



Figure R1: Example of the original auxiliary dataset (the first row), their noisy counterparts (the middle row), and directly construct data without an auxiliary dataset [R11] (the last row). @Reviewer iCGz

Table R1: Evaluating mixing unlearned data in the clean dataset on CIFAR10. The results demonstrate that mixing the unlearned samples into the constructed uploaded data for incremental learning negatively impacts the unlearning effect, as reflected by the increasing backdoor accuracy, but the model utility keeps. @Reviewer BZGC

On CIFAR10	Mixed 0% of Unlearned Data	2%	4%	6%	8%
Model Acc.	73.89%	73.85%	73.78%	73.25%	73.03%
Backdoor Acc.	9.40%	13.60%	33.40%	35.40%	43.26%
Running Time	6.63	6.72	6.83	7.01	7.16

Table R2: Evaluating learning rate on MNIST and CIFAR10. The results demonstrate that a larger learning rate can speed the convergence to achieve unlearning, costing less computation and achieving a better unlearning effect (low backdoor accuracy by removing). The tradeoff is that it slightly decreases the model utility at the same time, which is not too much on MNIST but a little worse on CIFAR10. @Reviewer iCGz, @Reviewer Gp18

	Metrics	Learning Rate: 0.0001	0.0002	0.0004	0.0006	0.0008
On MNIST	Model Acc.	98.52%	97.84%	96.72%	95.88%	95.37%
	Backdoor Acc.	9.67%	9.53%	9.17%	8.20%	8.67%
	Running Time	3.92	2.72	1.83	1.61	1.56
On Cifar10	Model Acc.	73.89%	72.98%	68.69%	65.23%	62.23%
	Backdoor Acc.	9.40%	6.20%	5.80%	4.00%	2.48%
	Running Time	6.63	3.72	2.83	2.51	2.23

Table R3: Membership inference attack accuracy after unlearning by OUBL. The results demonstrate that OUBL can effectively reduce the MI accuracy, achieving a significant unlearning performance. @Reviewer 5Whi

Dataset	Original Model	ASR and CSR, 1%	2%	3%	4%
On MNIST	63.86%	53.87%	53.61%	53.02%	52.86%
On CIFAR10	77.43%	61.47%	61.30%	61.10%	60.92%
On CelebA	58.37%	51.94%	51.32%	51.04%	50.69%

Table R4: The detailed running time. The results demonstrate that although we put more computational cost on the user side, it is affordable for users compared to the FL users in BFU. @Reviewer Ha5f

Dataset	Total running time of OUBL (second)	Unlearning update estimation (User side)	Unlearning noise generation (User side)	Unlearning by incremental learning (Server side)	Total running time of BFU
On MNIST	3.92	1.06	1.45	1.41	16.03
On CIFAR10	6.64	1.02	2.10	3.52	141.26
On CelebA	2.26	0.72	0.83	0.71	176.86

Table R5: Experimental results on Adult. The task of the Adult dataset is to predict whether an individual’s income exceeds \$50,000 per year, which is a binary classification. We first backdoor some samples in Adult by setting the “education-num” feature to 2 and changing the corresponding label. The aim of unlearning is to remove the influence of these backdoored samples, and the results are presented in the following table. Since the task on the Adult dataset is a binary task, dropping the backdoor accuracy of around 50% is similar to the random selection. Our method can effectively degrade the backdoor accuracy to around 50%, guaranteeing the effectiveness of unlearning. @Reviewer Gp18

On Adult	Original	ASR and CSR, 1%	2%	3%	4%
Model Acc.	85.32%	81.66%	81.69%	79.93%	80.79%
Backdoor Acc.	100.00%	54.81%	52.81%	50.02%	49.52%
Running Time	15.31	0.54	1.03	1.51	1.93

G SCENARIOS AND THREAT MODEL

@Reviewer BZGC, @Reviewer iCGz, @Reviewer 5Whi, @Reviewer EoYb, @Reviewer Ha5f, @Reviewer Gp18

Machine Unlearning Service Scenarios. To facilitate understanding, we introduce the problem in a Machine Learning as a Service (MLaaS) scenario. In the MLaaS setting, there are two key entities: a ML server that trains models as ML services, and users (data owners) who contribute their data for ML model training. In such scenarios, machine unlearning occurs when users realize that some of their previously contributed samples are private and wish to revoke these data contributions from the trained models.

