Approximate Machine Unlearning for High Dimensional R-ERM

Joint work with Arnab Auddy, Kamiar Rahnama Rad, Arian Maleki and Yongchan Kwon

Speaker: Haolin Zou

May 9, 2025

Columbia University

Table of contents

Motivation and Backgrounds

Motivation

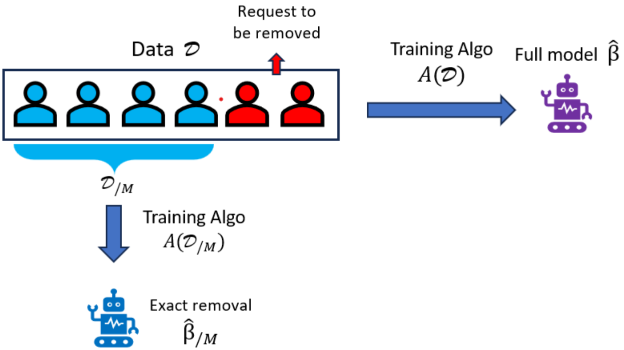
Consider a learning model:

Motivation

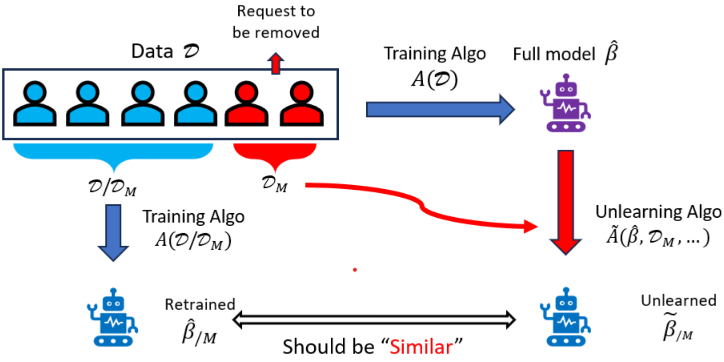
Consider a learning model:

Motivation

Now suppose a group of users request their data to be removed:

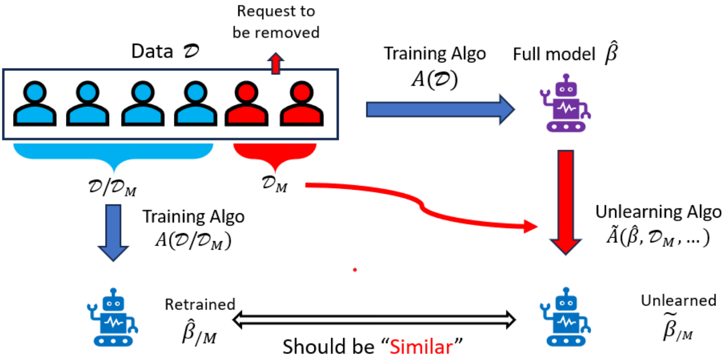


Now suppose a group of users request their data to be removed:



Motivation

Now suppose a group of users request their data to be removed:



Question: What is being **"similar"**?

Introduction

- Why removal?
  - 'Right to be forgotten' (California Consumer Privacy Act (CCPA), Act on the Protection of Personal Information (APPI) etc.)
  - Outlier removal.

Introduction

- Why removal?
  - 'Right to be forgotten' (California Consumer Privacy Act (CCPA), Act on the Protection of Personal Information (APPI) etc.)
  - Outlier removal.
- Why not retraining? - Expensive! (e.g. several weeks for GPT)

Introduction

- Why removal?
  - 'Right to be forgotten' (California Consumer Privacy Act (CCPA), Act on the Protection of Personal Information (APPI) etc.)
  - Outlier removal.
- Why not retraining? - Expensive! (e.g. several weeks for GPT)
- What has been done? - low dimensional $p \ll n$, gradient (GD) or Hessian (Newton) based methods.

