Hessian-Free Online Certified Unlearning

Xinbao Qiao[1] Meng Zhang[1] Ming Tang[2] Ermin Wei[3]

[1]Zhejiang University [2]Southern University of Science and Technology [3]Northwestern University

Empirical Risk Minimizer Rewighting

Original Empirical Risk Minimizer on the entire dataset:

$$\hat{\mathbf{w}} := \underset{\mathbf{w}\in\mathcal{W}}{\operatorname{argmin}} F_{\mathcal{S}} = \underset{\mathbf{w}\in\mathcal{W}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \ell\left(\mathbf{w}; z_i\right).$$

Retrained Empirical Risk Minimizer without sample u:

$$\hat{\mathbf{w}}^{-u} := \underset{\mathbf{w}\in\mathcal{W}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \ell\left(\mathbf{w}; z_i\right) + \omega\ell(\mathbf{w}; u).$$

Previous method:

$$\hat{\mathbf{w}}^{-u} - \hat{\mathbf{w}} \approx \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\nabla^2\ell(\hat{\mathbf{w}}; z_i)\right)^{-1}\nabla\ell(\hat{\mathbf{w}}; u).$$

Model Update Trajectory Rewighting

Original Model Update Trajectory on the entire dataset:

$$\mathbf{w}_{e,b+1} \leftarrow \mathbf{w}_{e,b} - \frac{\eta_{e,b}}{|\mathcal{B}_{e,b}|}\sum_{i\in\mathcal{B}_{e,b}}\nabla\ell(\mathbf{w}_{e,b}; z_i).$$

Retrained Model Update Trajectory without sample u:

$$\mathbf{w}^{-u}_{e,b+1} \leftarrow \mathbf{w}_{e,b} - \frac{\eta_{e,b}}{|\mathcal{B}_{e,b}|}\sum_{i\in\mathcal{B}_{e,b}}\left(\nabla\ell(\mathbf{w}_{e,b}; z_i) + 1_{\mathrm{b}=\mathrm{b}(u)}\omega\nabla\ell(\mathbf{w}_{e,b}; u)\right).$$

Proposed method:

$$\mathbf{w}^{-u}_{E,B} - \mathbf{w}_{E,B} \approx \sum_{e=0}^{E}\frac{\eta_{e,b(u)}}{|\mathcal{B}_{e,b(u)}|}\hat{\mathbf{H}}_{E,B-1\to e,b(u)+1}\cdot\nabla\ell(\mathbf{w}_{e,b(u)}; u)$$

where $\hat{\mathbf{H}}_{E,B-1\to e,b(u)+1} = (\mathbf{I} - \eta_{E,B-1}\mathbf{H}_{E,B-1})\cdots(\mathbf{I} - \eta_{e,b(u)+1}\mathbf{H}_{e,b(u)+1})$

INTRODUCTION

Machine unlearning strives to uphold data owners' right to be forgotten by enabling models to selectively forget specific data. In this context, **certified unlearning** provides theoretical guarantees, aiming to make the output distribution of the unlearned algorithm indistinguishable from that of retraining:

$$\mathrm{P}(\bar{\Omega}(\Omega(\mathcal{S}), \mathcal{T}(\mathcal{S})) \leq e^{\epsilon}\mathrm{P}(\bar{\Omega}(\Omega(\mathcal{S}\backslash U), \emptyset))) + \delta, \text{ and}$$

$$\mathrm{P}(\bar{\Omega}(\Omega(\mathcal{S}\backslash U), \emptyset) \leq e^{\epsilon}\mathrm{P}(\bar{\Omega}(\Omega(\mathcal{S}), \mathcal{T}(\mathcal{S})) + \delta$$

Recent advances suggest that certified unlearning can be achieved by pre-computing and storing statistics derived from second-order information, and implementing unlearning through Newton-style updates, as illustrated in the dark blue region on the left side of the figure.

However, existing algorithms have several fundamental challenges:

- existing algorithms lack the capability to handle multiple deletion requests in online because they require explicit precomputing Hessian or its inverse.
- the learned model needs to be empirical risk minimizer,
- with assumption of convexity to ensure the invertibility of Hessian.

