Autonomous Machine Unlearning via Projected Gradient Ascent

Securing Financial Privacy in Convex Optimization Models

Justin Minseob Seo · UC San Diego · miseo@ucsd.edu

Mentor: Jun-Kun Wang · jkw005@ucsd.edu

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

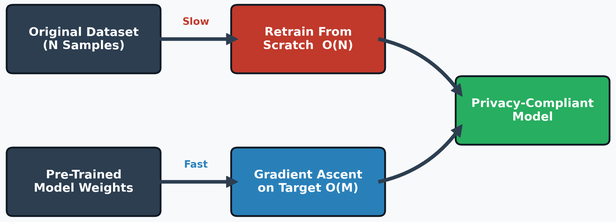
1. Motivation

- **Paradigm Shift:** Machine learning in finance makes user privacy a strict legal and ethical imperative.
- **Data Lifecycle:** Organizations need efficient ways to "unlearn" data from live production models.
- **Goal:** Validate Projected Gradient Ascent (PGA) as a verifiable alternative to full retraining.

2. The Problem: The "Right to be Forgotten"

- **The Mandate:** GDPR requires companies to delete user data upon request.
- **The Bottleneck:** Exact model retraining is computationally expensive and operationally wasteful.
- **The Flaw:** Current heuristics inject noise and lack verifiable mathematical guarantees.

The Unlearning Paradigm: Sidestepping the Retraining Bottleneck



3. Methods: Surgical Unlearning

- **Methodology:** We utilize **Projected Gradient Ascent (PGA)** on logistic regression models to reverse the learning process.
- **Objective:** Maximize error on the deleted data while preserving boundaries for the retained data.

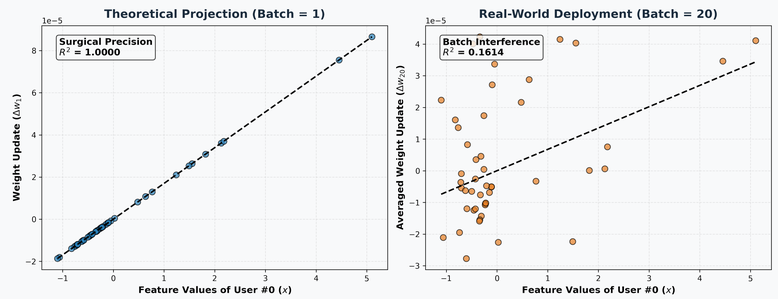
The Unlearning Update Rule:

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

4. Formal Geometric Audit

- **Theoretical Baseline ($N = 1$):** Unlearning a single user acts as a perfect mathematical projection. The gradient is strictly proportional to the feature vector ($R^2 = 1.0$).
- **Production Scaling ($N = 20$):** Batch unlearning reveals a **Batch Interference Effect**. Averaging competing gradients dilutes precision, reducing R^2 to 0.1614.
- **Stability:** Despite the scatter, **Cosine Similarity** remains locked at 0.427339, proving zero geometric drift in the trajectory.

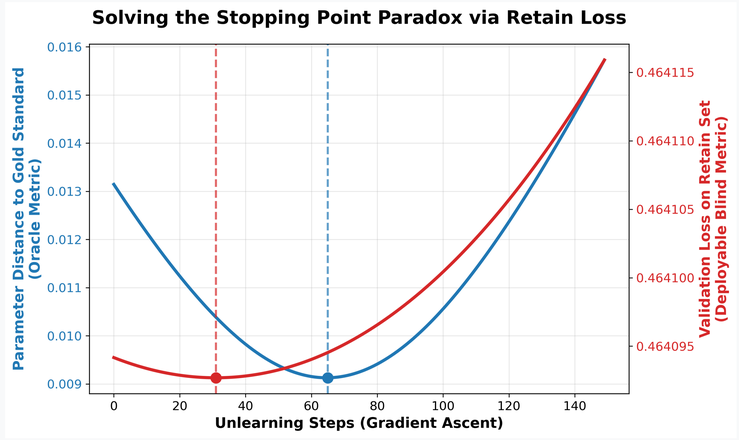
Geometric Audit: The Batch Interference Effect



5. Solving the Stopping Paradox

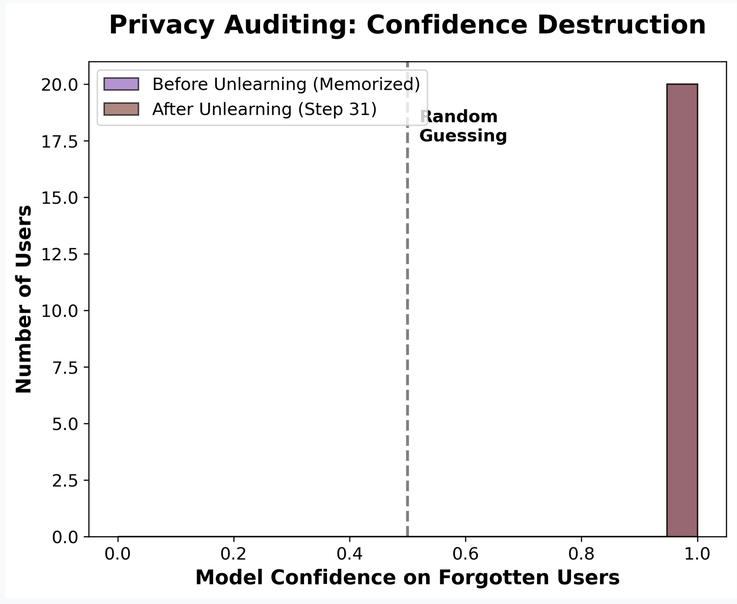
- **The Efficiency Window:** Ascending the gradient indefinitely causes catastrophic forgetting.
- **The Breakthrough:** Monitoring the validation loss on the retain set acts as a deployable "Blind Metric". This creates a **34 step Safety Window** where the model maintains utility.

Solving the Stopping Point Paradox via Retain Loss



6. Privacy Auditing: The "Amnesia" Test

- **The Problem:** How do we prove a model actually forgot a person?
- **The Test:** We measure prediction confidence. Extreme certainty means the profile is still memorized. True amnesia is represented by random guessing.
- **The Result:** Target confidence only dropped from **99.09%** to **98.93%** at Step 31.
- **Conclusion:** Highly confident profiles get trapped. Model utility breaks long before true amnesia is achieved.



7. Discussion & Next Steps

- **Discovery:** We mathematically validated PGA's stability but uncovered a fundamental **Privacy Utility Deadlock** in FinTech models.
- **Impact:** Standard gradient ascent is insufficient for outlier erasure. Future models must integrate weight resets to shatter memorization.

8. References

1. Bourtoule, L., et al. (2021). Machine unlearning. *IEEE Security and Privacy*.
2. Hofmann, H. (1994). Statlog (German Credit). UCI ML Repository.