Existing work focus on low dimensions

- Guo et al. (2019): (ϵ, δ)-Certified Removal, inspired by differential privacy, **randomized estimators**

Existing work focus on low dimensions

- Guo et al. (2019): (ϵ, δ)-Certified Removal, inspired by differential privacy, **randomized estimators**
  $\hat{\boldsymbol{\beta}}_{\setminus \mathcal{M}}$ and $\tilde{\boldsymbol{\beta}}_{\setminus \mathcal{M}}$ indistinguishable in distribution

Existing work focus on low dimensions

- Guo et al. (2019): (ϵ, δ)-Certified Removal, inspired by differential privacy, **randomized estimators**
  $\hat{\boldsymbol{\beta}}_{\setminus \mathcal{M}}$ and $\tilde{\boldsymbol{\beta}}_{\setminus \mathcal{M}}$ indistinguishable in distribution
  Theoretical guarantee for one Newton iteration, $p \ll n$

Existing work focus on low dimensions

- Guo et al. (2019): (ϵ, δ)-Certified Removal, inspired by differential privacy, **randomized estimators**
  $\hat{\boldsymbol{\beta}}_{\setminus \mathcal{M}}$ and $\tilde{\boldsymbol{\beta}}_{\setminus \mathcal{M}}$ indistinguishable in distribution
  Theoretical guarantee for one Newton iteration, $p \ll n$
- Sekhari et al. (2021): similar as above, adding an 'accuracy' metric using excess risk, $p \ll n$, one Newton iteration

Existing work focus on low dimensions

- Guo et al. (2019): (ϵ, δ)-Certified Removal, inspired by differential privacy, **randomized estimators**
  $\hat{\boldsymbol{\beta}}_{\setminus \mathcal{M}}$ and $\tilde{\boldsymbol{\beta}}_{\setminus \mathcal{M}}$ indistinguishable in distribution
  Theoretical guarantee for one Newton iteration, $p \ll n$
- Sekhari et al. (2021): similar as above, adding an 'accuracy' metric using excess risk, $p \ll n$, one Newton iteration
- Neel et al. (2021); Izzo et al. (2021): gradient descent based, $p \ll n$

Existing work focus on low dimensions

- Guo et al. (2019): (ϵ, δ)-Certified Removal, inspired by differential privacy, **randomized estimators**
  $\hat{\boldsymbol{\beta}}_{\setminus \mathcal{M}}$ and $\tilde{\boldsymbol{\beta}}_{\setminus \mathcal{M}}$ indistinguishable in distribution
  Theoretical guarantee for one Newton iteration, $p \ll n$
- Sekhari et al. (2021): similar as above, adding an 'accuracy' metric using excess risk, $p \ll n$, one Newton iteration
- Neel et al. (2021); Izzo et al. (2021): gradient descent based, $p \ll n$
- Xu et al. (2023): a comprehensive survey

Existing work focus on low dimensions

- Guo et al. (2019): (ϵ, δ)-Certified Removal, inspired by differential privacy, **randomized estimators**
  $\hat{\boldsymbol{\beta}}_{\setminus \mathcal{M}}$ and $\tilde{\boldsymbol{\beta}}_{\setminus \mathcal{M}}$ indistinguishable in distribution
  Theoretical guarantee for one Newton iteration, $p \ll n$
- Sekhari et al. (2021): similar as above, adding an 'accuracy' metric using excess risk, $p \ll n$, one Newton iteration
- Neel et al. (2021); Izzo et al. (2021): gradient descent based, $p \ll n$
- Xu et al. (2023): a comprehensive survey
- **Our Central Question:** Are existing unlearning methods reliable when $n, p \to \infty$ with $n/p \to \gamma_0 > 0$?

Our Key Findings

One step of Newton is **NOT** enough in high dimensions!

High Dimensional R-ERM, Certifiability and Accuracy

Formal Setup: Proportional High-dimensional R-ERM

- Dataset: $\mathcal{D} = \{(y_1, \mathbf{x}_1), \, (y_2, \mathbf{x}_2), \ldots, (y_n, \mathbf{x}_n)\}$.