METHODOLOGY

As illustrated in the orange region on the right side of the figure,

1. **Hessian-Vector Product** enables us to exploit the HVP technique to compute proposed mehtod without explicitly computing Hessian, which has the same order of magnitude as computing a gradient.
2. **Additivity** enables proposed mehtod to handle multiple deletion requests in an online manner. Previous Hessian-based methods, lacking additivity, fail to utilize HVP as the forgetting sample and unleaned model are unknown before a deletion request arrives.
3. **No need to Invert the Hessian matrix** allows the proposed method to avoid assuming convexity in the theoretical analysis, thereby extending it to non-convex settings.
4. **Pre-computing and pre-storage** of each sample's influence using the proposed method allow for nearly instantaneous data removal through simple vector additions upon receiving a deletion request.

THEORY

Theorem 1: Unlearning Guarantee. For any forgetting dataset U, the bound between the unlearning model and retrained model is,

$$\left\|\mathbf{w}^{-U}_{E,B} - \overline{\mathbf{w}}^{-U}_{E,B}\right\| \leq \mathcal{O}\left(\eta G m \rho^{nE/|\mathcal{B}|}\right)$$

Theorem 2: Generalization Guarantee. Unlearned model's excess risk bound is,

$$F\left(\tilde{\mathbf{w}}^{-U}_{E,B}\right) - \mathbb{E}\left[F\left(\mathbf{w}^*\right)\right] = \mathcal{O}\left(\frac{4L^2}{\lambda(n-m)} + \rho^{nE/|\mathcal{B}|}\left(\frac{2L^2}{\lambda} + \frac{2mL\eta G\sqrt{d}\sqrt{\ln(1/\delta)}}{\epsilon}\right)\right)$$

Theoritical Results:

Table 1: **Summary of Results.** Here, d, n, and m denote the model parameters size, training dataset size, and forgetting dataset size, respectively. The unlearning guarantee is derived under the same regularity conditions as prior second-order unlearning studies, with ρ being a constant less than 1.

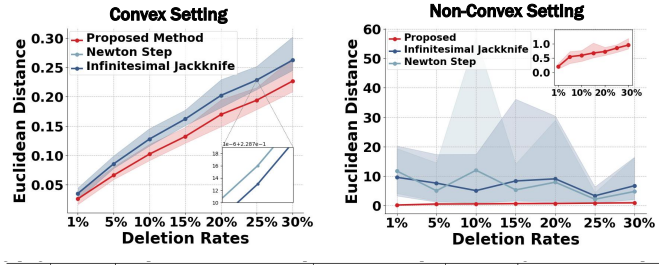
Methods	Computation time	Storage	Unlearning time	Unlearning guarantee
Influence Function[1]	$\mathcal{O}(nd^2)$	$\mathcal{O}(d^2)$	$\mathcal{O}(d^3 + md^2 + md)$	$\mathcal{O}(m^2/n^2)$
Infinitesimal Jackknife[2]	$\mathcal{O}(d^3 + nd^2)$	$\mathcal{O}(d^2)$	$\mathcal{O}(d^2 + md)$	$\mathcal{O}(m^2/n^2)$
Proposed method	$\mathcal{O}(n^2d)$	$\mathcal{O}(nd)$	$\mathcal{O}(md)$	$\mathcal{O}(m\rho^n)$

Reference

[1] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. **Remember what you want to forget: Algorithms for machine unlearning.** In Advances in Neural Information Processing Systems 34, NeurIPS, 2021.

[2] Vinith M. Suriyakumar, Ashia C. Wilson, et al. **Algorithms that approximate data removal: New results and limitations.** In Advances in Neural Information Processing Systems 35, NeurIPS, 2022.

EXPERIMENT



	Model	Unlearning Computation Runtime (Sec)	Unlearning Computation Speedup	Pre-Computaion Runtime (Sec)	Storage (GB)	Test Accuracy (%) Unlearned model
NS	LR	5.09×10^2	1.34×	2.55×10^3	0.23	86.25 (-1.50)
	CNN	5.81×10^3	0.12×	2.91×10^4	1.78	83.50 (-10.25)
IJ	LR	0.23×10^2	29.64×	2.55×10^3	0.23	86.25 (-1.50)
	CNN	1.91×10^2	3.63×	2.91×10^4	1.78	82.75 (-11.00)
HF	LR	**0.0073**	**93,376**×	$\mathbf{2.20\times10^2}$	**0.03**	**87.50** (-0.25)
	CNN	**0.0268**	**25,821**×	$\mathbf{5.34\times10^2}$	**0.08**	**91.50** (-2.25)

CONTACT:

Email: xinbaoqiao@zju.edu.cn

Homepage: xinbaoqiao.github.io

ICLR

ZJU-UIUC INSTITUTE

Zhejiang University-University of Illinois at Urbana-Champaign Institute

浙江大学伊利诺伊大学厄巴纳香槟校区联合学院