ICLR 2026

Convergent Differential Privacy Analysis for General Federated Learning

The University of Sydney & Shenzhen Campus of Sun Yat-sen University & Nanyang Technological University

Yan Sun[12], Qixin Zhang[14], Li Shen[3], Dacheng Tao*[4]

[1] Equal contributions
[2] The University of Sydney
[3] Shenzhen Campus of Sun Yat-sen University
[4] Nanyang Technological University
* Corresponding author

FL-DP Framework

Privacy-Preserving Mechanism for Local Datasets

01

DP & FL

Differential Privacy in Federated Learning

02

Object

Target of the Protection Mechanism

03

f-DP & GDP

Hypothesis-Testing-Based Privacy Metric

04

Convergent DP

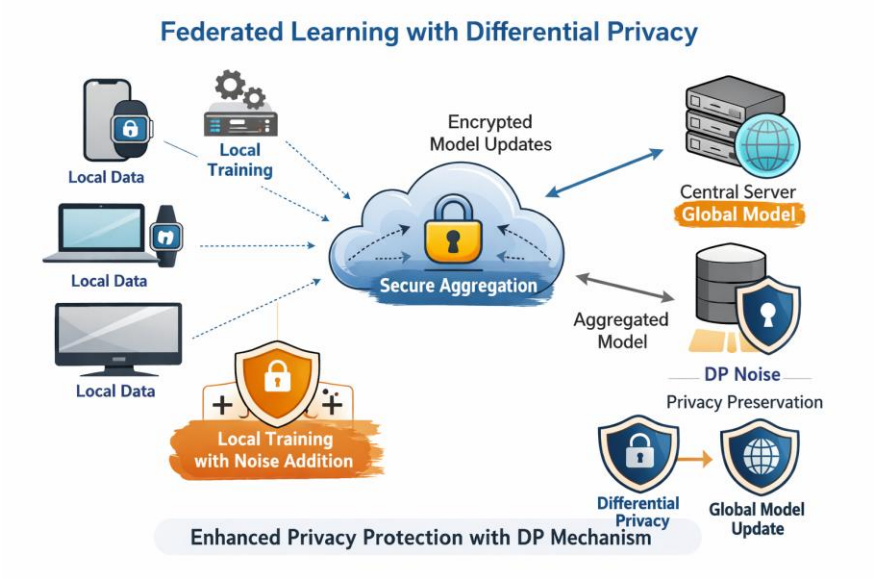
Convergent bound for Local Model Training

Federated Learning and Differential Privacy Mechanism

01

General FL

Federated learning enables collaborative model training across clients without sharing their raw local data.



02

DP

Differential privacy protects individual data by adding carefully calibrated noise to shared information.

1. DP & FL

1 Local Differential Privacy in FL

Each client adds noise to its model updates (e.g., gradients or parameters) before uploading them to the server for aggregation.

2 Central Differential Privacy in FL

Clients upload their model updates, and the server adds noise during or after aggregation (often combined with secure aggregation).

3 Gradient Clipping with Noise Injection

Client gradients are clipped to bound sensitivity, and calibrated noise (e.g., Gaussian noise) is added to satisfy DP guarantees.

3 Personalized DP in Federated Learning

Different clients adopt different privacy budgets or noise levels based on their data sensitivity or system requirements.

2. Object

In our analysis, the f-DP composition is applied to the sequence of perturbed model updates, each of which is a Gaussian mechanism. For Gaussian mechanisms, a valid coupling is guaranteed to exist: the optimal transport coupling that aligns the two output distributions via a shared Gaussian noise source. Formally, for adjacent datasets $\mathcal{C}$ and $\mathcal{C}'$, the mechanisms can be written as:

$$\mathcal{M}(C) = \text{Standard FL Update}(C) + Z \;\; and \;\; \mathcal{M}(C') = \text{Standard FL Update}(C') + Z,$$

where Standard FL Update() is generally defined by the algorithm itself, e.g. FedAvg and FedProx in our paper. Z is the shared Gaussian noise. This defines a measurable coupling, since the pair $(\text{Standard FL Update}(C) + Z, \text{Standard FL Update}(C') + Z)$ is a measurable mapping of Gaussian random variable. Under this coupling, the privacy-loss random variable has the explicit closed form used in f-DP analysis, and its mean shift is

$$\mu_t = \frac{\|\text{Standard FL Update}(C) - \text{Standard FL Update}(C')\|^2}{2\sigma^2},$$