Formal Setup: Proportional High-dimensional R-ERM

- Dataset: $\mathcal{D} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_n, \mathbf{x}_n)\}$.
- Model: Regularized-Empirical Risk Minimization (R-ERM)

$$\hat{\boldsymbol{\beta}} = A(\mathcal{D}) \triangleq \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i \le n} \ell(y_i | \boldsymbol{x}_i^\top \boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta})$$

Formal Setup: Proportional High-dimensional R-ERM

- Dataset: $\mathcal{D} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_n, \mathbf{x}_n)\}$.
- Model: Regularized-Empirical Risk Minimization (R-ERM)

$$\hat{\boldsymbol{\beta}} = A(\mathcal{D}) \triangleq \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i \leq n} \ell(y_i | \boldsymbol{x}_i^\top \boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta})$$

- Removal indices: $\mathcal{M} \subset \{1, 2, ..., n\}$, $|\mathcal{M}| = m$ (may increase with n)

Formal Setup: Proportional High-dimensional R-ERM

- Dataset: $\mathcal{D} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_n, \mathbf{x}_n)\}$.
- Model: Regularized-Empirical Risk Minimization (R-ERM)

$$\hat{\boldsymbol{\beta}} = A(\mathcal{D}) \triangleq \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i \leq n} \ell(y_i | \boldsymbol{x}_i^\top \boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta})$$

- Removal indices: $\mathcal{M} \subset \{1, 2, ..., n\}$, $|\mathcal{M}| = m$ (may increase with n)
- Exact removal:

$$\hat{\boldsymbol{\beta}}_{\setminus \mathcal{M}} = A(\mathcal{D} \setminus \mathcal{D}_{\mathcal{M}}) = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i \notin \mathcal{M}} \ell(y_i | \boldsymbol{x}_i^\top \boldsymbol{\beta}) + \lambda r(\boldsymbol{\beta});$$

Formal Setup: Proportional High-dimensional R-ERM

- Dataset: $\mathcal{D} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \dots, (y_n, \mathbf{x}_n)\}$.
- Model: Regularized-Empirical Risk Minimization (R-ERM)

$$\hat{\boldsymbol{\beta}} = A(\mathcal{D}) \triangleq \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i \leq n} \ell(y_i | \boldsymbol{x}_i^\top \beta) + \lambda r(\beta)$$

- Removal indices: $\mathcal{M} \subset \{1, 2, ..., n\}$, $|\mathcal{M}| = m$ (may increase with n)
- Exact removal:

$$\hat{\beta}_{\setminus \mathcal{M}} = A(\mathcal{D} \setminus \mathcal{D}_{\mathcal{M}}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i \notin \mathcal{M}} \ell(y_i | \boldsymbol{x}_i^\top \beta) + \lambda r(\beta);$$

High dimension

$p \to \infty, n \to \infty, p/n \equiv \gamma_0$ constant.

Certifiability and Accuracy

- Approximate removal:

$$\tilde{\boldsymbol{\beta}}_{\setminus \mathcal{M}} = \tilde{A}(\hat{\boldsymbol{\beta}}, \mathcal{D}_{\mathcal{M}}, ...)$$

- Need to add a perturbation $\boldsymbol{b}$ to obscure residual information about $\mathcal{D}_{\mathcal{M}}$ (similar to exponential & Gaussian mechanism in DP)

Certifiability and Accuracy

- Approximate removal:

$$\tilde{\boldsymbol{\beta}}_{\setminus \mathcal{M}} = \tilde{A}(\hat{\boldsymbol{\beta}}, \mathcal{D}_{\mathcal{M}}, ...)$$

- Need to add a perturbation $\boldsymbol{b}$ to obscure residual information about $\mathcal{D}_{\mathcal{M}}$ (similar to exponential & Gaussian mechanism in DP)
- A good removal: $\tilde{\boldsymbol{\beta}}_{\setminus \mathcal{M}} + \boldsymbol{b}$ similar to $\hat{\boldsymbol{\beta}}_{\setminus \mathcal{M}} + \boldsymbol{b}$

