LIA: Privacy-Preserving Data Quality Evaluation in Federated Learning Using a Lazy Influence Approximation — Supplementary Material —

Ljubomir Rokvic*, Panayiotis Danassis†, Sai Praneeth Karimireddy‡ and Boi Faltings*
*École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
†Telenor Research, Norway
‡University of California, Berkeley, USA
ljubomir.rokvic@epfl.ch, panayiotis.danassis@telenor.com, sp.karimireddy@berkeley.edu, boi.faltings@epfl.ch

I. Appendix / Supplemental Material

Contents

This appendix covers further details of our work, which have been omitted due to space limitations. Specifically:

1) In Section II we describe the methodology of our benchmarks.
2) In Section III we discuss the datasets used, the selected hyper-parameters, and other implementation details.
3) In Section IV we describe the potential positive and negative societal impact of our work.
4) In Section V we briefly discuss the limitations of our work.
5) In Sections VI we provide additional discussion of our simulation results.

II. Setting

We consider a classification problem from some input space $\mathcal{X}$ (e.g., features, images, etc.) to an output space $\mathcal{Y}$ (e.g., labels). In a Federated Learning setting, there is a center C that wants to learn a model $M(\theta)$ parameterized by $\theta \in \Theta$, with a non-negative loss function $L(z, \theta)$ on a sample $z = (\bar{x}, y) \in \mathcal{X} \times \mathcal{Y}$. Let $R(Z, \theta) = \frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$ denote the empirical risk, given a set of data $Z = \{z_i\}_{i=1}^{n}$. We assume that the empirical risk is differentiable in θ.The training data are supplied by a set of data holders.

A. Non-IID Setting

The main hurdle for Federated Learning is that not all data is IID. Heterogeneous data distributions are all but uncommon in the real world. To simulate a Non-IID distribution, we used Dirichlet distribution to split the training dataset as in related literature [3, 4, 7, 11]. This distribution is parameterized by α, which controls the concentration of different classes, as visualized in Figure 1. This work uses $\alpha \to 0.1$ for a non-IID distribution, as in related literature (e.g., [11]).

B. Exact Influence

In simple terms, influence measures the marginal contribution of a data point on a model's accuracy. A positive influence value indicates that a data point improves model accuracy, and vice-versa. More specifically, let $Z = \{z_i\}_{i=1}^{n}$, $Z_{+j} = Z \cup z_j$ where $z_j \notin Z$, and let

$$\hat{R} = \min_{\theta} R(Z, \theta) \quad \text{and} \quad \hat{R}_{+j} = \min_{\theta} R(Z_{+j}, \theta)$$

where $\hat{R}$ and $\hat{R}_{+j}$ denote the minimum empirical risk of their respective set of data. The *influence* of datapoint z_j on Z is defined as:

$$\mathcal{I}(z_j, Z) \triangleq \hat{R} - \hat{R}_{+j} \tag{1}$$

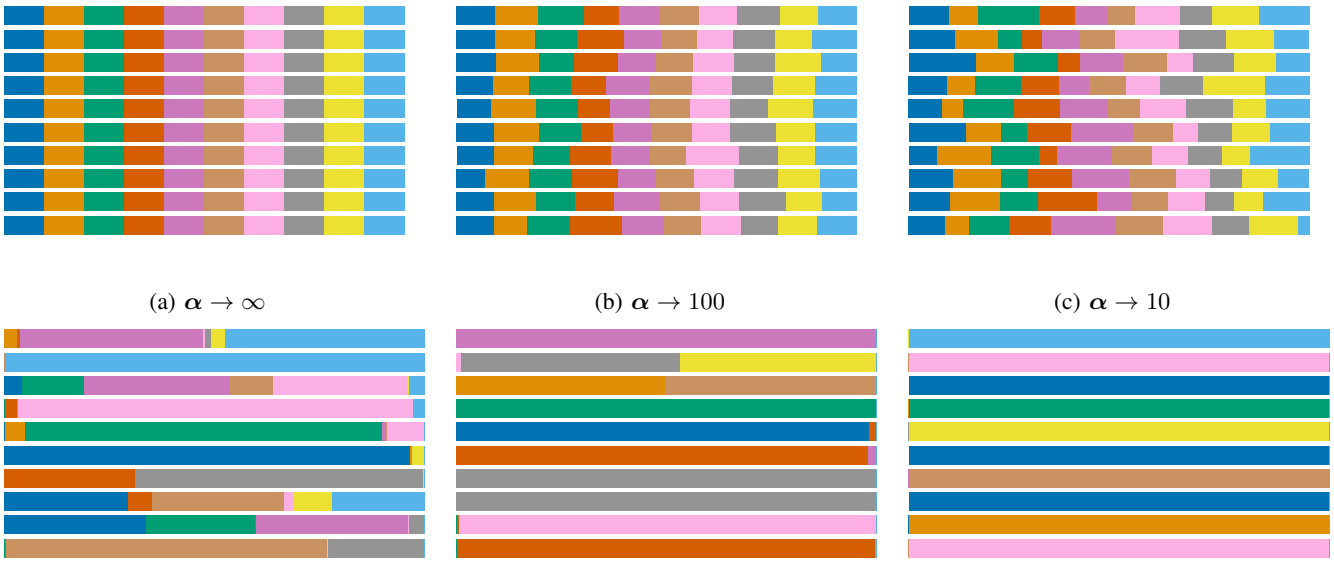
Despite being highly informative, influence functions have not achieved widespread use in Federated Learning (or Machine Learning in general). This is mainly due to the computational cost. Equation 1 requires complete retraining of the model, which is time-consuming, and very costly; especially for state-of-the-art, large ML models. Moreover, specifically in our setting, we do not have direct access to the training data. In the following section, we will introduce a practical approximation of the influence, applicable in Federated Learning scenarios.

