A Reliable Cryptographic Framework for Empirical Machine Unlearning Evaluation

Yiwen Tu[1*] Pingbang Hu[2*] Jiaqi W. Ma[2]
[1]University of Michigan, Ann Arbor [2]University of Illinois Urbana-Champaign

Motivation

Machine unlearning attracts many attentions. To evaluate it:

- *Membership Inference Attack* (MIA): most common choice

$\Rightarrow$ directly reflects **individual privacy risk**.

However, existing MIA-based evaluations are often not

1. **Well-calibrated** across different unlearning methods;
2. **Zero-grounded**: retraining is not always ranked highest;
3. **Comparable** across different attacks, yielding inconsistency.

Overview and Contributions

1. Formalize *unlearning sample inference game*, establishing a **novel unlearning evaluation metric** for data removal efficacy.
2. Demonstrate several **provable properties** of the proposed metric, dodging various pitfalls of existing MIA-based metrics.
3. Introduce a *SWAP* test for **efficient empirical** analysis.

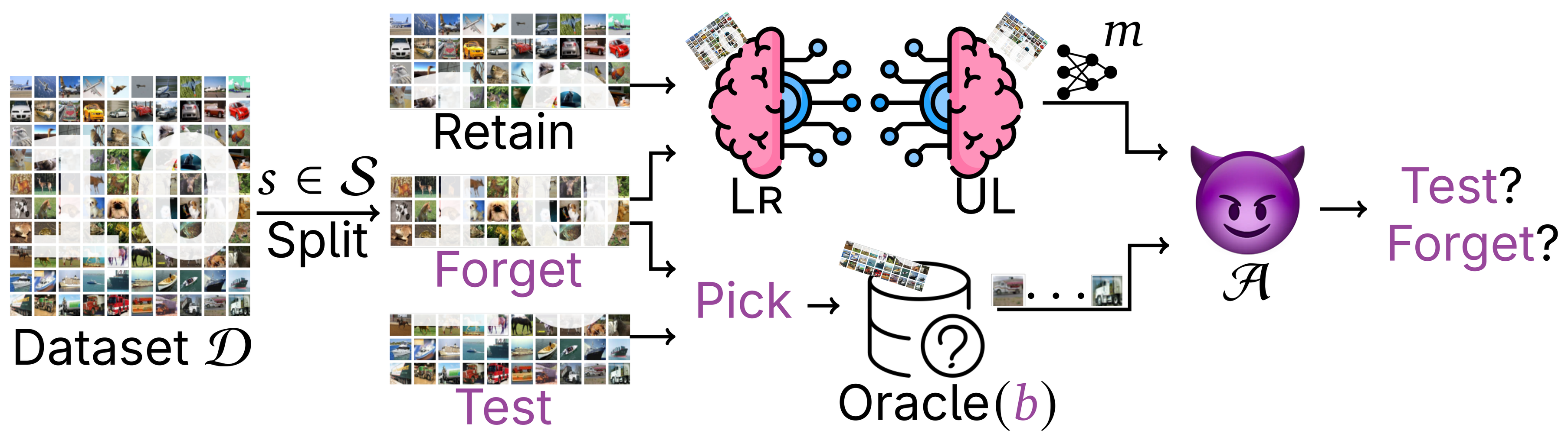
Machine Unlearning Evaluation as an Inference Game

We formulate unlearning as an **game** between

- UL: Unlearning algorithm (challenger), and
- $\mathcal{A}$: Membership-inference adversary $\mathcal{A}$.

Given a dataset $\mathcal{D}$, the *unlearning inference game* $\mathcal{G}$:

1. Split $\mathcal{D}$ into retain, **forget**, and **test** sets, forming a split $s \in \mathcal{S}$.
2. *Random oracle* $\mathcal{O}_s(b)$ with a **secret bit** $b \in \{0, 1\}$ is instantiated.
3. UL outputs unlearned model m, and $\mathcal{A}$ attempts to infer b.