Certifiability and Accuracy

- Approximate removal:

$$\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} = \tilde{A}(\hat{\boldsymbol{\beta}}, \mathcal{D}_{\mathcal{M}}, ...)$$

- Need to add a perturbation $\boldsymbol{b}$ to obscure residual information about $\mathcal{D}_{\mathcal{M}}$ (similar to exponential & Gaussian mechanism in DP)
- A good removal: $\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b}$ similar to $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b}$
  - **Certifiability:** "Indistinguishability"
    (ϕ, ϵ)-Probabilistically-certified Approximate Removal (PAR)

$$e^{-\epsilon} \leq \frac{p(\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b}|\mathcal{D})}{p(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b}|\mathcal{D})} \leq e^{\epsilon} \;\; w.p. \geq 1 - \phi$$

Certifiability and Accuracy

- Approximate removal:

$$\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} = \tilde{A}(\hat{\boldsymbol{\beta}}, \mathcal{D}_{\mathcal{M}}, \ldots)$$

- Need to add a perturbation $\boldsymbol{b}$ to obscure residual information about $\mathcal{D}_{\mathcal{M}}$ (similar to exponential & Gaussian mechanism in DP)
- A good removal: $\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b}$ similar to $\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b}$
  - **Certifiability:** "Indistinguishability"
    (ϕ, ϵ)-Probabilistically-certified Approximate Removal (PAR)

$$e^{-\epsilon} \leq \frac{p(\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b} | \mathcal{D})}{p(\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b} | \mathcal{D})} \leq e^{\epsilon} \quad w.p. \geq 1 - \phi$$

  - **Accuracy:** Generalization Error Divergence (GED)

$$\mathrm{GED} := |\ell(y_{\mathrm{new}} | \boldsymbol{x}_{\mathrm{new}}^{\top}(\tilde{\boldsymbol{\beta}}_{\setminus\mathcal{M}} + \boldsymbol{b})) - \ell(y_{\mathrm{new}} | \boldsymbol{x}_{\mathrm{new}}^{\top}\hat{\boldsymbol{\beta}}_{\setminus\mathcal{M}})| \to_p 0$$

    for a new observation $(y_{\mathrm{new}}, \boldsymbol{x}_{\mathrm{new}})$.

Main Theoretical Results: One Newton Step Not Enough

Newton Method + Laplacian Perturbation

- Newton Method: initialize at $\tilde{\beta}^{(0)}_{\setminus\mathcal{M}} = \hat{\beta}$.

$$\tilde{\beta}^{(T)}_{\setminus\mathcal{M}} = \tilde{\beta}^{(t-1)}_{\setminus\mathcal{M}} - \left(\nabla^2 L_{\setminus\mathcal{M}}(\tilde{\beta}^{(t-1)}_{\setminus\mathcal{M}})\right)^{-1} \nabla L_{\setminus\mathcal{M}}(\tilde{\beta}^{(t-1)}_{\setminus\mathcal{M}})$$

  - $L_{\setminus\mathcal{M}}$: objective function for $\hat{\beta}_{\setminus\mathcal{M}}$

- Then add Isotropic Laplacian noise to ensure certifiability:

$$\tilde{\beta}^{(T)}_{\setminus\mathcal{M}} + \mathbf{b}, \quad \text{with } p(\mathbf{b}) \propto \exp\left(-\frac{\epsilon}{r_{t,n}}\|\mathbf{b}\|\right).$$

L2 error of Newton estimator

Lemma

Fix any number of Newton steps $T \geq 1$, $\epsilon > 0$, $m = o(n^{1/3})$:

$$\max_{t \leq T} ||\tilde{\beta}^{(T)}_{\setminus \mathcal{M}} - \hat{\beta}_{\setminus \mathcal{M}}||_2 = O_p\left(\left(\frac{m^3}{n}\right)^{2^{T-2}} \mathrm{polylog}(n)\right)$$

L2 error of Newton estimator

Lemma

Fix any number of Newton steps $T \geq 1$, $\epsilon > 0$, $m = o(n^{1/3})$:

$$\max_{t \leq T} ||\tilde{\beta}^{(T)}_{\setminus \mathcal{M}} - \hat{\beta}_{\setminus \mathcal{M}}||_2 = O_p\left(\left(\frac{m^3}{n}\right)^{2^{T-2}} \mathrm{polylog}(n)\right)$$

Main Assumptions:

- L is strongly-convex
- Smoothness, Gaussian design $\boldsymbol{X}$, ℓ, r and their derivatives have polynomial growth, bounded SNR.

