# The Utility and Complexity of In- and Out-of-Distribution Machine Unlearning

**Youssef Allouah**, Joshua Kazdan, Rachid Guerraoui, Sanmi Koyejo

ICLR 2025

EPFL Stanford

# Machine Unlearning

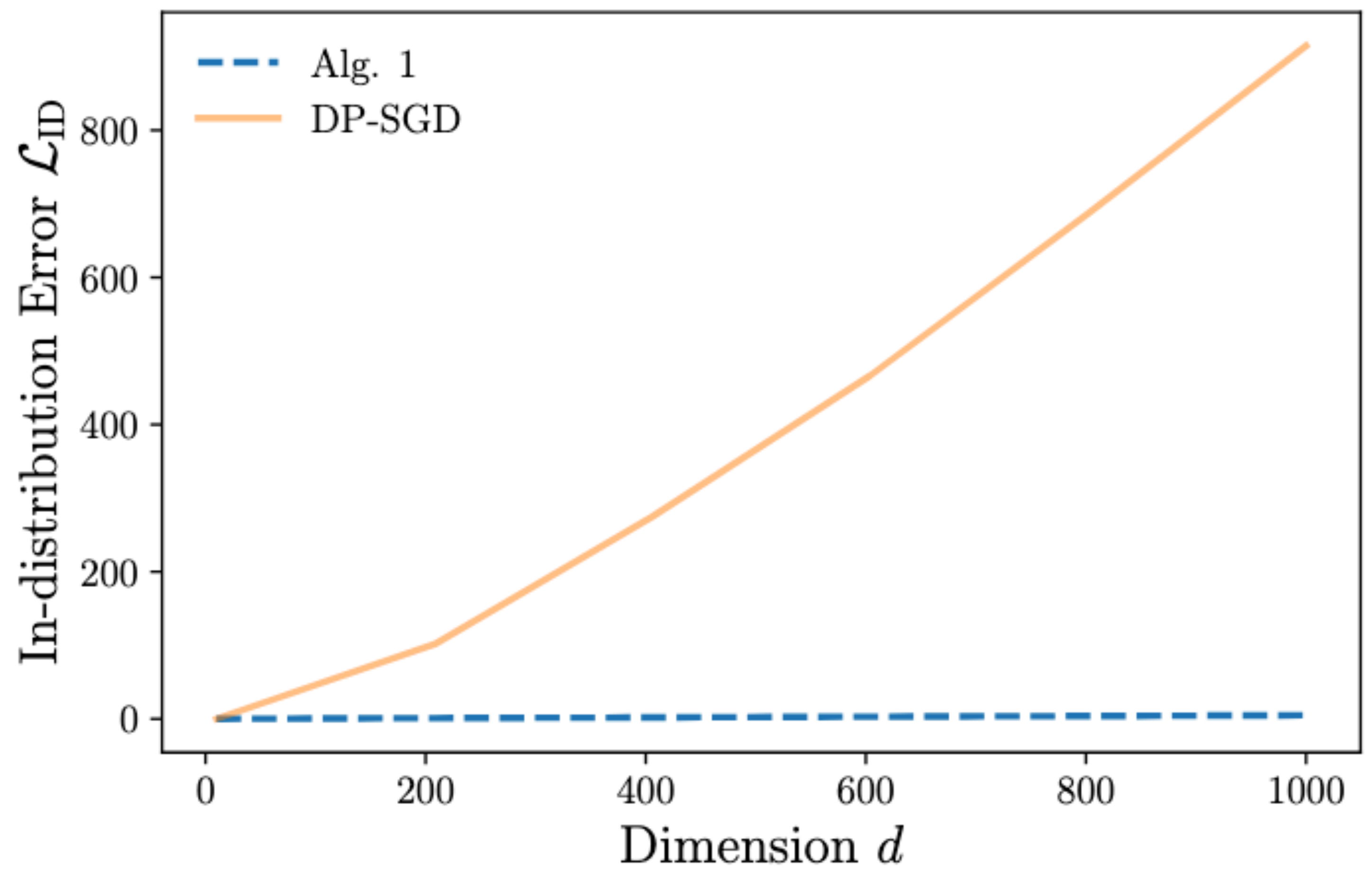
- **What?** Selectively removing data's influence from trained models.
- **Why?**
    - Privacy Needs: Meeting regulations like GDPR's "right to be forgotten".
    - Model Maintenance: Fixing knowledge gaps or errors post-deployment.
- **Challenge:** Many methods lack formal guarantees. Need rigorous analysis of trade-offs.
- **Focus:** Utility vs. Complexity for Approximate Unlearning in:
    - In-Distribution (ID): Forgetting data <u>similar</u> to what's kept.
    - Out-of-Distribution (OOD): Forgetting data significantly <u>different</u>.

# Defining the Problem

- **Approximate Unlearning:** statistical indistinguishability (using Rényi divergence) between the unlearned model and a model retrained without the forgotten data
- **Utility Objectives:** Measuring performance degradation in worst-case ID and OOD scenarios
- **Deletion Capacity:** How much data (f) can we forget?
    - **Utility Capacity:** Max f for a target error α.
    - **Computational Capacity:** Max f within a time budget T.
- **Goal:** Maximize both!