The ML Server’s Ability. We assume the ML server is honest but curious [R1]: while it honestly hosts and provides ML services, including model training and updating, it may still be curious about private information, such as **unlearned data** and **unlearning intentions**, if there are other operations. Informing the server of unlearning intentions to customize unlearning operations is considered a privacy threat because it reveals users’ unlearning purposes, potentially enabling the server to prepare targeted unlearning attacks [R1,R2]. Therefore, in our setting, we assume the ML server has only the learning algorithm \mathcal{A} and the model with parameters θ to meet strict privacy requirements. The ML server will not conduct unlearning operations other than training the model using the learning algorithm \mathcal{A} for model updating.

Moreover, we assume the ML server does not store the original training data and cannot access the erased data, which should not be exposed to the server again due to privacy concerns. This assumption is reasonable in both real-world and privacy-preserving MLaaS scenarios. In real-world applications, vast amounts of data are generated daily, leading to the need of prompt model updates. Consequently, many models are trained using incremental or continual learning techniques [R3,R4]. Therefore, the server does not retain the entire raw data due to its large size [R5,R6]. In privacy-preserving scenarios, the ML server is restricted from directly accessing private training data from users due to privacy concerns [R7,R8].

The Users’ Ability. The training data D was collected from all users and was used to train the model θ_o . The unlearning user has the erased data $D_u \subset D$. To estimate the unlearning update as

the target for unlearning noise generation in our method, we assume the unlearning user can access the trained model θ_o , which is a common setting even in many privacy-preserving scenarios such as FL. We assume the unlearning user has auxiliary clean examples D_a so that they can synthesize a new dataset based on it with the unlearning noise, replacing the erased data D_u for achieving the unlearning effect with only incremental learning using the synthesized dataset.

H DISCUSSION ABOUT DISTINGUISHING BENIGN UNLEARNING USERS AND MALICIOUS USERS @REVIEWER ICGZ

To distinguish a benign user who wants to delete their data from a malicious user and who wants to upload noisy gradients to sabotage the model performance, we can only propose some possible ways for the server to distinguish these two kinds of users. The most significant difference is the purposes of the unlearning user and the malicious user. Unlearning users want to remove some knowledge of their data from the model, and they also want to preserve the model’s utility. Therefore, most clean samples and the auxiliary data they choose are in the same distribution as the genuine samples, and the synthesized noise should not influence the utility of the remaining dataset, as shown in the second objective of Eq.(4). However, the purpose of the malicious user is to sabotage the model performance. Their uploaded data will not be consistent with the genuine samples, so they can degrade model utility. We believe checking the similarity between the uploaded samples and genuine samples would be a possible solution. However, detailed poisoning attacking methods may need different solutions, and the problem is valuable to investigate in future work.

I DISCUSSION ABOUT DIFFERENCE BETWEEN EXISTING UNLEARNING METHODS @REVIEWER GP18

Compared with existing representative approximate unlearning methods [R19, R20, R21], our method also has the following differences. Specifically, the key techniques used in [R20] are the Hessian approximation and Fisher information, which is similar to our unlearning update estimation method that is also based on the Hessian matrix. The difference is that we use Hessian-vector products (HVPs) while [R20] uses the Fisher information to improve the efficiency. The HVPs solution is more efficient and more suitable to our scenarios in which the unlearning user cannot access the remaining dataset. [R19] and [R21] are approximate unlearning methods based on techniques called error maximizing. They generate error-maximizing noise for the unlearned samples to remove the influence from the model. One significant advantage of [R19] and [R21] is that they do not require access to the remaining training dataset. Compared with them, we put more effort into designing the method to further hide the unlearning data and the unlearning intentions from the server.

J ADDITIONAL EXPERIMENTS ON THE MORE PRACTICAL BLACK-BOX SCENARIOS @REVIEWER HA5F

To prove the feasibility of our method in the more practical black-box scenarios, we conducted additional experiments on the black-box setting on MNIST and CIFAR10. In this setting, the unlearning user cannot access the server’s current model. The unlearning user only knows the type of the model (MLP on MNIST and CNN on CIFAR10 in our experiment), and the user only has the erasing data and the auxiliary data. We set the size of auxiliary data to 1% of the server-side training data. Other unlearning settings are the same as the main setting in the paper, where we first backdoor the erasing data for model training and aim to unlearn these backdoored erasing data.

With these settings, the unlearning user trains a shadow model (θ_s) with 94.55% accuracy on MNIST and 42.57% accuracy on CIFAR10. By contrast, the accuracy of the server’s model (θ_o) trained with the entire dataset is 98.74% on MNIST and 78.80% on CIFAR10. Since both models are optimized on the erasing dataset, the proposed efficient unlearning update estimation (EUUE) method is effective for simulating the update of the unlearning data based on the shadow model. Hence, we can generate effective noise for the incremental learning data to approach the influence of unlearning. Then, we upload the constructed data to the server side for incremental learning, aiming to achieve the unlearning effect at the same time. We present the results as follows in Table [R6](#).

