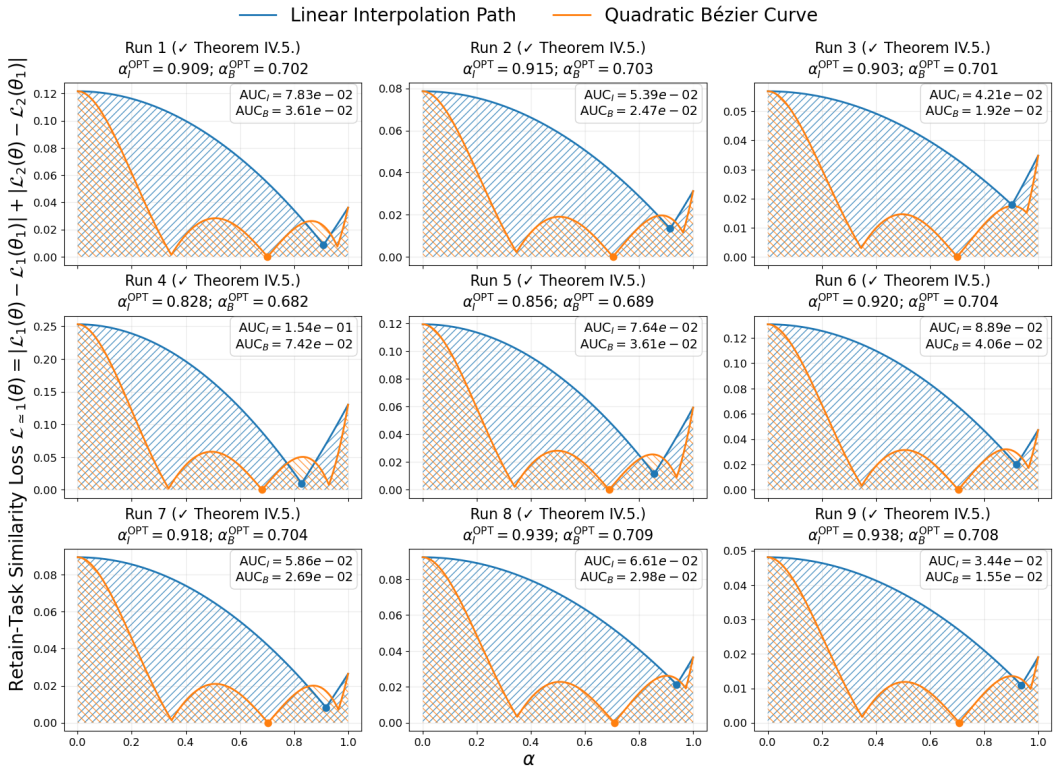

Figure 12: Machine unlearning setting 2: $\theta_R = \bar{\theta}_{1',2'}$. Feature correlation $\rho = 0$, task correlation $\tau = 1$.