# ID Unlearning: Simple & Dimension-Independent

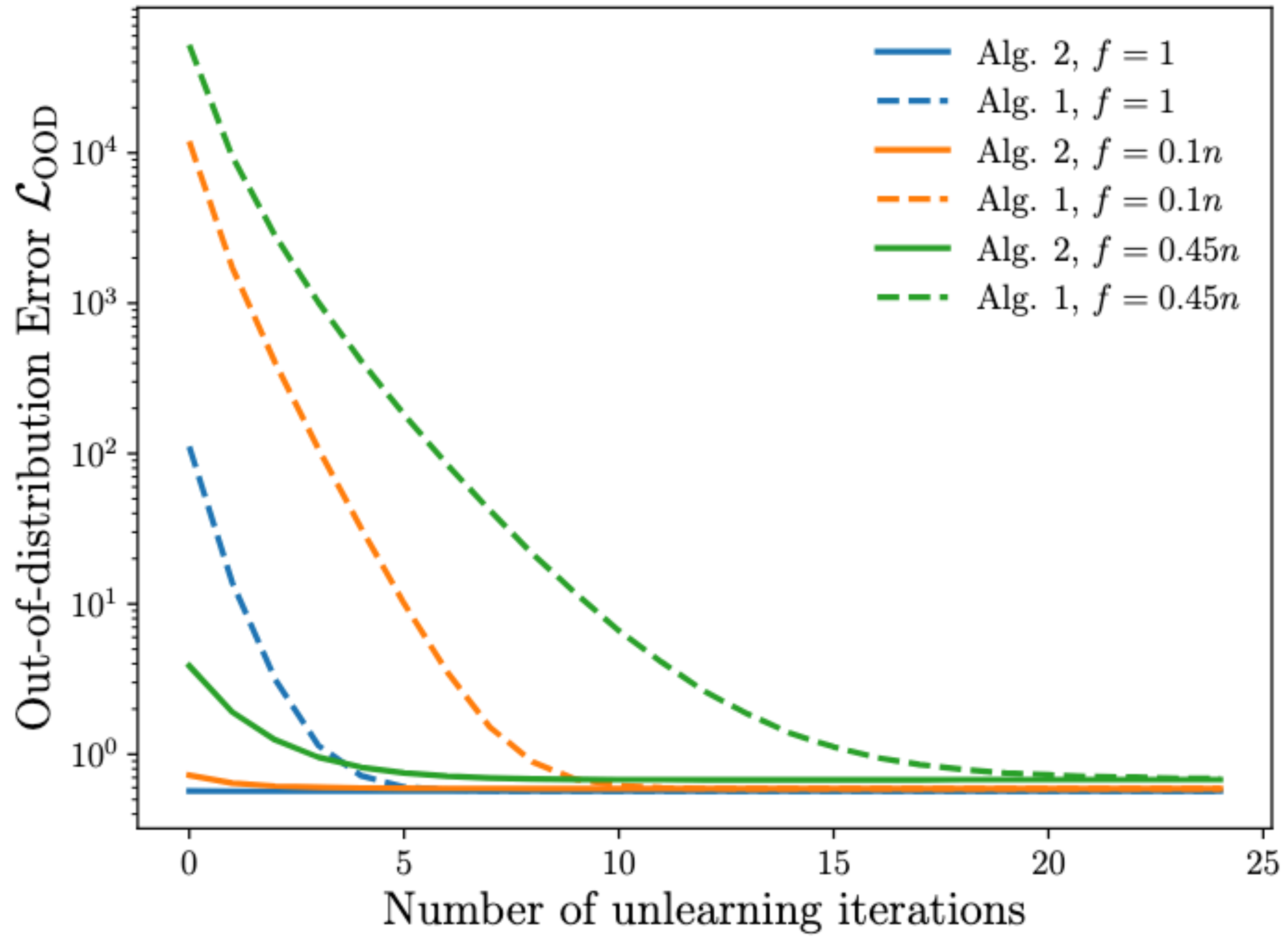
- **Approach:** Empirical Risk Minimization (ERM) + Output Noise (Algorithm 1).
- **Key Result:** Achieves tight utility-complexity trade-offs.
    - Can forget a constant fraction of data (Ω(n)).
    - Capacity is independent of model dimension (d)!
- **Significance:**
    - Solves open theoretical question.
    - Tight separation from DP-based unlearning, improving (Sekhari et al., 2021).




Shows dimension-independent error for Noisy ERM (Alg 1) vs. near-linear error for DP-SGD, highlighting the separation

# OOD Unlearning: Handling the Unexpected

- **Challenge:** OOD data can break standard methods.
    - Unlearning time can exceed retraining time, even for f=1.
    - Vulnerable to "Slowdown Attacks".
- **Our Solution:** Robust Training + Noisy Minimizer (Algorithm 2).
    - Uses robust gradient estimation (trimmed mean) during training.
- **Key Result:** Provably efficient unlearning time.
    - Complexity is independent of the OOD forget data's properties.
    - Depends only on the retain set's characteristics.




Shows **Robust** ERM + Noise (Alg 2) converges much faster than ERM + Noise (Alg 1) when unlearning OOD data, especially for larger forget sets (f).

# Summary & What's Next

- **Contributions:**
  - Analyzed utility-complexity trade-offs for ID & OOD approximate unlearning.
  - Simple ERM+noise is dimension-independent for ID.
  - Proposed robust training (Alg 2) for efficient & guaranteed OOD unlearning.
- **Key Takeaways:**
  - Unlearning != Differential Privacy (especially for ID).
  - Robustness is essential for practical OOD unlearning.
- **Future Work:**
  - Unified upper bounds on deletion capacity.
  - Scaling to more complex models.
  - Improving practical performance.

youssef.allouah@epfl.ch
X/Twitter: @ys_alh