2. Object

We consider the general finite-sum minimization problem in the classical federated learning:

$$w^\star \in \arg\min_w f(w) \triangleq \frac{1}{m}\sum_{i\in\mathcal{I}} f_i(w), \tag{1}$$

where $f_i(w) = \mathbb{E}_{\varepsilon\sim\mathcal{D}_i}[f_i(w,\varepsilon)]$ denotes the local population risk. $w \in \mathbb{R}^d$ denotes d-dim learnable parameters. $\varepsilon \sim \mathcal{D}_i$ denotes that the private dataset on client i is sampled from distribution $\mathcal{D}_i$. We consider the general heterogeneity, i.e. $\mathcal{D}_i$ can differ from $\mathcal{D}_j$ if $i \neq j$, leading to $f_i(w) \neq f_j(w)$.

Noisy-FedAvg: we consider that each local client performs a fundamental gradient descent as follows:

$$w_{i,k+1,t} = w_{i,k,t} - \eta_{k,t} g_{i,k,t}, \tag{2}$$

where $g_{i,k,t} = \nabla f_i(w_{i,k,t},\varepsilon)/\max\{1, \frac{\|\nabla f_i(w_{i,k,t},\varepsilon)\|}{V}\}$, and V is a constant coefficient.

Noisy-FedProx: The vanilla local training in `FedProx` is based on solving the following surrogate:

$$\min_w f_i(w) + \frac{\alpha}{2}\|w - w_t\|^2. \tag{3}$$

To generally compare with `Noisy-FedAvg`, we consider an iterative form of gradient descent as:

$$w_{i,k+1,t} = w_{i,k,t} - \eta_{k,t}\left[g_{i,k,t} + \alpha(w_{i,k,t} - w_t)\right]. \tag{4}$$

3. f-DP & GDP

Definition 1 *We denote heterogeneous datasets on the client i by $\mathcal{S}_i = \{\varepsilon_{ij}\}$ and let the union of all local datasets be $\mathcal{C} = \{\mathcal{S}_i\}$. We say two unions are adjacent datasets if they only differ by one data sample. For instance, there exists the union $\mathcal{C}' = \{\mathcal{S}_i'\}$. $(\mathcal{C}, \mathcal{C}')$ are adjacent datasets if there exists the index pair $(i^\star, j^\star)$ such that all other data samples are the same except for $\varepsilon_{i^\star j^\star} \neq \varepsilon'_{i^\star j^\star}$.*

Definition 2 *A randomized mechanism $\mathcal{M}$ is (ϵ, δ)-DP if for any event E the following satisfies:*

$$P(\mathcal{M}(\mathcal{C}) \in E) \leq e^{\epsilon} P(\mathcal{M}(\mathcal{C}') \in E) + \delta. \tag{5}$$

Definition 2 is the widely used (ϵ, δ)-DP, which is a lossy relaxation in the DP analysis since its probabilistic gaps. To bridge the discrepancy of precise DP definitions, statistic analysis demonstrates that DP could be naturally deduced by hypothesis-testing problems (Wasserman & Zhou, 2010; Kairouz et al., 2015). From the perspective of attackers, DP means the difficulty in distinguishing $\mathcal{C}$ and $\mathcal{C}'$ under the mechanism $\mathcal{M}$. They can generally consider the following problem:

Given $\mathcal{M}$, is the underlying union $\mathcal{C}$ (H_0) or $\mathcal{C}'$ (H_1)?

