Gaussian certified unlearning in high dimensions

A hypothesis testing approach

Aaradhya Pandey
joint with Arnab Auddy, Haolin Zou, Arian Maleki, Sanjeev Kulkarni

Operations Research and Financial Engineering,
Princeton University

ICLR 2026

What is machine unlearning?

Machine Unlearning: Motivation and Goal

- **What is the motivation behind machine unlearning?**
  - Companies collect sensitive user data to train their ML models.
  - Users may later request that their personal records be deleted.
  - Companies must delete both the data and its statistical influence from trained models.
  - Re-training from scratch for every deletion request is prohibitively expensive.
- **What is the goal of machine unlearning?**
  - Develop **efficient algorithms** to remove the influence of selected data points.
  - It avoids full retraining while preserving **generalization capabilities** of the model.
  - It allows **unlearning** by enabling the users to exercise their right to be forgotten.

Mathematical Framework: Dataset and Training

- **Dataset** $D_n = \{(x_i, y_i)\}_{i=1}^n$ consists of feature-response pairs $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$.
- We assume they are IID samples from a generalized linear model $p(x)\, q(y \mid x^\top \beta^*)$.
- We assume the learning algorithm is **Regularized empirical risk minimization**.

$$\textbf{RERM: } \hat{\boldsymbol{\beta}} = A(\mathscr{D}_n) := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \quad L(\boldsymbol{\beta}) := \sum_{i=1}^n \ell(y_i, x_i^\top \boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta}).$$

Mathematical Framework: Retraining and Unlearning

- The **ideal retraining** for a given deletion request $M \subset [n]$ with $|M| = m$ takes the form

$$\textbf{Retraining: } \hat{\boldsymbol{\beta}}_{\setminus M} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} L_{\setminus M}(\boldsymbol{\beta}) = \lambda r(\boldsymbol{\beta}) + \sum_{i \notin M} \ell(y_i, x_i^\top \boldsymbol{\beta}).$$

- **Unlearning** is a randomized procedure $\bar{A}$ having access to the trained model $\hat{\boldsymbol{\beta}}$

$$\textbf{Unlearning: } \tilde{\boldsymbol{\beta}}_{\setminus M} := \bar{A}(\hat{\boldsymbol{\beta}}, D_M, T(D_n), b)$$

- We assume it has access to D_M, Hessian information $T(D_n)$, and external noise b.

$$\tilde{\boldsymbol{\beta}}_{\setminus M} := \bar{A}(\hat{\boldsymbol{\beta}}, D_M, T(D_n), b)$$

We need **efficient**, **certified**, and **accurate** unlearning algorithm $\bar{A}$.

Certifiability Through Hypothesis Testing

- We **introduce** f**-certifiability** through a hypothesis testing problem for an adversary.
- An adversary observing the output $\tilde{\boldsymbol{\beta}}_{\setminus M}$ tries to distinguish between the two.

$$H_0 : \tilde{\boldsymbol{\beta}}_{\setminus M} \sim P_{\text{re}} \quad \text{vs.} \quad H_1 : \tilde{\boldsymbol{\beta}}_{\setminus M} \sim P_{\text{un}}$$

- $P_{\text{re}} \overset{d}{=} \bar{A}(\hat{\boldsymbol{\beta}}_{\setminus M}, \emptyset, T(D_{\setminus M}), b)$ is $\bar{A}$ run on $D_{\setminus M}$ with no deletion requests.
- $P_{\text{un}} \overset{d}{=} \bar{A}(\hat{\boldsymbol{\beta}}, D_M, T(D), b)$ is $\bar{A}$ run on D with deletion request M.
- Certifiability holds if and only if the adversary cannot distinguish H_0 from H_1.