C. Influence Approximation

The first-order Taylor approximation of influence, adopted by [5] (based on [1]), to understand the effects of training points on the predictions of a *centralized* ML model. To the best of our knowledge, this is the current state-of-the-art approach to utilizing the influence function in ML. Thus, it is worth taking the time to understand the challenges that arise if we adopt this approximation in the Federated Learning setting.

Let $\hat{\theta} = \arg\min_{\theta} R(Z, \theta)$ denote the empirical risk minimizer. The approximate influence of a training point z_j on the validation point z_{val} can be computed without having to re-train the model, according to the following equation:

$$\mathcal{I}_{appr}(z_j, z_{val}) \triangleq -\nabla_{\theta} L(z_{val}, \hat{\theta}) H_{\hat{\theta}}^{-1} \nabla_{\theta} L(z_j, \hat{\theta}) \tag{2}$$



(a) $\boldsymbol{\alpha} \to \infty$ (b) $\boldsymbol{\alpha} \to 100$ (c) $\boldsymbol{\alpha} \to 10$

(d) $\boldsymbol{\alpha} \to 0.1$ (e) $\boldsymbol{\alpha} \to 0.01$ (f) $\boldsymbol{\alpha} \to 0$

Fig. 1: Dirichlet distribution visualisation for 10 classes, parametrized by $\boldsymbol{\alpha}$. $\boldsymbol{\alpha}$ controls the concentration of different classes. Each row represents a participant, each color a different class, and each colored segment the amount of data the participant has from each class. For $\boldsymbol{\alpha} \to \infty$, each participant has the same amount of data from each class (IID distribution). For $\boldsymbol{\alpha} \to 0$, each participant only holds data from one class. In this work, we use $\boldsymbol{\alpha} \to 0.1$ for a non-IID distribution.

where $H_{\hat{\theta}}^{-1}$ is the inverse Hessian computed on all the model's training data. The advantage of Equation 2 is that we can answer counterfactuals on the effects of up/down-scaling a training point, without having to re-train the model. One can potentially average over the validation points of a participant, and/or across the training points in a batch of a contributor, to get the total influence.

D. Challenges

Consider Figure 2 as a motivating example. In this scenario, we have participants with corrupted data. Even a very robust model (ViT) loses performance when corruption is involved. This can also be observed in the work of [6]. Filtering those corrupted participants (orange line) restores the model's performance.

While Equation 2 can be an effective tool in understanding centralized machine learning systems, it is *ill-matched* for Federated Learning models, for several key reasons.

To begin with, evaluating Equation 2 requires *forming and inverting* the Hessian of the empirical risk. With n training points and $\theta \in \mathbb{R}^m$, this requires $O(nm^2 + m^3)$ operations [5], which is *impractical* for modern-day deep neural networks with millions of parameters. To overcome these challenges, [5] used implicit Hessian-vector products (HVPs) to more efficiently approximate $\nabla_\theta L(z_{val}, \hat{\theta}) H_{\hat{\theta}}^{-1}$, which typically requires $O(p)$ [5]. While this is a somewhat more efficient computation, it is *communication-intensive*, as it requires *transferring all of the (either training or validation) data* at each FL round. Most importantly, it *can not provide any privacy* to the users' data, an important, inherent requirement/constraint in FL.

