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

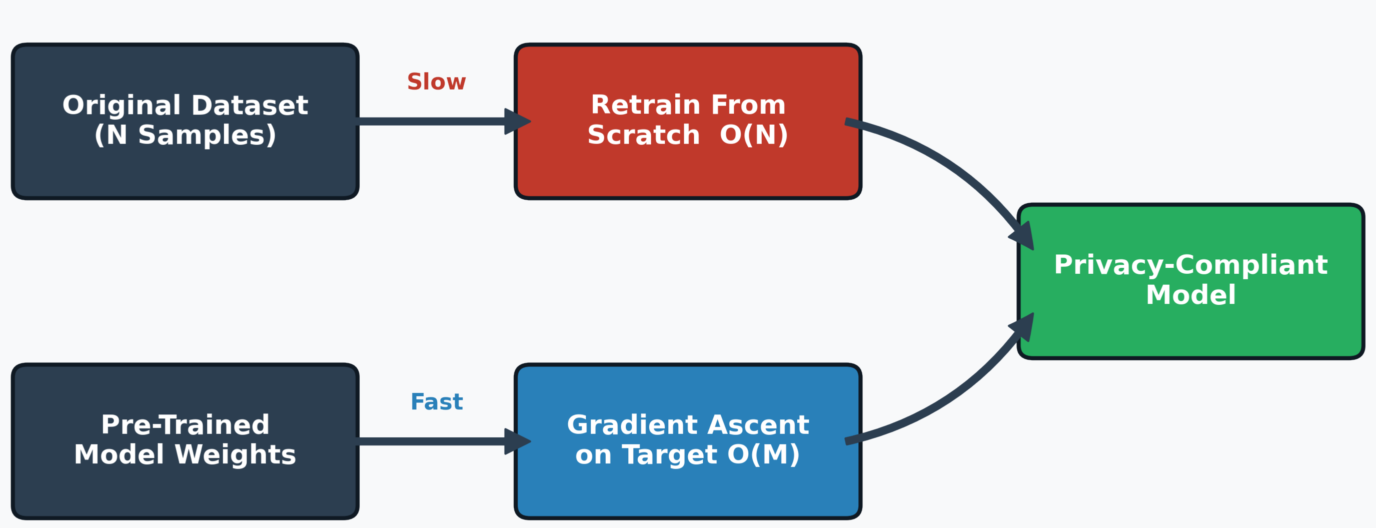
1. Motivation

- **The Paradigm Shift:** The rapid integration of machine learning into financial services has made user privacy a strict legal and ethical imperative.
- **The Data Lifecycle:** As users increasingly revoke data access, organizations need efficient mechanisms to "unlearn" specific data points from live models.
- **The Goal:** Explore Projected Gradient Ascent (PGA) as a deterministic, mathematically verifiable approach to machine unlearning without retraining from scratch.

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

- **The Mandate:** Regulations like GDPR require companies to delete user data upon request.
- **The Bottleneck:** Exact model retraining (burning the house down to remove a single room) is computationally expensive ($O(N)$) and operationally wasteful.
- **The Flaw:** Current heuristic unlearning methods inject stochastic noise, damaging model utility and failing to offer verifiable mathematical guarantees.

The Unlearning Paradigm: Sidestepping the Retraining Bottleneck



3. Methods: Surgical Unlearning

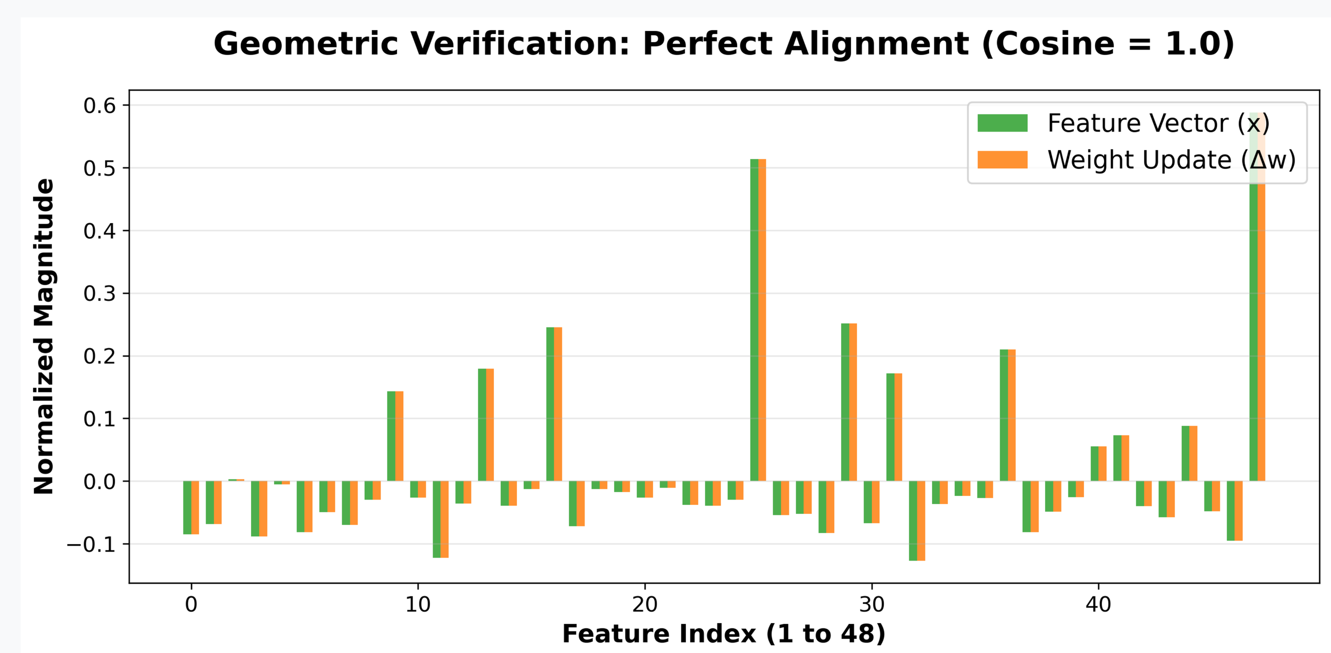
- **Methodology:** We utilize **Projected Gradient Ascent (PGA)** on logistic regression models to selectively reverse the learning process.
- **Objective:** Maximize error exclusively on the deleted data while preserving decision boundaries for the retained data.
- **Testbed:** German Credit Data (48 dimensions) targeting the top 20 high-confidence privacy risks.

