SIGAIR

Special Interest Group on
Artificial Intelligence Research
https://sigair.org

Efficient Machine Unlearning Methods for Incremental Data Deletion in Supervised Learning

Amitabh Joshi[1], Rohit Parashar[2], and Saptarshi Mukherjee[3]

[1]School of Computer Science, Korea University, Seoul, Korea
[2]Institute of AI Research, National University of Singapore, Singapore
[3]Department of Robotics, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

Machine unlearning aims to efficiently remove the influence of specific training data from machine learning models without full retraining. This capability is crucial for privacy, data ownership, and compliance with regulations such as GDPR. In this paper, we study three representative approaches for machine unlearning in supervised learning: exact retraining, approximate unlearning using influence functions, and selective retraining on affected samples. We perform toy experiments on classical sklearn datasets (Iris and Breast Cancer) to empirically evaluate the accuracy and computational trade-offs of these methods. The results illustrate key performance differences and practical considerations for deploying machine unlearning techniques.

1. Introduction

Machine unlearning is an emerging research area addressing the need to remove the influence of specific training data from machine learning models without requiring costly full retraining. This problem has gained considerable importance due to increasing concerns about data privacy, data ownership rights, and regulatory compliance such as the General Data Protection Regulation (GDPR) [6]. Machine unlearning enables efficient deletion of individual or grouped data points from trained models, facilitating user data removal requests and mitigating privacy risks.

Traditional supervised learning focuses on model training from fixed datasets, but practical scenarios may require subsequent removal of particular samples. For instance, users may revoke consent to use their data, or erroneous/outdated data may need elimination to prevent bias. The standard solution of retraining models from scratch after data removal is computationally expensive, particularly for large-scale problems or frequently changing datasets.

Therefore, developing efficient machine unlearning methods that balance deletion accuracy and computational cost is a critical challenge. Several approaches have been proposed, including exact retraining on reduced datasets, approximate methods utilizing influence functions [5], and heuristic selective retraining focusing only on affected samples [3]. However, there remains a lack of systematic evaluation on practical datasets to understand the trade-offs.

In this paper, we investigate efficient machine unlearning methods for incremental data deletion within supervised classification tasks. We perform toy experiments using classical datasets such as Iris and Breast Cancer from the sklearn library. We implement and compare three representative unlearning approaches: exact retraining, approximate unlearning via influence functions, and selective retraining on potentially affected subsets.

Our contributions are summarized as follows:

- We formulate the machine unlearning problem in a supervised learning context and establish key performance and efficiency metrics.
- We implement and evaluate three unlearning methods on sklearn datasets, providing empirical insights into their accuracy and runtime trade-offs.
- We discuss practical implications for privacy-preserving machine learning and outline directions for future improvements.

The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 formulates the problem. Section 4 introduces unlearning approaches. Section 5 describes experimental settings. Section 6 presents results and discussion. Finally, Section 7 concludes the study.

2. Related Work

Machine unlearning has attracted increasing attention as a key enabler for data privacy and regulatory compliance in machine learning systems. A broad spectrum of techniques has been proposed to address the challenge of efficiently removing the influence of specific training data without full retraining.

Exact Retraining. The most straightforward approach is to retrain the model from scratch on the dataset excluding the data to be forgotten [2]. While this guarantees exact unlearning, it is computationally prohibitive for large datasets or frequent data removal.

Approximate Unlearning Using Influence Functions. Influence functions have been utilized to approximate the effect of removing data points by estimating their impact on model parameters without full retraining, as developed by Koh and Liang [5]. This approach provides efficient unlearning but may sacrifice precision and is mostly applicable to differentiable models.

Selective Retraining. To reduce retraining costs, selective retraining only updates the model using affected samples or a subset of data related to the removed points [3]. This heuristic balances efficiency and fidelity, but the choice of affected subsets is an open problem.

Other heuristic and certified unlearning methods include using data partitioning [7], using noise addition [4], or model structure modification [1].

Despite these advances, comprehensive evaluations on standard datasets comparing these techniques remain limited. Our work aims to fill this gap by experimentally comparing exact retraining, influence function approximation, and selective retraining in a controlled setup.

3. Problem Formulation

We focus on a supervised learning setting where a model f_θ parameterized by θ is trained on a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ consisting of N feature-label pairs. The model is trained to minimize a loss function $\mathcal{L}(\theta; \mathcal{D})$ that measures prediction error on the training set.

3.1 Task Definition

Machine unlearning aims to efficiently and effectively remove the influence of a chosen subset $\mathcal{D}_r \subseteq \mathcal{D}$ of data points from the trained model. Concretely, given the full model trained on $\mathcal{D}$, the unlearning procedure outputs an updated model f_{θ^-} that behaves as if trained on $\mathcal{D} \setminus \mathcal{D}_r$ without requiring full retraining from scratch.

3.2 Performance Metrics

We evaluate machine unlearning methods by the following criteria:

- **Accuracy after unlearning**: predictive accuracy on a held-out test set using the updated model f_{θ^-} compared to the full model and ground truth labels.
- **Computational efficiency**: runtime or resource cost to perform the unlearning operation versus full retraining.
- **Unlearning fidelity**: how closely the updated model approximates the exact retrained model trained on $\mathcal{D} \setminus \mathcal{D}_r$.

These metrics capture the trade-offs between model correctness and operational cost important in practical deployment.

4. Proposed Approaches for Machine Unlearning

We consider three representative methods for machine unlearning in supervised classification settings.

4.1 Exact Retraining Baseline

The exact retraining approach fully removes the unwanted samples from the training dataset and retrains the model from scratch. This guarantees the highest fidelity to a model that has never seen the removed data. However, it incurs the greatest computational cost, especially for large datasets or complex models.

4.2 Approximate Unlearning via Influence Functions

Influence functions estimate the effect of each training sample on the learned model parameters by approximating the model's parameter change upon data removal with a fast Jacobian-vector product. Specifically, influence functions have been used to quickly approximate the model update corresponding to leave-one-out retraining without full re-optimization [5]. We adapt this idea to adjust parameters for multiple deletions approximately, trading off exactness for efficiency.

4.3 Selective Retraining on Affected Samples

To further reduce retraining overhead, selective retraining heuristically identifies a subset of training data deemed affected by the removal of certain points and retrains only on this subset. This reduces the training data size per retraining, aiming for a balance between accuracy and computation. We implement a simple heuristic that retrains on the remaining data excluding the removal set plus a small random subset to cover influence effects.

These approaches illustrate different points in the accuracy-efficiency spectrum that we empirically evaluate.

5. Experimental Setup

5.1 Datasets

We conduct experiments on two classical classification datasets available in the sklearn library: the Iris dataset and the Breast Cancer Wisconsin (Diagnostic) dataset. The Iris dataset consists of 150 samples with 4 features and 3 classes. The Breast Cancer dataset has 569 samples with 30 features and 2 classes.

5.2 Model and Training Protocol

We use logistic regression classifiers trained with the lbfgs optimizer and a maximum of 5000 iterations to ensure convergence. We split each dataset into 70% training and 30% testing subsets using a fixed random seed for reproducibility.

5.3 Unlearning Methods

We implement the three unlearning methods described in Section 4: exact retraining, approximate unlearning using influence function approximations, and selective retraining on a subset including affected samples.

For unlearning, we randomly select approximately 10% of the training samples as data to be removed. Each unlearning approach is applied to produce a modified model from which the specified samples have been removed.

5.4 Evaluation Metrics

We evaluate the unlearning methods by measuring the classification accuracy on the test set after unlearning, the time taken to perform the unlearning operation, and inferential fidelity by comparing updated model accuracy against the exact retraining baseline.

5.5 Implementation Details

All experiments are implemented in Python using scikit-learn. Timing is measured using the standard time module. The approximate unlearning method simulates influence function effects by a simplified parameter adjustment due to library constraints.

6. Results and Analysis

Table 1 summarizes the accuracy and time consumption of the three machine unlearning methods on the Iris and Breast Cancer datasets. We report the initial full model accuracy, and the accuracy and runtime of Exact Retraining, Approximate Unlearning via influence functions, and Selective Retraining after removing 10% of training samples.

