Autonomous Machine Unlearning via Projected Gradient Ascent: Securing Financial Privacy in Convex Optimization Models

Justin Minseob Seo
miseo@ucsd.edu

Jun Kun Wang
jkw005@ucsd.edu

Abstract

The rapid integration of machine learning into financial services makes user privacy a strict legal and ethical imperative. Under regulations like the GDPR, users possess the Right to be Forgotten. Exact model retraining guarantees absolute privacy but scales at an impossible computational cost. We investigate whether Projected Gradient Ascent (PGA) can serve as a mathematically verifiable proxy for exact retraining in convex environments. Utilizing Logistic Regression on the Statlog German Credit Dataset, we isolate the unlearning dynamics from stochastic noise. Our geometric audits prove that while single user unlearning achieves surgical precision, production scaling reveals a Batch Interference Effect that dilutes precision. We further identified a 34 step Safety Window and an autonomous stopping criterion to prevent catastrophic forgetting. Finally, Membership Inference Attacks exposed a fundamental Privacy Utility Deadlock. Highly confident outliers become trapped on the flat plateau of the sigmoid curve, preventing true amnesia without destroying the global model utility.

Website:
https://juseotin.github.io/Optimization_Dynamics_MachineUnlearning
Code:
https://github.com/juseotin/Optimization_Dynamics_MachineUnlearning

1 Introduction 2
2 Methods and Data Framework 2
3 Formal Geometric Audit 2
4 Solving the Stopping Point Paradox 3
5 Privacy Auditing and The Memorization Trap 4
6 Conclusion 5
References 5
Appendices A1

1 Introduction

The primary challenge in machine unlearning is balancing operational efficiency with mathematical guarantees. When users revoke their data, exact retraining provides a perfect privacy guarantee Bourtoule et al. (2021). However, this burns the model down to remove a single data point, proving computationally prohibitive at an $O(N)$ scale.

Previous explorations into deep neural networks revealed significant issues. The stochastic nature of nonconvex loss landscapes makes it difficult to distinguish between true unlearning and random model degradation. To resolve this, we focus exclusively on strictly convex optimization. By utilizing Logistic Regression, we operate in a deterministic landscape with a single global minimum, allowing us to mathematically isolate and verify the precise geometry of the unlearning mechanism.

2 Methods and Data Framework

2.1 The Financial Testbed

We shifted our experimental testbed to real world financial data using the Statlog German Credit Dataset Hofmann (1994). Following one hot encoding, the dataset consists of 48 distinct financial features such as savings account balances and employment duration. The target is a binary classification task predicting Good or Bad credit risk. A standard scaler was applied to ensure all features maintain a mean of zero and unit variance, which is critical for stabilizing gradient math during aggressive unlearning.

2.2 Algorithm: Projected Gradient Ascent

We utilize Projected Gradient Ascent to surgically reverse the learning process. Instead of minimizing error, PGA maximizes the Binary Cross Entropy loss strictly on a targeted forget batch of 20 highly memorized outliers.

The update rule at step t is defined as:

$$w_{t+1} = w_t + \eta \nabla \mathscr{L}_{\text{forget}}(w_t) \tag{1}$$

3 Formal Geometric Audit

To verify that the unlearning process is mathematically sound and not merely adding random noise, we analyzed the geometric relationship between the input features x and the resulting weight updates Δw, as visualized in Figure 1.