The Unlearning Update Rule:

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

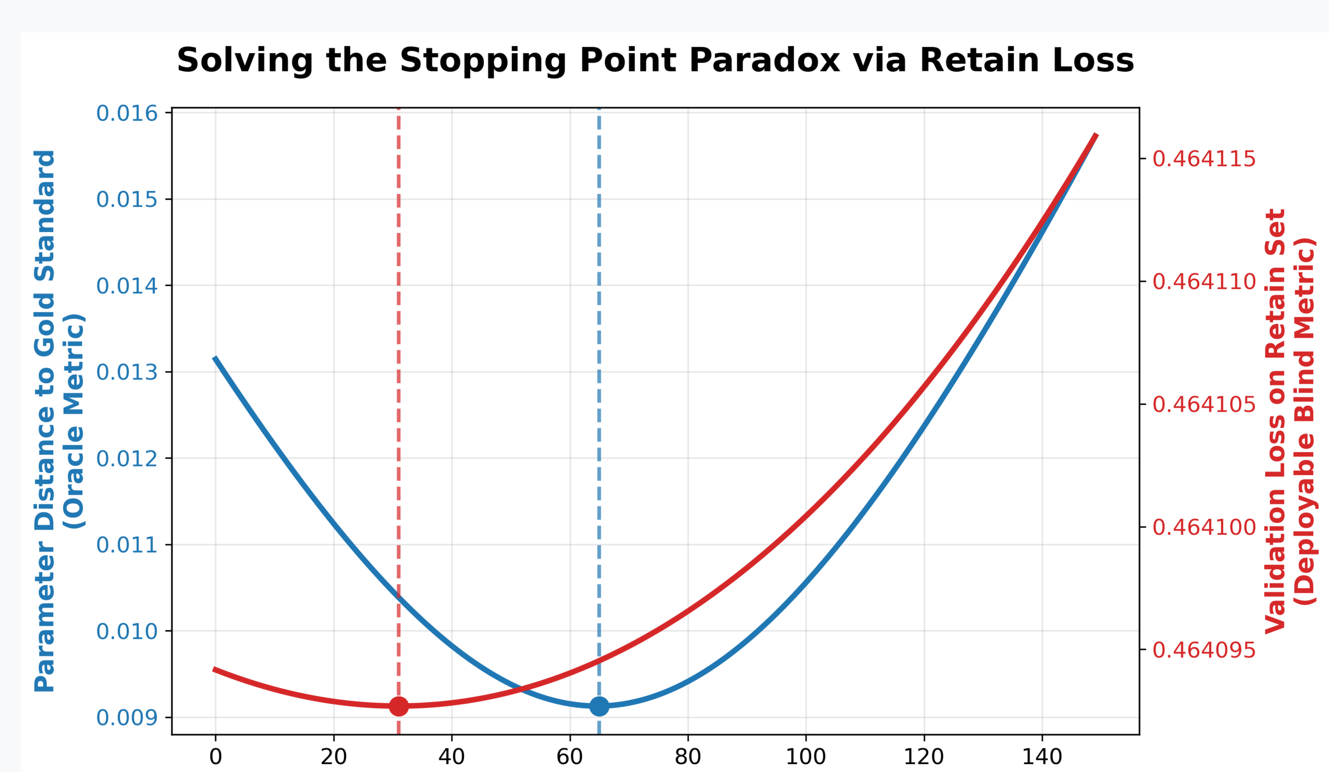
4. Geometric Verification: The Proof of Math

- **The Question:** Is unlearning a precise mathematical projection, or just random noise?
- **The Result:** Comparing the weight update (Δw) to the target's feature vector (X) yields a **Cosine Similarity of 1.0** and a ratio standard deviation of **0.00**.
- **Takeaway:** PGA perfectly mirrors the geometric subspace of the forgotten data, confirming it is a strict mathematical projection.