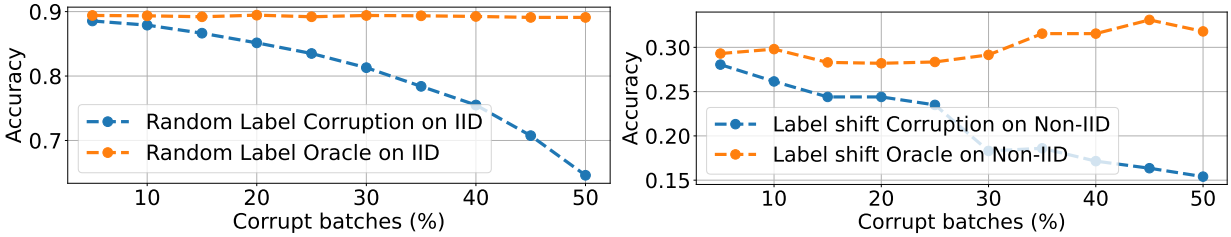
Finally, to compute Equation 2, the loss function has to be strictly convex and twice differentiable (which is not always the case in modern ML applications). A proposed solution is to swap out non-differentiable components for smoothed approximations [5], but there is no quality guarantee of the influence calculated in this way.

III. Implementation Details

This section describes the base model used in our simulations and all hyper-parameters. Specifically, we used a Visual Image Transformer (VIT) [2, 9]. The basis of our model represents a model pre-trained on ImageNet-21k at 224x224 resolution and then fine-tuned on ImageNet 2012 at 224x224 resolution. All hyper-parameters added or changed from the default VIT hyper-parameters are listed in Table I with their default values. The following hyper-parameters have been added to support our evaluation technique:

- **Random Factor**: this coefficient represents the amount of corrupted data inside a corrupted batch.
- **Final Evaluation Size**: an a priori separated batch of test data to evaluate model performance.
- **Parameters to Change**: number of parameters (and biases) in the last layer of the model.

For the HAR dataset we have used a simple two-layer fully connected neural network. This network has not been pretrained like the previous one. With this network we chose



(a) Random Label IID

(b) Label Shift Non-IID ($\alpha = 0.01$)

Fig. 2: Model accuracy relative to different mislabel rates (5% - 50%). These models have been trained over 25 communication rounds and 100 participants. We compare a centralized model with no filtering of mislabeled data (blue) to an FL model under perfect (oracle) filtering (orange). Note that the lower accuracy on the Non-IID setting is due to the fact that we are considering the most extreme non-IID case. This is where the majority of the participants have access to at most 1 class.

	CIFAR10	CIFAR100	HAR
No. of Participants	100	100	500
Batch Size	100	250	2000
Validation Size	50	50	500
Random Factor	0.9	0.9	1.0
Warm-up Size	600	4000	7352
Final Evaluation Size	2000	2000	2947
Load Best Model	False	False	False
Parameters to Change	7690	76900	36358
Learning Rate	0.001	0.001	0.001
Train Epochs	3	3	20
Weight Decay	0.01	0.01	None

TABLE I: Table of hyper-parameters.

to omit regularization and the detailed hyper-parameters are listed in Table I. [10]

Regarding reproducibility, we ran the provided (in the supplementary material) code for each dataset with seeds from the range of $0 - 7$.

A. Termination Condition

Different termination conditions have been used for our proposed solution and to retrain the exact influence. Our solution has only one termination condition, that is the number of local epochs k.

B. Computational Resources

All simulations were run on two different systems:

1) Intel Xeon E5-2680 – 12 cores, 24 threads, 2.5 GHz – with 256 GB of RAM, Titan X GPU (Pascal)
2) EPFL RCP Cluster with a100 and h100 GPUs.

IV. Societal Impact

Privacy advocacy movements have, in recent years, raised their voices about the potential abuse of these systems. Additionally, legal institutions have also recognized the importance of privacy, and have passed regulations in accordance, for example, the General Data Protection Regulation (GDPR). Our work provides practical privacy guarantees to protect all parties, with minimal compromise on performance. Furthermore, we allow data holders and collectors to be paid for their contribution in a joint model, instead of simply taking the data. Such incentives could potentially help speed up the growth of underdeveloped countries, and provide more high-quality data to researchers (as an example application, consider paying low-income farmers for gathering data in crop disease prevention [8]).

	Corruption				
	0%	10%	20%	30%	40%
Accuracy	85.2 +- 2.1 %	86.4 +- 1.2 %	87.2 +- 1.3 %	87.4 +- 1.1 %	89.0 +- 1.2 %
Recall	-	97.0 +- 0.5 %	96.4 +- 0.7%	98.1 +- 0.6%	97.5 +- 0.7 %

TABLE II: Filtering performance on varying percentages of corrupt participants on the Human Activity Recognition dataset for *highly non-IID*. Following the same strict *worst-case differential privacy* guarantees ($\varepsilon \leq 1, \delta = 10^{-5}$).

V. Limitations

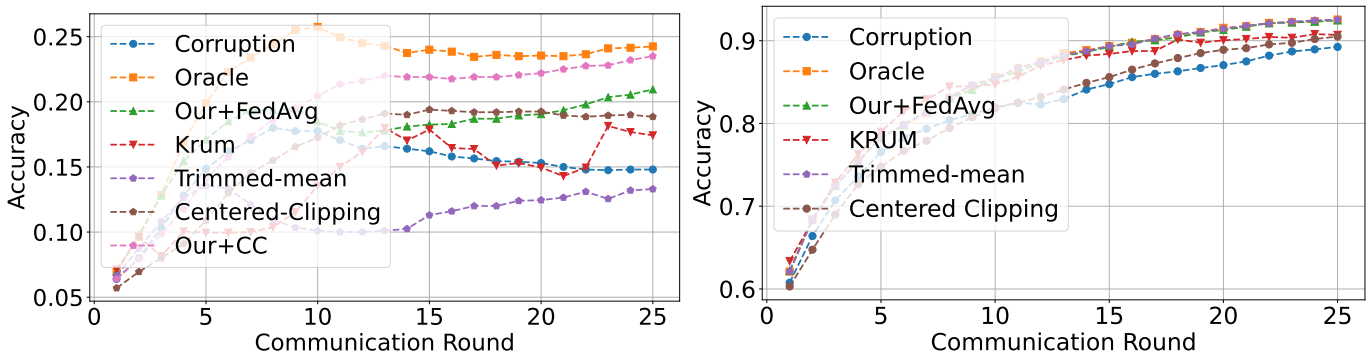
The main limitation of our approach is that if the optimizer does not produce a good enough gradient, we cannot get a good approximation of the direction the model is headed for. The result of this is a lower score, and therefore a potentially inaccurate prediction.

Another potential limitation is the filtering of "good" data. These data may be correctly labeled, but including it does not essentially provide any benefit to the model, as can be shown by the accuracy scores in Figure 3. While this allows us to train models of equal performance with a fraction of the data, some participants may be filtered out, even though they contribute accurate data. This might deter users from participating in the future.

VI. Numerical Results

We provide detailed results that include both the means and standard deviations. The metrics can be found in Table III. The following subsections provide a more comprehensive analysis of the results, summarized in the main text due to space limitations.

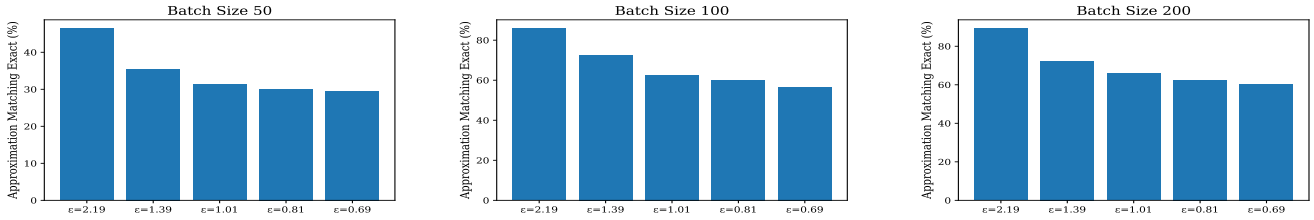
We have conducted initial testing on CIFAR10 and CIFAR100 to explore the impact of various parameters on our model's performance. Our results, illustrated in Figures, 6, and 7 demonstrate the effect of both learning rate and number of epochs on filtration performance.



(a) Label shift Non-IID ($\alpha = 0.01$)

(b) Random Label IID

Fig. 3: Model accuracy over 25 communication rounds with a 30% mislabel rate on CIFAR-10. We compare a centralized model with no filtering (blue) to an FL model under perfect (oracle) filtering (orange), KRUM (red), Trimmed-mean (purple), and our approach (green). Note that the jagged line for KRUM is because only a single gradient is selected instead of performing FedAvg.



(a)

(b)

(c)

Fig. 4: Effect of batch size to the percentage of the times the sign of the proposed Lazy Influence Approximation (LIA) matches the sign of the exact influence, for varying differential privacy guarantees (ε in the x-axis). Comparing 4a to the other two figures, we can observe that there is a certain threshold of data that needs to be passed for *lazy influence* to be effective. After this threshold has been reached, adding more data only gives marginal improvement as can be seen by comparing 4b and 4c.

		Filtration Metrics		
	Distribution	Recall	Precision	Accuracy
CIFAR 10	IID	97.08 ± 3.51 %	91.91 ± 7.15 %	96.38 ± 2.83 %
	Non-IID	93.75 ± 5.12 %	69.02 ± 6.28 %	85.00 ± 3.28 %
CIFAR 100	IID	99.17 ± 2.20 %	97.96 ± 2.30 %	99.12 ± 1.27 %
	Non-IID	92.50 ± 5.71 %	55.41 ± 3.94 %	75.12 ± 3.76 %
HAR	IID	100.00 ± 0.0%	100.00 ± 0.0%	100.00 ± 0.0%
	Non-IID	95.77 ± 2.1%	71.71 ± 1.3%	87.40 ± 1.1%

TABLE III: Filtering performance on various datasets, including *real-data* on Human Activity Recognition, for IID and *highly non-IID* setting ($\alpha \to 0.1$, i.e., 3 classes per participant for HAR). 100 participants, 30% mislabeling rate. *Strict worst-case differential privacy* guarantees ($\varepsilon \leq 1, \delta = 10^{-5}$).

We observe a balance between recall and accuracy that varies based on the parameters. This balance can be seen in both the CIFAR10 and CIFAR100 datasets. Additionally, the best parameters for IID and Non-IID may differ. For instance, the best recall for Non-IID and IID is achieved with different parameter pairs, and CIFAR100 also has a distinct parameter pair for IID compared to Non-IID.

Finally, we examine the impact of various privacy guarantees (ε) and larger problem dimensions in Figure 5. Our findings show that a smaller federation is needed to achieve the same level of performance when data is IID, compared to when it is Non-IID.

REFERENCES

[1] R Dennis Cook and Sanford Weisberg. *Residuals and influence in regression*. Chapman and Hall, 1982.

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[3] Haley Hoech, Roman Rischke, Karsten Müller, and Wojciech Samek. Fedauxfdp: Differentially private one-shot federated distillation, 2022.

[4] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 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] Anran Li, Lan Zhang, Juntao Tan, Yaxuan Qin, Junhao Wang, and Xiang-Yang Li. Sample-level data selection for federated learning. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021.

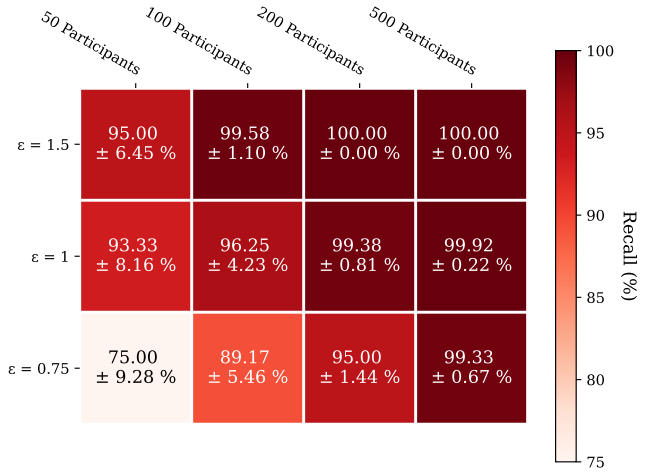
[7] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 2020.

[8] Sharada P Mohanty, David P Hughes, and Marcel Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:1419, 2016.

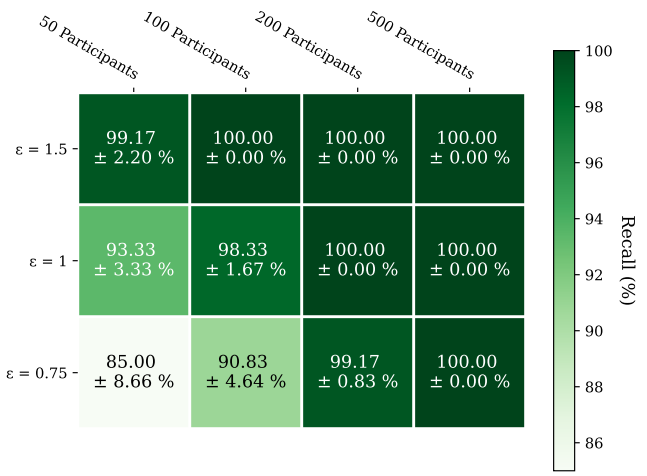
[9] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020.

[10] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. Opacus: User-friendly differential privacy library in pytorch. *CoRR*, abs/2109.12298, 2021.

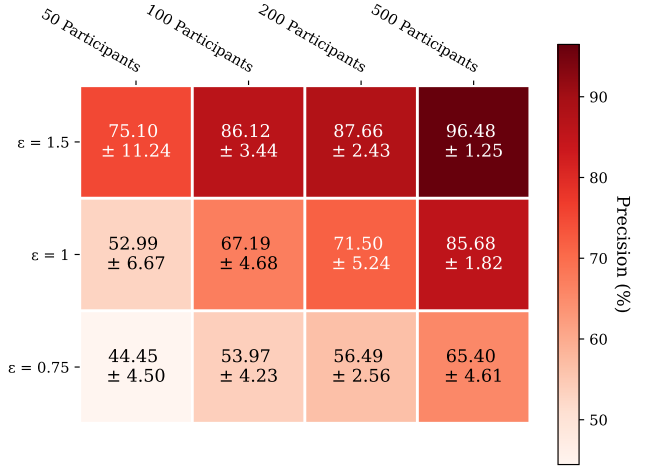
[11] Yaodong Yu, Alexander Wei, Sai Praneeth Karimireddy, Yi Ma, and Michael Jordan. TCT: Convexifying federated learning using bootstrapped neural tangent kernels. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.



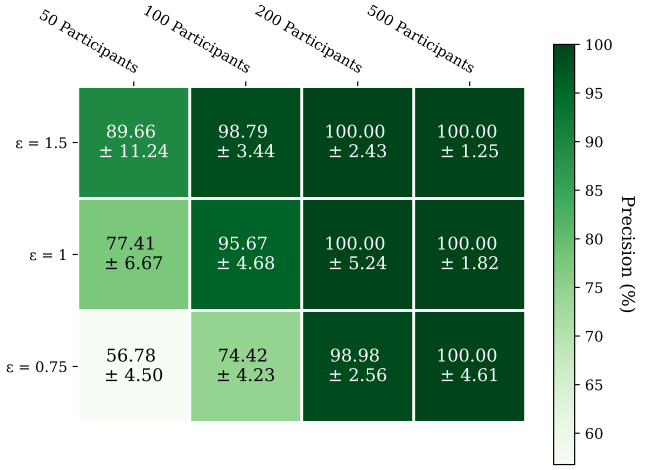
(a) Recall



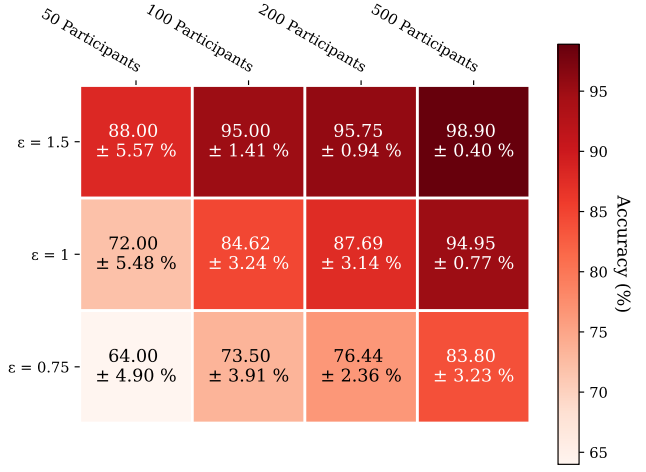
(b) Recall



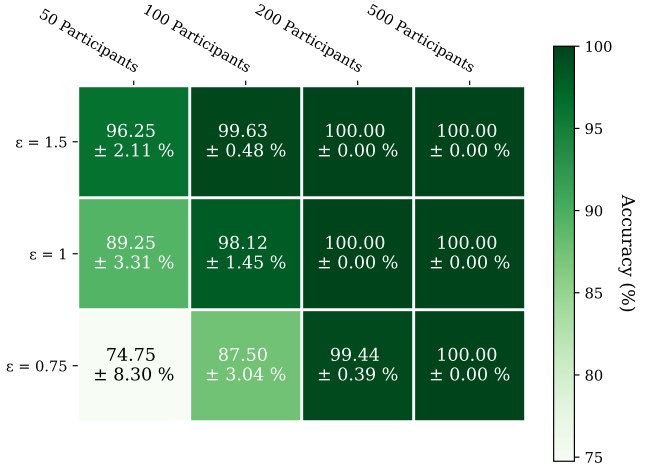
(c) Precision



(d) Precision

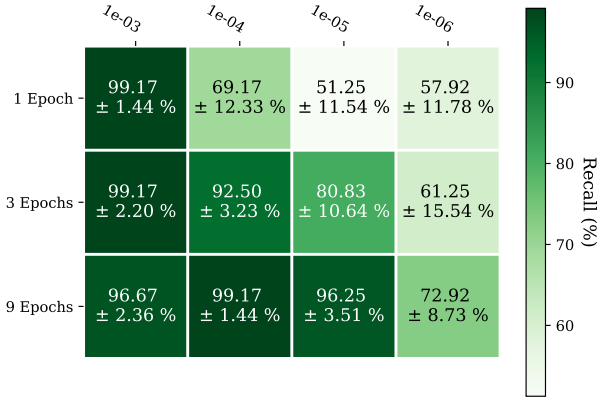


(e) Accuracy

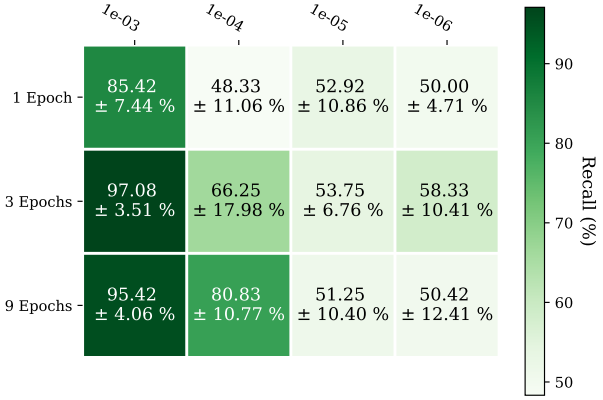


(f) Accuracy

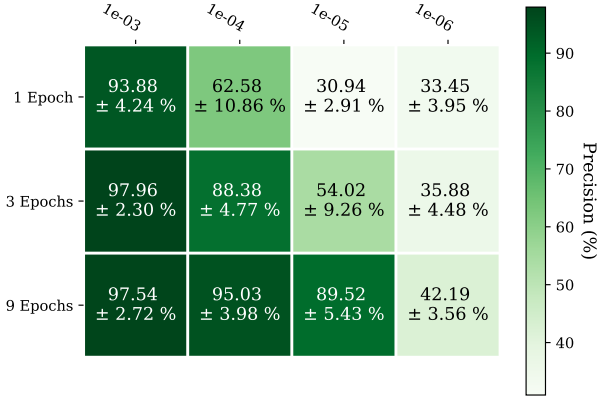
Fig. 5: Recall (top), Precision (middle), and Accuracy(Bottom) on CIFAR 10, non-IID (left), IID (right), for increasing problem size (number of participants), and varying privacy guarantees (ε – lower ε provides stronger privacy).