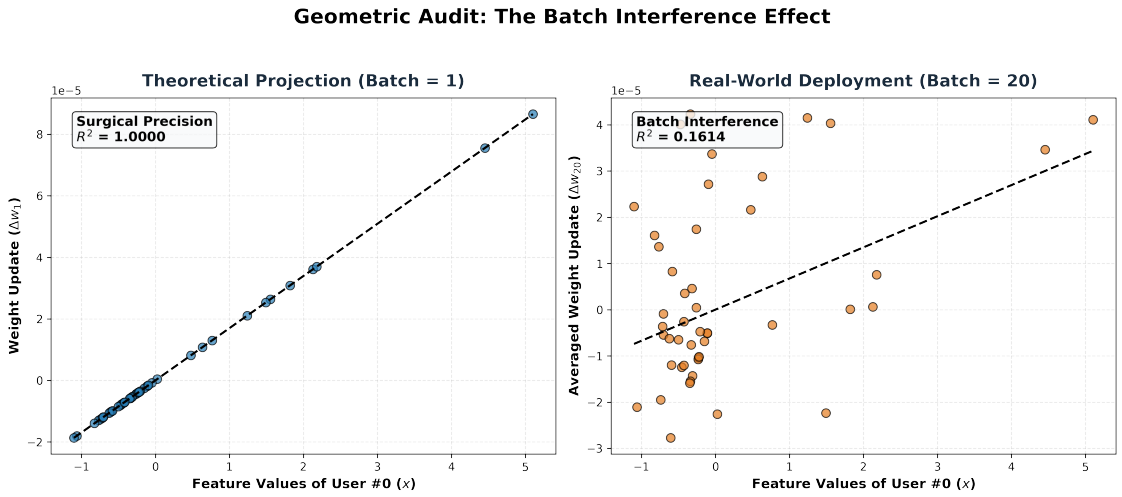

Figure 1: A geometric comparison of weight updates. The theoretical baseline ($N = 1$) shows perfect linear projection, while the production scale ($N = 20$) demonstrates the scattering caused by the Batch Interference Effect.

3.1 Theoretical Baseline ($N = 1$)

We first isolated the update for a single user. Because the pure unlearning gradient is strictly proportional to the feature vector, the weight update acts as a perfect mathematical projection. Running an Ordinary Least Squares regression between the features and the weight updates yielded an R^2 of exactly 1.0.

3.2 Production Scaling and Batch Interference ($N = 20$)

In real world deployments, financial institutions process unlearning requests in batches. When scaling our target batch to 20 users, the algorithm must average the competing gradients of distinct financial profiles. This compromise heavily dilutes the surgical precision for any individual user within the batch, reducing the R^2 to 0.1614. We define this degradation as the Batch Interference Effect. Despite this interference, the Cosine Similarity of the trajectory remained perfectly locked at 0.427339, proving that the algorithm maintains absolute directional stability without random geometric drift.

4 Solving the Stopping Point Paradox

Unlearning is a highly nonlinear process. Ascending the gradient indefinitely pushes the weights too far and inevitably leads to catastrophic forgetting of the retained dataset (Figure 2).

By measuring the Euclidean distance to a perfectly retrained Gold Standard oracle model, we identified that the optimal unlearned state occurs at Step 65. However, real world

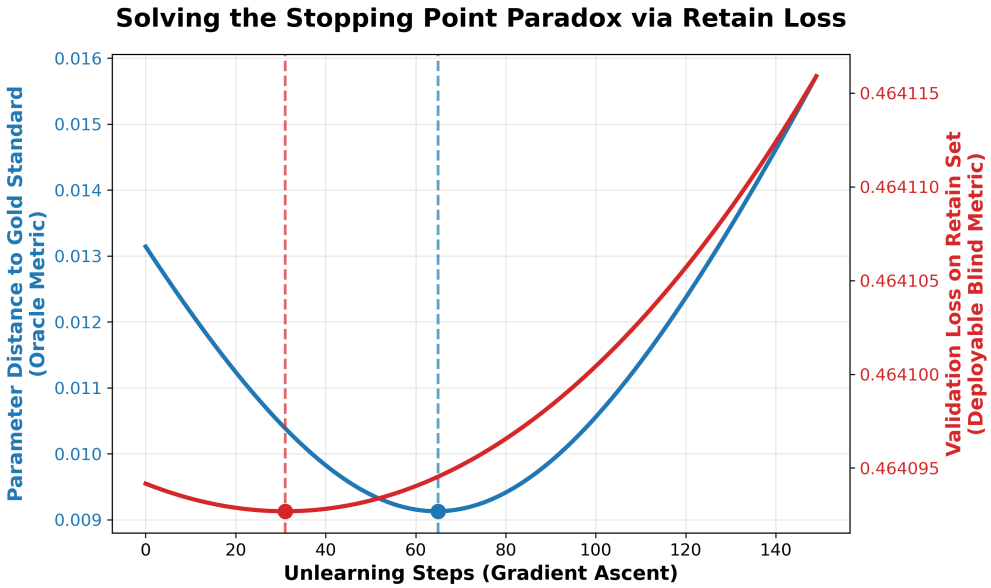

Figure 2: The Stopping Paradox. The model achieves optimal proximity to the Gold Standard at Step 65, but general utility (Validation Loss) begins degrading catastrophically after Step 31.

production systems do not have access to an oracle model. To solve this, we implemented a deployable blind metric by monitoring the Validation Loss on the Retain Set.

This analysis revealed a critical 34 step Safety Window. The model's actual utility begins to degrade rapidly after Step 31. This proves that systems can autonomously use validation loss to halt unlearning precisely at Step 31, preserving peak global utility without requiring computationally expensive oracles.

5 Privacy Auditing and The Memorization Trap

Geometric alignment to a Gold Standard is a necessary but insufficient condition for true privacy. To verify if true amnesia was achieved, we audited the optimal Step 31 model using a Membership Inference Attack (Figure 3).

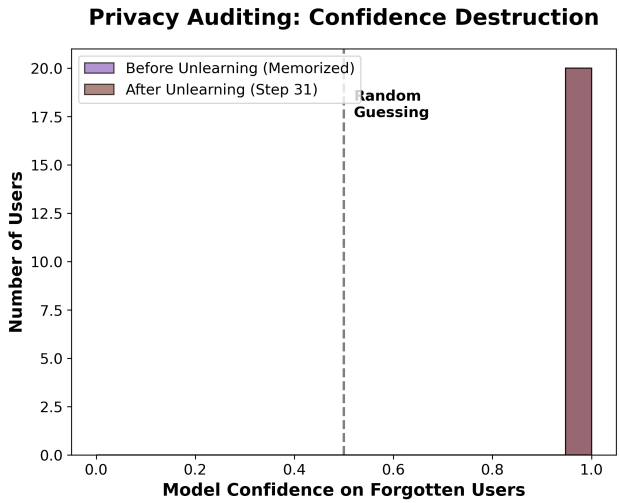

Figure 3: Membership Inference Attack evaluation. Target user probabilities remain dangerously high, illustrating the Privacy Utility Deadlock at the edge of the sigmoid curve.

True amnesia requires the model to evaluate a targeted user with the same uncertainty as a completely unseen control profile, approaching a 50 percent random guessing threshold. Our MIA audit revealed a fundamental vulnerability. At Step 31, the model's confidence on the targeted forget batch only dropped from 99.09% to 98.93%.

Because these targeted users were extreme outliers, they were deeply memorized and pushed to the flat extreme edges of the logistic sigmoid curve. On this plateau, mathematical gradients vanish. Consequently, standard gradient ascent becomes trapped in a Privacy Utility Deadlock. It is mathematically impossible to push these high confidence profiles down to a safe privacy threshold without blowing past the Safety Window and destroying the entire model.

6 Conclusion

This research successfully isolates and verifies the geometric trajectory of convex machine unlearning. While Projected Gradient Ascent is mathematically perfect in isolation, its real world viability is severely limited by two phenomena. First, the Batch Interference Effect inherently dilutes surgical precision when processing multiple users. Second, the Privacy Utility Deadlock proves that standard gradient ascent is completely insufficient for erasing high confidence outliers. Future architectures must explore dynamic learning rates or weight reinitialization to shatter deep memorization within the safe deployable window.

References

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

Hofmann, Hans. 1994. "Statlog (German Credit Data)." UCI Machine Learning Repository. [Link]

Appendices

A.1 Training Details A1

A.1 Training Details

The unlearning procedures were conducted utilizing standard Convex Optimization libraries in Python, specifically relying on NumPy for gradient operations and scikit learn for the baseline Logistic Regression implementations. Hyperparameter tuning was conducted to isolate a stable learning rate η that prevented mathematical overflow during the aggressive gradient ascent phase.