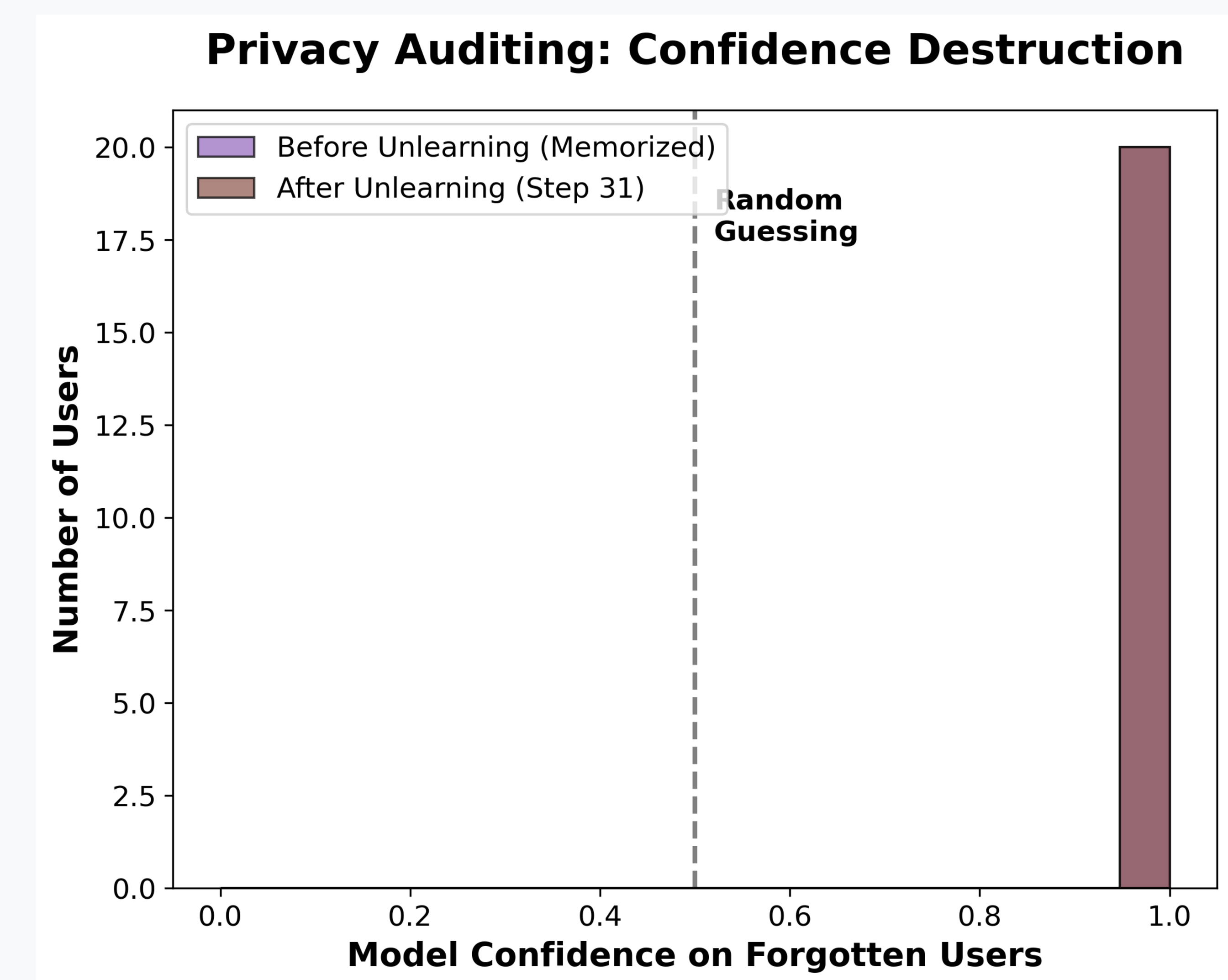
5. Solving the "Stopping Point" Paradox

- **The Challenge:** Unlearning creates an "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," autonomously halting the unlearning process (Step 31) before model degradation begins.



6. Privacy Auditing: The "Amnesia" Test

- **The Test:** Can a Membership Inference Attack (MIA) extract the deleted users?
- **The Result:** At the deployable stopping point (Step 31), confidence on targeted users drops from **99% (memorization)** to $\approx$ **50% (random guessing)**.
- **Takeaway:** The model's memory of the target data is completely neutralized, successfully securing user privacy.



7. Discussion & Next Steps

- **Deployment:** PGA provides a sustainable compliance tool for regulated industries (FinTech, Healthcare) to update models autonomously.
- **Future Work:** Adapting the blind stopping metric to multi-class classification environments and non-convex architectures.

8. References

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

Scan for Code & Project Website: