A Principled Framework for Parameter Estimation in Federated Unlearning

Antonio Balordi[a, b,*]

[a]CASD - Italian Defense University, Rome, Italy
[b]Models and Algorithms for Data and Text Mining Laboratory (MADLab), University of Milano-Bicocca, Milan, Italy
ORCID (Antonio Balordi): https://orcid.org/0009-0009-6710-8087

Abstract. Privacy regulations require the erasure of data from deep learning models, a challenge that is amplified in Federated Learning, where retraining from scratch is often infeasible. My research work aims to define an efficient Federated Unlearning framework based on information theory, modeling leakage as a parameter estimation problem. The idea is to use second-order Hessian information to identify and selectively reset only parameters most sensitive to the data being forgotten, followed by minimal federated retraining. This model-agnostic approach supports client and categorical unlearning without requiring server access to raw client data after initial information aggregation. Evaluations on benchmark datasets demonstrate strong privacy (MIA success near random, categorical knowledge erased), high performance (Normalized Accuracy concerning the retrained benchmarks of ≈ 0.9), and significant efficiency over full retraining, offering a practical solution for data forgetting in FL.

Keywords. Federated Unlearning · Parameter estimation · Right To Be Forgotten · Parameter Estimation · Hessian computation.

1 Introduction

Deep learning models, fueled by vast and ever-growing datasets, are achieving unprecedented success across diverse domains, revolutionizing industries from healthcare to finance. However, this deep dependency creates a pressing privacy paradox: the very models we rely on for their learning capabilities can become liabilities, inadvertently memorizing and potentially exposing sensitive user information [13, 3]. With stringent international privacy regulations, such as GDPR and CCPA championing data sovereignty and firmly establishing individuals' *Right To Be Forgotten* [16, 4], the imperative for machine learning systems to unlearn specific data is no longer an academic concern but a critical operational and ethical need for organizations worldwide.

Consider a collaborative diagnostic model for detecting a rare disease, trained using Federated Learning (FL) across a consortium of hospitals. One of the hospitals, which initially consented to their medical imagery being used, later decided to withdraw from the study, invoking their "Right to Be Forgotten". The consortium is now legally and ethically obligated to ensure the global model no longer retains any information learned from this specific sensitive data.

This demand for controlled forgetting presents a relevant challenge, particularly when amplified by the decentralized architecture of Federated Learning. In FL environments, data is intentionally kept decentralized across multiple client devices to enhance privacy and security. While this is a strength, it complicates the task of data removal. The most straightforward solution, which involves completely retraining the global model from the ground up without the data of the requesting client, is an economically and computationally expensive non-starter in most practical federated learning (FL) systems. Its prohibitive costs, especially concerning communication and computation, become insurmountable in dynamic settings with high client turnover or when dealing with large-scale models and datasets [12, 2, 14]. This doctoral research project ventures into this complex and crucial terrain. I aim to develop innovative Federated Unlearning (FU) approaches that effectively reconcile the stringent demands of data privacy and erasure with the operational realities and inherent efficiencies expected of distributed artificial intelligence.

In my work, I introduced an innovative framework that involves defining information leakage as a parameter estimation problem. My research proposes a targeted and efficient algorithm for unlearning. By utilizing second-order Hessian diagonal information (obtained after initial training) to pinpoint the parameters most influential in unlearning the data, the method selectively resets only this critical subset. A brief, minimal retraining step completes the process. This approach demonstrates considerable versatility, as it applies to various network architectures and handles both client-level and categorical unlearning needs. Designed for the federated context, it facilitates server-side unlearning with significantly reduced retraining overhead while upholding client data privacy through reliance on aggregated Hessian statistics.

In Balordi et al. [1], we presented the mathematical framework and validated it through a series of experiments.

2 Related Works

Federated Unlearning (FU) extends Machine Unlearning (MU) to federated learning (FL) environments, but faces unique challenges:

- **Distributed Data & Privacy Constraints:** Centralized MU techniques like SISA (sharded retraining) or influence functions do not directly translate to FL. In FL, data remain on clients, communication is costly, and direct data access is prohibited, making

[*] Email: antonio.balordi@stud.unicasd.it

shard-based retraining or expensive influence-function computations impractical [2, 9, 6].