Question. How can we measure the performance of UL and $\mathcal{A}$?

Cryptographic Advantage and Unlearning Quality

Intuitively, how well UL can **fool** $\mathcal{A}$ measures the performance.

- This is a well-known concept in cryptography: *advantage*.

For our game $\mathcal{G}$, the advantage $\operatorname{Adv}(\mathcal{A}, \mathrm{UL})$ of $\mathcal{A}$ against UL is

$$\frac{1}{|\mathcal{S}|}\left|\sum_{s\in\mathcal{S}} \Pr_{\substack{m\sim\mathbb{P}(\mathrm{UL},s)\\ \mathcal{O}=\mathcal{O}_s(0)}}(\mathcal{A}^{\mathcal{O}}(m)=1) - \sum_{s\in\mathcal{S}} \Pr_{\substack{m\sim\mathbb{P}(\mathrm{UL},s)\\ \mathcal{O}=\mathcal{O}_s(1)}}(\mathcal{A}^{\mathcal{O}}(m)=1)\right|.$$

Definition (Unlearning Quality). For any UL, its *Unlearning Quality* $\mathcal{Q}$ under a game $\mathcal{G}$ is defined as

$$\mathcal{Q}(\mathrm{UL}) := 1 - \sup_{\mathcal{A}} \operatorname{Adv}(\mathcal{A}, \mathrm{UL}),$$

Theoretical Guarantees for $\mathcal{Q}$

Theorem (Zero Grounding). For any adversary $\mathcal{A}$, we have $\operatorname{Adv}(\mathcal{A}, \textsc{Retrain}) = 0$. Hence, $\mathcal{Q}(\textsc{Retrain}) = 1$.

This guarantees that the retraining method is always the best.

Theorem (Calibrated Guarantees). Given an (ϵ, δ)-certified removal UL for some $\epsilon, \delta > 0$, for any $\mathcal{A}$ against UL, we have

$$\operatorname{Adv}(\mathcal{A}, \mathrm{UL}) \le 2 \cdot \left(1 - \frac{2 - 2\delta}{e^{\epsilon} + 1}\right) \Rightarrow \mathcal{Q}(\mathrm{UL}) \ge \frac{4 - 4\delta}{e^{\epsilon} + 1} - 1$$

Hence, $\mathcal{Q}$ calibrates with other known privacy metrics faithfully.

SWAP Test: Approximation Algorithm for $\mathcal{Q}$

To efficiently evaluate the $\mathcal{Q}$, we propose a *SWAP* test:

- Consider only ***swapped** splits* s, s' between forget and test set.
- Approximate $\operatorname{Adv}(\mathcal{A}, \mathrm{UL})$ by only few swap pairs.

Theorem (SWAP's Zero Grounding). For any $\mathcal{A}$ and swap splits $s, s' \in \mathcal{S}$, $\overline{\operatorname{Adv}}_{\{s,s'\}}(\mathcal{A}, \textsc{Retrain}) = 0$.

It turned out that SWAP is not only sufficient, but necessary.

Theorem (Blowup without SWAP). For two non-swapped splits $s_1, s_2 \in \mathcal{S}$, there exists $\mathcal{A}$ such that $\overline{\operatorname{Adv}}_{\{s_1,s_2\}}(\mathcal{A}, \mathrm{UL}) = 1$ for **any** UL. Particularly, $\overline{\operatorname{Adv}}_{\{s_1,s_2\}}(\mathcal{A}, \textsc{Retrain}) = 1$.

Experimental Results: Model Trained with Different Privacy

Consider unlearning on models trained ***with DP budgets*** ϵ.

Unlearning Quality is negatively correlated with DP budget ϵ!