General f-Certifiability: Proposed Definition

- An algorithm $\bar{A}$ is (ϕ, f)-certifiable if with probability $\geq 1-\phi$ over the data D

f-Certifiability: $\inf_{|M|\leq m} \min\big(T(P_{\text{re}}, P_{\text{un}})(\alpha),\ T(P_{\text{un}}, P_{\text{re}})(\alpha)\big) \geq f(\alpha)\ \ \forall \alpha \in [0,1].$

- $f = f_{\varepsilon,\delta}(\alpha) = \max\{0, 1-\delta-e^{\varepsilon}\alpha, e^{-\varepsilon}(1-\delta-\alpha)\}$ recovers (ε,δ)-certifiability.
- **Trade-off function:** For two distributions P, Q any type-I error $\alpha \in [0,1]$:

Neyman–Pearson curve: $T(P,Q)(\alpha) := \inf_{0\leq\phi\leq 1} \big\{\mathbb{E}_Q[1-\phi] \,\big|\, \mathbb{E}_P[\phi] \leq \alpha\big\}.$

- A higher Neyman-Pearson curve $\Leftrightarrow$ harder to distinguish $\Leftrightarrow$ better certifiability.

Gaussian Certifiability: A Canonical Choice in High-Dimensions

- Log-concave mechanisms "behave" like a scaled Gaussian mechanism as $p \uparrow \infty$.
- Gaussian mechanism is the canonical mechanism to achieve certifiability in HD.
- Gaussian certifiability with $f = f_{G,\varepsilon}$ *tightly* characterizes the Gaussian mechanism.

The Gaussian trade-off curve: $f_{G,\varepsilon}(\alpha) := T\big(\mathcal{N}(0,1), \mathcal{N}(\varepsilon,1)\big)(\alpha)$

Accuracy Through Generalization error divergence

- An algorithm that outputs pure noise achieves perfect certifiability but zero utility.
- We need a criterion to measure the generalization capabilities of the unlearned model.
- We measure this for a fresh test point (x_0, y_0) independent of D (Zou et al. 2025):

$$\mathrm{GED}_\ell(A, \bar{A}; M, D_n) := \mathbb{E}\left[\left|\ell(y_0, x_0^\top \hat{\boldsymbol{\beta}}_{\setminus M}) - \ell(y_0, x_0^\top \tilde{\boldsymbol{\beta}}_{\setminus M})\right| \,\middle|\, D_n\right]$$

- GED$= 0$ means the unlearned model generalizes identically to the retrained model.

Newton-Based Unlearning Algorithm

- **Step 1 — Newton approximation.**

$$\hat{\boldsymbol{\beta}}^{(1)}_{\setminus M} = \hat{\boldsymbol{\beta}} - \left[\nabla^2 L_{\setminus M}(\hat{\boldsymbol{\beta}})\right]^{-1} \nabla L_{\setminus M}(\hat{\boldsymbol{\beta}})$$

  - Since $\nabla L(\hat{\boldsymbol{\beta}}) = 0$, $\nabla L_{\setminus M}(\hat{\boldsymbol{\beta}}) = -\sum_{i \in M} \dot{\ell}(y_i, x_i^\top \hat{\boldsymbol{\beta}})\, x_i$ involves only the deleted data.

- **Step 2 — Gaussian noise injection.**

$$\tilde{\boldsymbol{\beta}}_{\setminus M} = \hat{\boldsymbol{\beta}}^{(1)}_{\setminus M} + b, \qquad b \sim \mathcal{N}\left(0, \tfrac{R^2}{\varepsilon^2} I_p\right)$$

  - We choose $R = C_1(n)\sqrt{C_2(n) m^3/(2\lambda \nu n)}$ for some $C_1, C_2 = O(\text{polylog}(n))$.
  - These constants in the noise calibration can be computed for a given problem.