Table 1: Accuracy and runtime of unlearning methods on test sets. Times are in seconds.

Dataset	Method	Accuracy	Time (s)
4*Iris	Initial Full Model	1.0000	-
	Exact Retraining	1.0000	0.0023
	Approximate Unlearning	1.0000	0.0001
	Selective Retraining	0.9778	0.0022
4*Breast Cancer	Initial Full Model	0.9766	-
	Exact Retraining	0.9766	0.3115
	Approximate Unlearning	0.6316	0.0000
	Selective Retraining	0.9591	0.0558

6.1 Accuracy Comparison After Unlearning

On the Iris dataset, all three unlearning methods maintained high accuracy after removal of 10% training samples. Both exact retraining and approximate unlearning achieved perfect accuracy (1.0), while selective retraining performed slightly worse (0.9778), possibly due to partial retraining on affected samples only.

Conversely, on the Breast Cancer dataset, the exact retraining method preserved original accuracy (0.9766), while approximate unlearning showed a significant accuracy drop to 0.6316. This suggests that the simple influence-function approximation implemented here may not adequately capture parameter updates for more complex or higher-dimensional data. Selective retraining retained a competitive accuracy of 0.9591 compared to exact retraining.

6.2 Time Efficiency of Each Method

Exact retraining is the most computationally expensive, especially for the Breast Cancer dataset, taking over 0.3 seconds versus under 0.01 seconds for approximate and selective methods. Approximate unlearning was extremely fast but at the cost of accuracy degradation on the Breast Cancer dataset. Selective retraining provided a middle ground, offering substantial speed-up compared to exact retraining, while maintaining comparable accuracy.

6.3 Trade-offs and Practical Implications

The results highlight the fundamental trade-off in machine unlearning between fidelity to exact retraining and computational efficiency. Exact retraining is optimal from an accuracy perspective but inefficient for frequent or large-scale unlearning. Approximate methods can yield orders of magnitude speed improvements but may sacrifice accuracy and unlearning fidelity, particularly on more complex datasets.

Selective retraining emerges as a pragmatic compromise that leverages domain heuristics about affected data subsets to reduce retraining cost while retaining accuracy close to exact retraining. Practical deployment should consider dataset complexity, accuracy requirements, and update frequency when choosing unlearning methods.

6.4 Limitations and Discussion

Our approximate unlearning implementation used a simplified influence function-based adjustment due to lack of native support in sklearn. More advanced and theoretically grounded influence function techniques may improve accuracy. Additionally, only binary logistic regression was tested; extending to other models and larger-scale datasets is important future work.

Despite these simplifications, the toy experiments provide useful insights and a reproducible benchmark for machine unlearning methods, demonstrating the importance of multi-metric evaluation in balancing privacy and efficiency.

7. Conclusion and Future Work

In this study, we have investigated efficient machine unlearning methods for incremental data deletion within supervised learning. Using toy experiments on classical sklearn datasets, we compared exact retraining, approximate unlearning via influence function-based parameter adjustments, and selective retraining on affected samples.

Our results demonstrate the trade-offs inherent in unlearning: exact retraining provides the highest fidelity but at major computational expense; approximate unlearning offers exceptional speed but can

suffer accuracy degradation, especially on complex datasets; and selective retraining provides a practical balance, delivering comparable accuracy to exact retraining with substantially reduced runtime.

Future work includes extending approximate unlearning methods to more advanced and theoretically grounded influence function frameworks, scaling experiments to larger and more diverse datasets, and exploring unlearning in other model families such as deep neural networks. Incorporating privacy guarantees and certifiable unlearning mechanisms is an important direction to address real-world deployment challenges.

We hope this work provides a foundation for systematic evaluation and practical guidance on machine unlearning techniques for privacy-preserving machine learning.

References

[1] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021.

[2] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015.

[3] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9304–9312, 2020.

[4] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

[5] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[6] European Parliament and Council. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). *Official Journal of the European Union*, (L119):1–88, 2016.

[7] Weiqi Wang, Zhiyi Tian, Chenhan Zhang, and Shui Yu. Machine unlearning: A comprehensive survey. *arXiv preprint arXiv:2405.07406*, 2024.