UL	ϵ = 50	150	600	∞
None	$0.972^{\dagger}$	$0.960^{\dagger}$	0.932^{*}	$0.587^{\dagger}$
NegGrad	0.980^{*}	$0.975^{\dagger}$	$0.953^{\dagger}$	$0.628^{\dagger}$
RetrFinal	$0.972^{\dagger}$	$0.964^{\dagger}$	$0.939^{\dagger}$	$0.576^{\dagger}$
FtFinal	0.973^{*}	$0.963^{\dagger}$	$0.939^{\dagger}$	$0.574^{\dagger}$
Fisher	0.973^{*}	$0.967^{\dagger}$	0.942^{*}	$0.709^{\dagger}$
SalUn	0.979^{*}	$0.972^{\dagger}$	0.945^{*}	0.689^{*}
SSD	0.996^{*}	0.988^{*}	$0.981^{\dagger}$	$0.888^{\dagger}$
Retrain	0.998^{*}	0.996^{*}	0.997^{*}	0.993^{*}

(A) $\mathcal{Q}$ score *versus* DP budgets.

UL	ϵ = 50	150	600	∞
None	$0.451^{\dagger}$	$0.433^{\dagger}$	$0.454^{\dagger}$	$0.380^{\dagger}$
NegGrad	$0.476^{\dagger}$	$0.482^{\dagger}$	$0.466^{\dagger}$	$0.299^{\dagger}$
RetrFinal	$0.485^{\dagger}$	$0.485^{\dagger}$	$0.472^{\dagger}$	$0.248^{\dagger}$
FtFinal	$0.485^{\dagger}$	$0.485^{\dagger}$	$0.472^{\dagger}$	$0.247^{\dagger}$
Fisher	$0.475^{\dagger}$	$0.484^{\dagger}$	$0.463^{\dagger}$	$0.325^{\dagger}$
SalUn	0.488^{*}	0.491^{*}	$0.477^{\dagger}$	0.268^{*}
SSD	0.480^{*}	0.480^{*}	$0.468^{\dagger}$	0.244^{*}
Retrain	0.479^{*}	0.491^{*}	0.492^{*}	0.488^{*}

(B) MIA score *versus* DP budgets.

† indicates standard error of the mean is < 0.01, and * for < 0.005.

Next, we consider applying unlearning on different ***dataset sizes***.

Unlearning Quality maintains a consistent ranking of UL!

UL	Dataset percentage (%) 0.1	0.4	0.8	1.0
RetrFinal	$0.340_{\pm 0.017}$	$0.586_{\pm 0.015}$	$0.621_{\pm 0.014}$	$0.634_{\pm 0.025}$
FtFinal	$0.131_{\pm 0.011}$	$0.585_{\pm 0.016}$	$0.619_{\pm 0.014}$	$0.634_{\pm 0.024}$
Fisher	$0.751_{\pm 0.024}$	$0.679_{\pm 0.005}$	$0.734_{\pm 0.006}$	$0.791_{\pm 0.020}$
NegGrad	$0.124_{\pm 0.010}$	$0.564_{\pm 0.018}$	$0.603_{\pm 0.014}$	$0.656_{\pm 0.035}$
SalUn	$0.476_{\pm 0.014}$	$0.617_{\pm 0.016}$	$0.689_{\pm 0.013}$	$0.748_{\pm 0.004}$
SSD	$0.975_{\pm 0.008}$	$0.939_{\pm 0.025}$	$0.929_{\pm 0.021}$	$0.928_{\pm 0.015}$
Retrain	$0.999_{\pm 0.000}$	$0.997_{\pm 0.001}$	$0.993_{\pm 0.001}$	$0.993_{\pm 0.001}$

- **Well-calibrated**: $\mathcal{Q}$ not only calibrates under ϵ, but also other hyperparameters such as dataset percentage.
- **Zero-grounded**: For all settings, $\mathcal{Q}(\textsc{Retrain}) \approx 1$.
- **Comparable**: While MIA score is inconsistent, $\mathcal{Q}$ unifies it.

Next Step

1. Efficient adaptation to foundation models unlearning?
2. More complicated unlearning scenarios, such as non-i.i.d. unlearning and feature unlearning?