Main Theorem: Certifiability and Accuracy

Theorem

For any fixed number of Newton steps $t \geq 1$*, if* $p(\boldsymbol{b}) \propto e^{-\frac{\epsilon}{r_{t,n}}||\boldsymbol{b}||}$ *with*

$$r_{t,n} \simeq \left(\frac{m^3}{n}\right)^{2^{t-2}} \mathrm{polylog}(n),$$

and $|\mathcal{M}| = m = o(n^{1/3})$*: then under high dimensions* ($p \propto n$)*:*

- **Certifiability:** $\tilde{\beta}^{(T)}_{\setminus\mathcal{M}} + \boldsymbol{b}$ *achieves* (ϕ_n, ϵ)*-PAR with* $\phi_n \to 0$*.*
- **Accuracy:**

$$\mathrm{GED} = O_p\left(\frac{\sqrt{mp}}{\epsilon} r_{t,n} \mathrm{polylog}(n)\right)$$

Implications and comparisons

- We need T to satisfy

$$T \geq 1 + \log_2 \left(\frac{\alpha + 1}{1 - 3\alpha} \right)$$

  where $\alpha := \log(m)/\log(n) < \frac{1}{3}$
- In low dimensions, $t = 1$ Newton step suffices
- However, under high dimensions (when $p \propto n$) even for $m = 1$ a single step leads to too high a noise level, but two steps are enough.

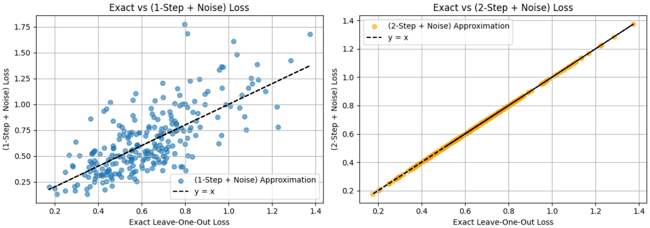
Numerical Experiments

One step is not enough in high dimensions

Numerical Experiment:

- Logistic + Ridge model, $m = 1$
- Plot $\ell(y_i|\boldsymbol{x}_i^\top(\tilde{\beta}_{\setminus i}^{(t)} + \boldsymbol{b}))$ against $\ell(y_i|\boldsymbol{x}_i^\top \hat{\beta}_{\setminus i})$
- Left: one step Newton. Right: two steps.

$n = 250,\ p = 500,\ df/p = 0.13,\ \lambda = 1$



Conclusions and Future Work

- We propose a certifiable data removal (machine unlearning) framework (certifiability and accuracy) that is suitable for high-dimensional settings.
- The proposed perturbation–based Newton method requires more than one update step to be both certified and accurate.
- Theoretical analysis under high dimensions and numerical experiments support the need for multiple Newton steps.
- Future work: extensions to non-smooth models, alternative forms of perturbation, gradient descent, efficient sequential removal, etc.

References

C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

Z. Izzo, M. Anne Smart, K. Chaudhuri, and J. Zou. Approximate data deletion from machine learning models. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2008–2016. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/izzo21a.html.

S. Neel, A. Roth, and S. Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pages 931–962. PMLR, 2021.

A. Sekhari, J. Acharya, G. Kamath, and A. T. Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

H. Xu, T. Zhu, L. Zhang, W. Zhou, and P. S. Yu. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1), Aug. 2023. ISSN 0360-0300. doi: 10.1145/3603620. URL `https://doi.org/10.1145/3603620`.