- **Retraining-Based/Approximation Methods:** Methods that approximate full retraining (e.g., via knowledge distillation) still require multiple communication rounds and heavy computation. Their unlearning guarantees are hard to quantify, and accuracy can degrade when unlearning large data fractions or in complex models [17, 18, 6].
- **Gradient/Update Manipulation:** Techniques that attempt to "undo" forgotten clients' contributions by inverting or perturbing gradients are largely heuristic. They may force mispredictions but do not reliably erase memorized information, leaving models vulnerable to membership inference or reconstruction attacks [7, 3].
- **Parameter Masking/Perturbation:** Approaches that prune or add noise to parameters associated with forgotten data struggle to identify the correct subset of weights, as dependencies in deep FL models are complex. Tuning these modifications is non-trivial and often degrades utility [5, 10, 15].

Overall, FU must account for FL's distributed nature, communication constraints, and strict privacy requirements [11, 8]. Existing FU categories—retraining approximations, gradient manipulations, and parameter perturbations—each have significant practical and theoretical limitations when applied to large-scale, deep federated models.

3 A Precision-Based Unlearning Framework

The prohibitive cost of full retraining in Federated Learning (FL) demands a more intelligent approach to unlearning. Our research proposes a novel framework that moves away from brute-force methods and instead treats unlearning as a form of precise *model surgery*. The core philosophy is to operate directly on the final, fully-trained model to surgically identify, erase, and then reintegrate knowledge, all while respecting the foundational privacy and communication constraints of the federated environment. This process is structured into three distinct, sequential steps.

3.1 Quantifying Parameter Influence via Second-Order Statistics

The first and most critical challenge is to determine which parts of the model have been most influenced by the data we now wish to forget (the *target dataset*). Accomplishing this without violating client privacy is paramount.

The approach sidesteps the need for direct data access by leveraging second-order information. We task the clients with a specific local computation. For each client, its local dataset is partitioned into two subsets: the data to be forgotten ($D^{(f)}$) and the data to be retained ($D^{(r)}$). The client then computes the diagonal of the Hessian of the loss function separately for each subset. Conceptually, the Hessian diagonal measures the *curvature of the loss landscape* for each parameter. A high value indicates that a parameter is highly sensitive and critical to minimizing loss for that specific data. By computing this for both the "forget" and "retain" data, clients generate a rich, privacy-safe summary of parameter importance.

These two Hessian diagonals per client are then sent to the central server. The server aggregates them to form a global view, allowing it to calculate a final *information score* for each parameter in the model. This score effectively quantifies how much more *influential* a given parameter is for the forgotten data compared to the retained data. This step directly addresses the *Inscrutable Local Datasets* challenge, widely discussed in the literature [12], by replacing risky data sharing with the communication of aggregated statistical summaries.

3.2 Targeted Knowledge Erasure through Parameter Resetting

With the information scores in hand, the server can now pinpoint the exact locations in the model where knowledge of the target dataset is most heavily encoded. Instead of attempting a complex and often unstable reversal of past gradient updates—a process complicated by the *Iterative Learning* nature of FL—we perform a direct and decisive action.

We introduce a key hyperparameter, the *removal percentage* ($\alpha_{removal}$), which defines the fraction of parameters to be targeted. Within each layer of the model, the server identifies the parameters with the highest information scores. These parameters—the ones most "responsible" for memorizing the target data—are then *reset to their original, random initial values* from before training began. This action effectively induces amnesia in the most critical parts of the network, severing the learned connections that were formed based on the now-unwanted data. This approach is highly flexible, as the target dataset can be defined to correspond to a specific client, a subset of samples, or an entire data class, thus addressing all primary FU objectives.