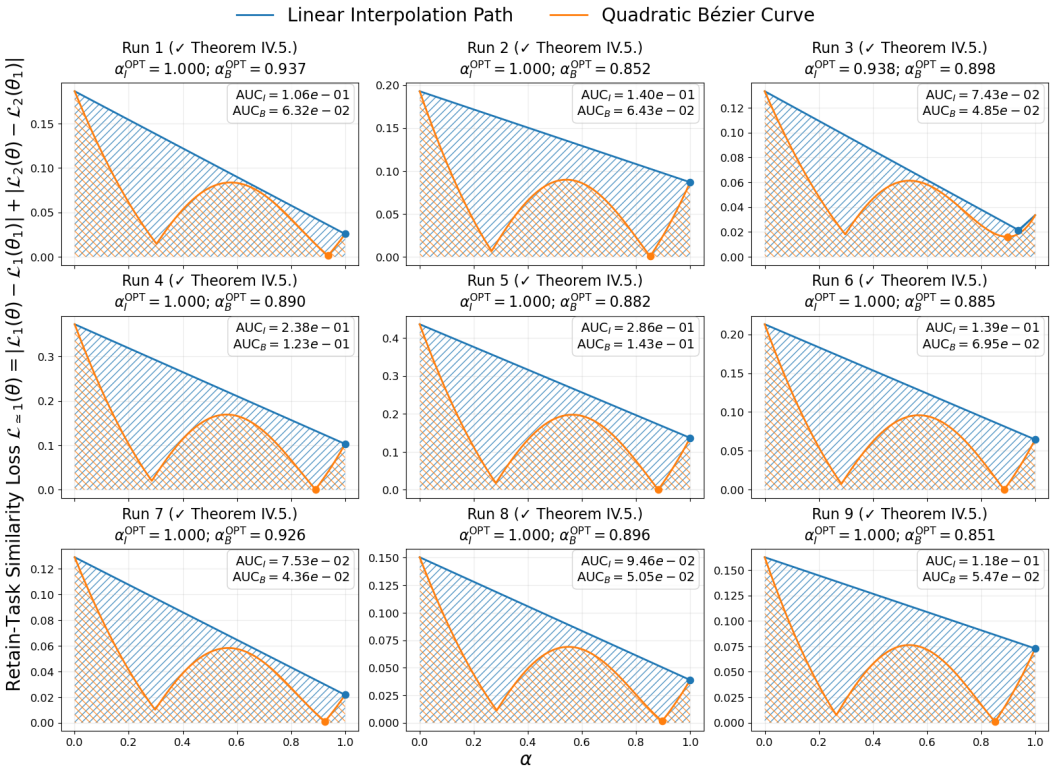

Figure 13: Machine unlearning setting 2: $\theta_R = \bar{\theta}_{1',2'}$. Feature correlation $\rho = 1$, task correlation $\tau = 0$.

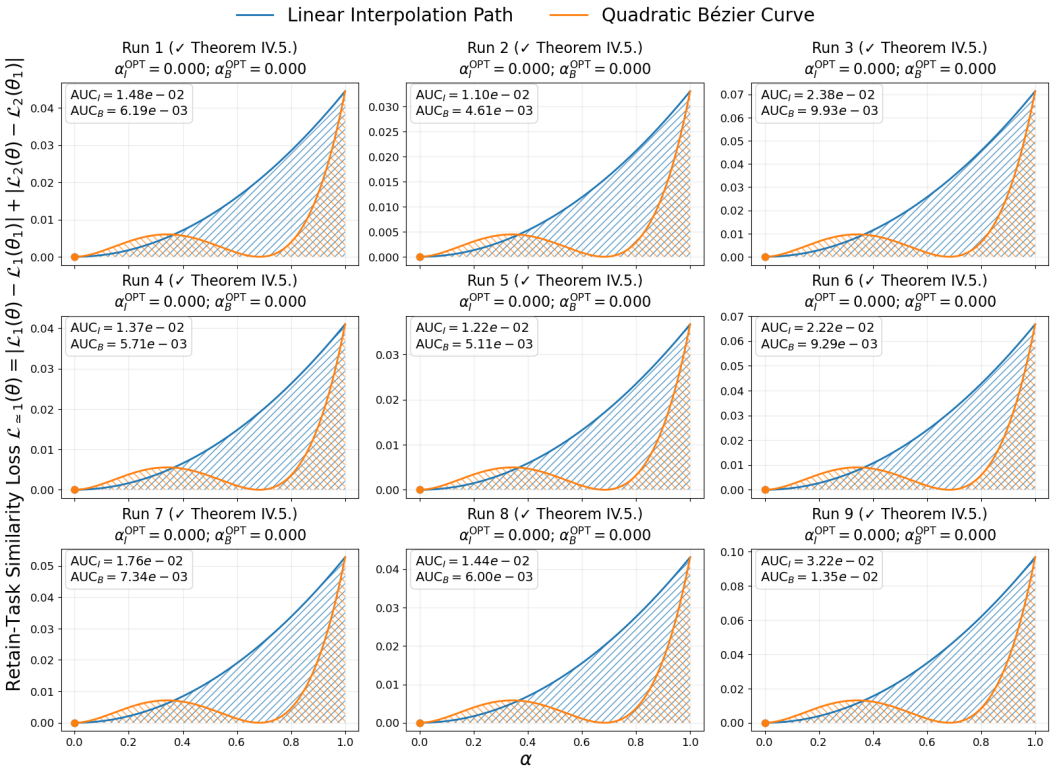

Figure 14: Machine unlearning setting 2: $\theta_R = \theta_2$. Feature correlation $\rho = 1$, task correlation $\tau = 1$.