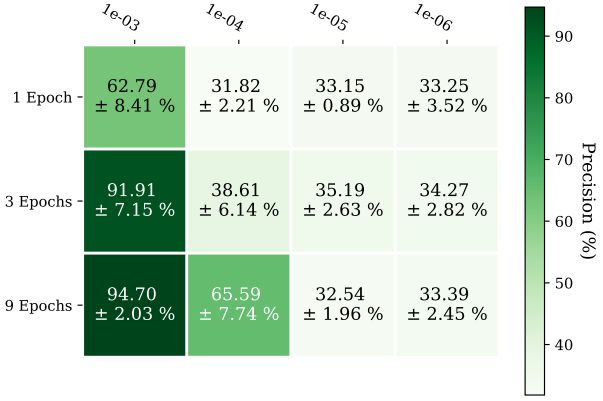
(a) Recall



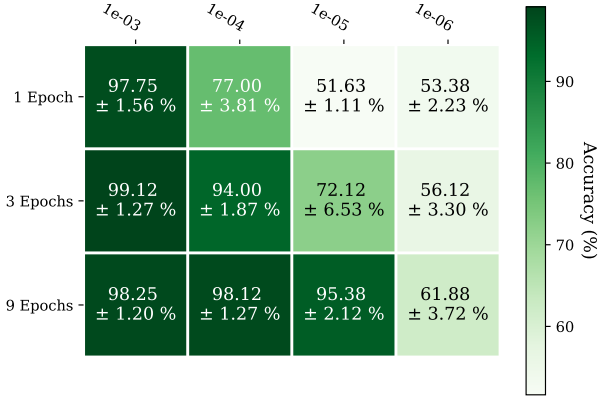
(b) Recall



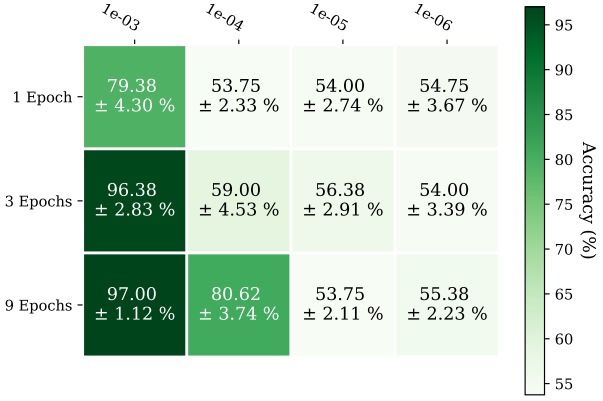
(c) Precision



(d) Precision

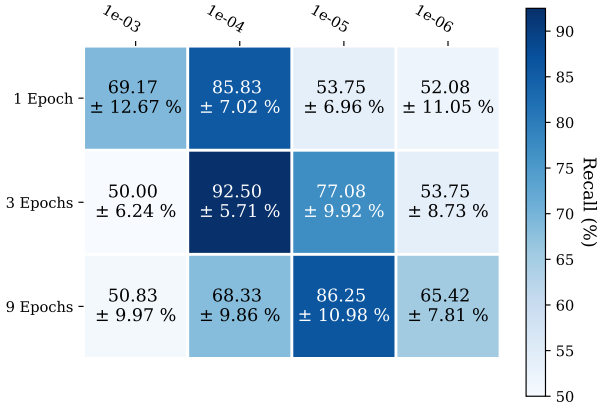


(e) Accuracy

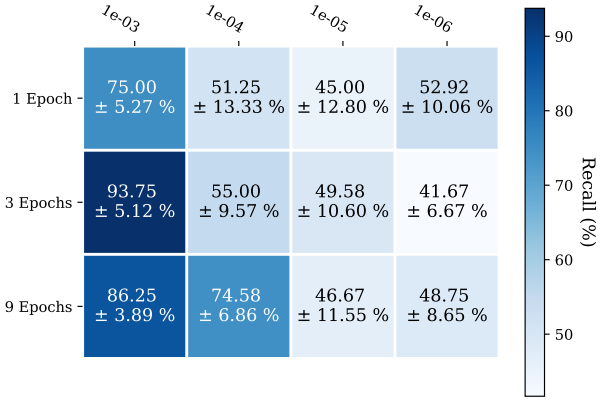


(f) Accuracy

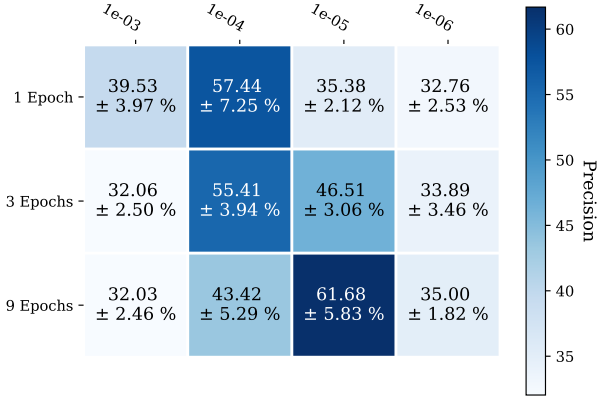
Fig. 6: Recall (top), Precision (middle), and Accuracy(Bottom) on CIFAR 100 (left) and CIFAR 10 (right), IID, for different parameter pairs of learning rate and epoch count.