Main Theoretical Results: Certifiability and Accuracy [Pandey *et al.*, 2025]

Gaussian certifiability: Under some assumptions on ℓ, r, and D, the one-step noisy Newton $\tilde{\boldsymbol{\beta}}_{\setminus M}$ with noise $b \sim \mathcal{N}(0, R^2\varepsilon^{-2}I_p)$ achieves $(\phi_n, f_{G,\varepsilon})$ certifiability with

$$R = C_1(n)\sqrt{\frac{C_2(n)m^3}{2\lambda\nu n}}, \qquad \phi_n = nq_n^{(y)} + 8n^{-3} + ne^{-p/2} + 2e^{-p} \to 0.$$

Vanishing GED: With the same assumptions and the choice of R as above, we have with probability $\geq 1-(n+1)q_n^{(y)} - 14n^{-3} - ne^{-p/2} - 2e^{-p} - e^{-(1-\log 2)p}$:

$$\mathrm{GED}(\tilde{\boldsymbol{\beta}}_{\setminus M}, \hat{\boldsymbol{\beta}}_{\setminus M}) \leq C_1(n)\sqrt{C_2(n)}\left(\tfrac{1}{\varepsilon} + \tfrac{1}{\sqrt{p}}\right)\sqrt{\tfrac{m^3(m+2)}{\lambda\nu n}} \cdot \mathrm{polylog}(n).$$

Overall message and comparison with previous works

- One noisy Newton step achieves both **Gaussian certifiability** and vanishing GED.
- It happens for $m = o(n^{1/4-\alpha})$ for any $\alpha > 0$, even when $n, p \to \infty$ with $n/p \to \gamma$.
- [Zou *et al.*, 2025] showed at least two Newton steps are required to achieve (ϕ, ε)**-certifiability** with vanishing GED. We show that one step achieves both.
- This discrepancy is due to the sub-optimality of the notion of (ϕ, ε)**-certifiability** under noise additions, which **Gaussian certifiability** is able to overcome optimally.
- Gaussian mechanism is the canonical mechanism to achieve certifiability in HD.
- The Gaussian certifiability is the canonical notion of certifiability in high dimensions.

Experiments and Future Directions

- **Real data experiments:** With low-dimensional *MNIST* ($p = 784$), all three notions of certifiability are comparable, whereas for high-dimensional *IMDb* ($p = 3161$), the Gaussian framework outperforms (ε, δ) unlearning, confirming the advantage in HD.

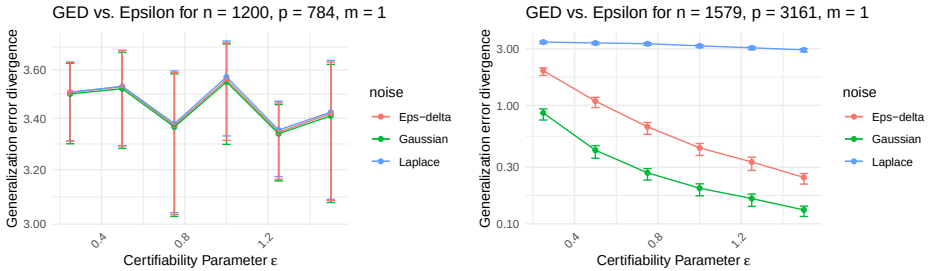

Figure 1: Mean GED across ε for (left) MNIST data with $p = 784$ and (right) IMDb data with $p = 3161$. We set $\lambda = 0.01$. Experiments were repeated 20 times across random draws of the test

Interesting Future Directions to Pursue

- Extend the current certifiability framework with
  - Non-convex loss,
  - First-order methods,
  - Online unlearning,
  - Beyond $p \sim n$ regime.

References I

[Pandey *et al.*, 2025] A. Pandey, A. Auddy, H. Zou, A. Maleki, and S. Kulkarni.
Gaussian Certified Unlearning in High Dimensions: A hypothesis testing approach.

[Dong *et al.*, 2022] J. Dong, A. Roth, and W. J. Su.
Gaussian Differential Privacy, *J. R. Stat. Soc. Series B* **84**(1), 337, 2022.

[Zou *et al.*, 2025] H. Zou, A. Auddy, Y. Kwon, K. Rahnama Rad, and A. Maleki.
Certified Data Removal Under High-dimensional Settings.
arXiv preprint arXiv:2505.07640, 2025.