3.3 Efficient Model Recovery via Focused Fine-Tuning

Resetting parameters inevitably results in a temporary decline in the model's overall performance. The final step is to recover this performance efficiently, without re-introducing the influence of the forgotten data. A complete retraining epoch on the remaining data would be computationally expensive and slow.

Our framework employs a far more efficient method. We utilize a custom wrapper module, which we term `RelearnNet`, to manage the retraining process. This module *freezes all the parameters that were not reset*, preserving the vast majority of the knowledge learned from the retained data. It then exposes *only* the newly reset parameters as trainable. A single, brief federated fine-tuning epoch is then conducted. During this epoch, clients use only their retained data ($D^{(r)}$) to compute updates, and only the reset parameters are adjusted.

The underlying assumption, validated by our experiments, is that the model's overall structure remains close to an optimal state for the retained data. The reset parameters need to be "re-integrated" into this existing, well-trained structure. A single, focused epoch is sufficient for them to learn appropriate values and restore the model's predictive accuracy, completing the unlearning process with a minimal fraction of the cost of full retraining. This addresses the challenge of efficiency head-on, making unlearning a practical operation rather than a catastrophic expense.

4 Future Works

We plan to extend our framework in several directions:

- Adapting it to fully non-deterministic federated training, leveraging richer second-order information (e.g., block-diagonal or full-Hessian approximations).
- Explore adaptive hyperparameter tuning for unlearning rates and fine-tuning durations.

- Incorporate formal privacy accounting (such as differential privacy) to achieve provable guarantees and optimize the privacy–performance trade-off in federated unlearning.

References

[1] A. Balordi, L. Manini, F. Stella, and A. Merlo. Tackling federated unlearning as a parameter estimation problem, 2025. URL https://arxiv.org/abs/2508.19065.

[2] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021.

[3] N. Carlini, C. Liu, Úlfar Erlingsson, J. Kos, and D. Song. The secret sharer: Evaluating and testing unintended memorization in neural networks, 2019. URL https://arxiv.org/abs/1802.08232.

[4] C. C. Code. Assembly bill no. 1008, 2023.

[5] A. Golatkar, A. Achille, and S. 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.

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

[7] A. Halimi, S. Kadhe, A. Rawat, and N. Baracaldo. Federated unlearning: How to efficiently erase a client in fl? *arXiv preprint arXiv:2207.05521*, 2022.

[8] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

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

[10] Z. Ma, Y. Liu, X. Liu, J. Liu, J. Ma, and K. Ren. Learn to forget: Machine unlearning via neuron masking. *IEEE Transactions on Dependable and Secure Computing*, 20(4):3194–3207, 2022.

[11] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[12] N. Romandini, A. Mora, C. Mazzocca, R. Montanari, and P. Bellavista. Federated unlearning: A survey on methods, design guidelines, and evaluation metrics. *IEEE Transactions on Neural Networks and Learning Systems*, page 1–21, 2024. ISSN 2162-2388. doi: 10.1109/tnnls.2024.3478334. URL http://dx.doi.org/10.1109/TNNLS.2024.3478334.

[13] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models, 2017. URL https://arxiv.org/abs/1610.05820.

[14] S. Su, B. Li, and X. Xue. One-shot federated learning without server-side training. *Neural Networks*, 164:203–215, July 2023. ISSN 0893-6080. doi: 10.1016/j.neunet.2023.04.035. URL http://dx.doi.org/10.1016/j.neunet.2023.04.035.

[15] A. Thudi, G. Deza, V. Chandrasekaran, and N. Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319. IEEE, 2022.

[16] P. Voigt and A. Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer Cham, 01 2017. ISBN 978-3-319-57958-0. doi: 10.1007/978-3-319-57959-7.

[17] C. Wu, S. Zhu, and P. Mitra. Federated unlearning with knowledge distillation. *arXiv preprint arXiv:2201.09441*, 2022.

[18] G. Ye, T. Chen, Q. V. Hung Nguyen, and H. Yin. Heterogeneous decentralised machine unlearning with seed model distillation. *CAAI Transactions on Intelligence Technology*, 9(3):608–619, 2024.