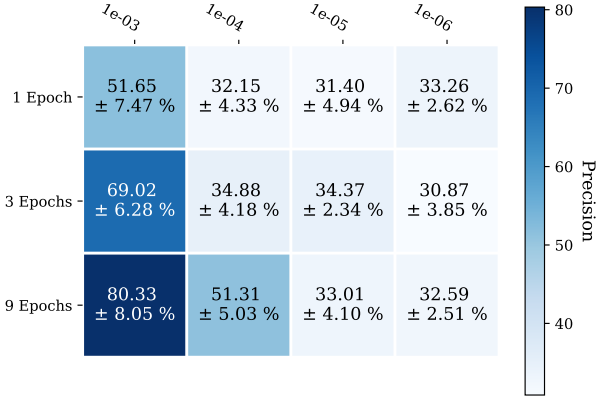
(a) Recall



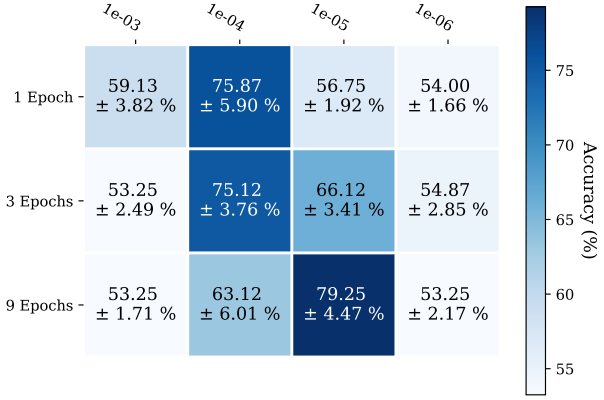
(b) Recall



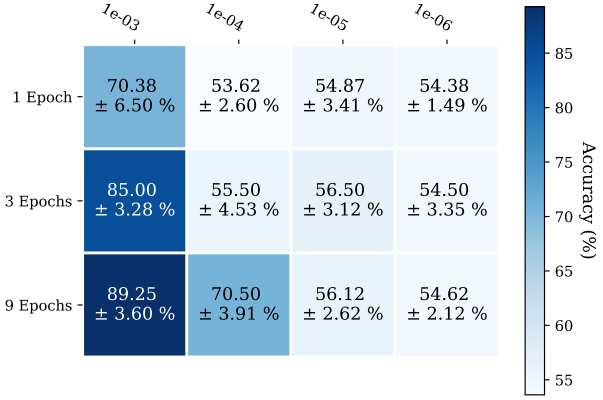
(c) Precision



(d) Precision



(e) Accuracy



(f) Accuracy

Fig. 7: Recall (top), Precision (middle), and Accuracy(Bottom) on CIFAR 100 (left), and CIFAR 10 (right), non-IID, for different parameter pairs of learning rate and epoch count.