Table R6: Additional experiments on the black-box setting. On both datasets, OUBL achieves effective unlearning performance, effectively removing the backdoor influence. The backdoor removal effectiveness in the black-box setting is slightly lower than in the white-box setting. However, the negative impact on the model utility is also mitigated. These experimental results demonstrate the feasibility of OUBL in a more practical scenario, which lets the unlearning user not rely on the assumption of white-box access to the trained model in the federated learning scenarios. @Reviewer Ha5f

	Metrics	USR = 1%	2%	3%	4%	5%
On MNIST	Model Acc.(white-box)	98.52%	98.55%	98.15%	98.19%	95.43%
	Model Acc. (black-box)	98.26%	98.20%	98.31%	98.27%	98.54%
	Backdoor Acc. (white-box)	9.67%	10.08%	9.83%	10.42%	10.57%
	Backdoor Acc. (black-box)	12.33%	9.58%	11.67%	10.64%	11.83%
On Cifar10	Model Acc.(white-box)	73.89%	74.57%	74.50%	75.15%	75.99%
	Model Acc. (black-box)	76.06%	75.98%	74.93%	75.06%	74.68%
	Backdoor Acc. (white-box)	9.40%	7.30%	7.87%	8.70%	7.24%
	Backdoor Acc. (black-box)	13.20%	10.20%	8.40%	10.25%	8.28%

K ADDITIONAL REFERENCES

- [R1]. Hu, Hongsheng, et al. "Learn what you want to unlearn: Unlearning inversion attacks against machine unlearning." IEEE Symposium on Security and Privacy (SP) (2024).
- [R2]. Chen, Min, et al. "When machine unlearning jeopardizes privacy." Proceedings of the 2021 ACM SIGSAC conference on computer and communications security. 2021.
- [R3]. Rolnick, David, et al. "Experience replay for continual learning." Advances in neural information processing systems 32 (2019).
- [R4]. Lopez-Paz, David, and Marc'Aurelio Ranzato. "Gradient episodic memory for continual learning." Advances in neural information processing systems 30 (2017).
- [R5]. Wu, Yue, et al. "Large scale incremental learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
- [R6]. Wang, Zifeng, et al. "Learning to prompt for continual learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
- [R7]. Naseri, Mohammad, Jamie Hayes, and Emiliano De Cristofaro. "Local and central differential privacy for robustness and privacy in federated learning." NDSS (2022).
- [R8]. Bonawitz, Keith, et al. "Practical secure aggregation for privacy-preserving machine learning." proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017.
- [R9]. Gao, Xiangshan, et al. "Verifi: Towards verifiable federated unlearning." IEEE Transactions on Dependable and Secure Computing (2024).
- [R10]. Guo, Xintong, et al. "Fast: Adopting federated unlearning to eliminating malicious terminals at server side." IEEE Transactions on Network Science and Engineering (2023).
- [R11]. Zhu, Ligeng, Zhijian Liu, and Song Han. "Deep leakage from gradients." Advances in neural information processing systems 32 (2019).
- [R12]. Kurmanji, M., Triantafillou, P., Hayes, J. and Triantafillou, E., 2024. Towards unbounded machine unlearning. Advances in neural information processing systems, 36.
- [R13]. Chen, Min, et al. "When machine unlearning jeopardizes privacy." Proceedings of the 2021 ACM SIGSAC conference on computer and communications security. 2021.
- [R14]. Hu, Hongsheng, et al. "Membership inference via backdooring." IJCAI (2022).
- [R15]. Bertram, Theo, et al. "Five years of the right to be forgotten." Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019.
- [R16]. Oh, Seong Joon, Bernt Schiele, and Mario Fritz. "Towards reverse-engineering black-box neural networks." Explainable AI: interpreting, explaining and visualizing deep learning (2019): 121-144.

- [R17]. Salem, Ahmed, et al. "Updates-Leak: Data set inference and reconstruction attacks in online learning." 29th USENIX security symposium (USENIX Security 20). 2020.
- [R18]. Wang, Binghui, and Neil Zhenqiang Gong. "Stealing hyperparameters in machine learning." 2018 IEEE symposium on security and privacy (SP). IEEE, 2018.
- [R19] Chundawat, Vikram S., et al. "Zero-shot machine unlearning." IEEE Transactions on Information Forensics and Security 18 (2023): 2345-2354.
- [R20] Golatkar, Aditya, Alessandro Achille, and Stefano Soatto. "Eternal sunshine of the spotless net: Selective forgetting in deep networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
- [R21] Tarun, Ayush K., et al. "Fast yet effective machine unlearning." IEEE Transactions on Neural Networks and Learning Systems (2023).