To exactly quantify the difficulty of its answer, Dong et al. (2022) propose that distinguishing these two hypotheses could be best delineated by the optimal trade-off between the possible type I and type II errors. Specifically, by considering rejection rules $0 \leq \chi \leq 1$, type I and type II errors can be:

$$E_I = \mathbb{E}_{\mathcal{M}(\mathcal{C})}[\chi], \qquad E_{II} = 1 - \mathbb{E}_{\mathcal{M}(\mathcal{C}')}[\chi], \tag{6}$$

Here, we abuse $\mathcal{M}(\mathcal{C})$ to represent its probability distribution. To measure the fine-grained relationships between these two testing errors, f-DP is introduced.

Definition 3 (Trade-off function) *For any two probability distributions P and Q, the trade-off function is defined as: $T(P;Q)(\gamma) = \inf\{1 - \mathbb{E}_Q[\chi] \mid \mathbb{E}_P[\chi] \leq \gamma\}$, where the infimum is taken over all measurable rejection rules.*

$T(P;Q)(\gamma)$ is convex, continuous, and non-increasing. For any possible rejection rules, it satisfies $T(P;Q)(\gamma) \leq 1 - \gamma$. It functions as the clear boundary between the achievable and unachievable selections of type I and type II errors, essentially distinguishing the difficulties between these two hypotheses. This relevant statistical property provides a stricter definition of privacy, which mitigates the excessive relaxation of privacy based on composition analysis in existing approaches.

Definition 4 (f-DP and GDP) *A mechanism $\mathcal{M}$ is f-DP if $T(\mathcal{M}(\mathcal{C}), \mathcal{M}(\mathcal{C}'))(\gamma) \geq f(\gamma)$ for all possible adjacent datasets $\mathcal{C}$ and $\mathcal{C}'$. When f measures two Gaussian distributions, namely Gaussian-DP (GDP), denoted as $T_G(\mu)(\gamma) \triangleq T(\mathcal{N}(0,1), \mathcal{N}(\mu,1))(\gamma)$ for $\mu \geq 0$.*

Lemma 1 (Post-processing) *If a randomized mechanism $\mathcal{M}$ is f-DP, any post processing mechanism based on $\mathcal{M}$ is still at least f-DP, i.e. $T(P';Q') \geq T(P;Q)$ for any post-processing mapping which leads to $P \to P'$ and $Q \to Q'$.*

Intuitively, post-processing mappings bring some changes in the original distributions. However, such changes can not allow the updated distributions to be much easier to discern. This lemma also widely exists in other DP relaxations and stands as one of the foundational elements in current privacy analyses. In f-DP, this lemma also clearly demonstrates that the difficulty of hypothesis testing problems can not be simplified with the addition of known information, which still preserves the original distinguishability.

Lemma 2 (Composition) *We have a series of mechanisms $\mathcal{M}_i$ and a joint serial composition mechanism $\mathcal{M}$. Let each private mechanism $\mathcal{M}_i(\cdot, y_1, \cdots, y_{i-1})$ be f_i-DP for all $y_1 \in Y_1, \cdots, y_{i-1} \in Y_{i-1}$. Then the n-fold composed mechanism $\mathcal{M} : X \to Y_1 \times \cdots \times Y_n$ is $f_1 \otimes \cdots \otimes f_n$-DP, where $\otimes$ denotes the joint distribution. For instance, if $f = T(P;Q)$ and $g = T(P';Q')$, then $f \otimes g = T(P \times P'; Q \times Q')$.*

Lemma 3 (GDP $\to$ (ϵ, δ)-DP) *A μ-GDP mechanism with a trade-off function $T_G(\mu)$ is also $(\epsilon, \delta(\epsilon))$-DP for all $\epsilon \geq 0$ where*

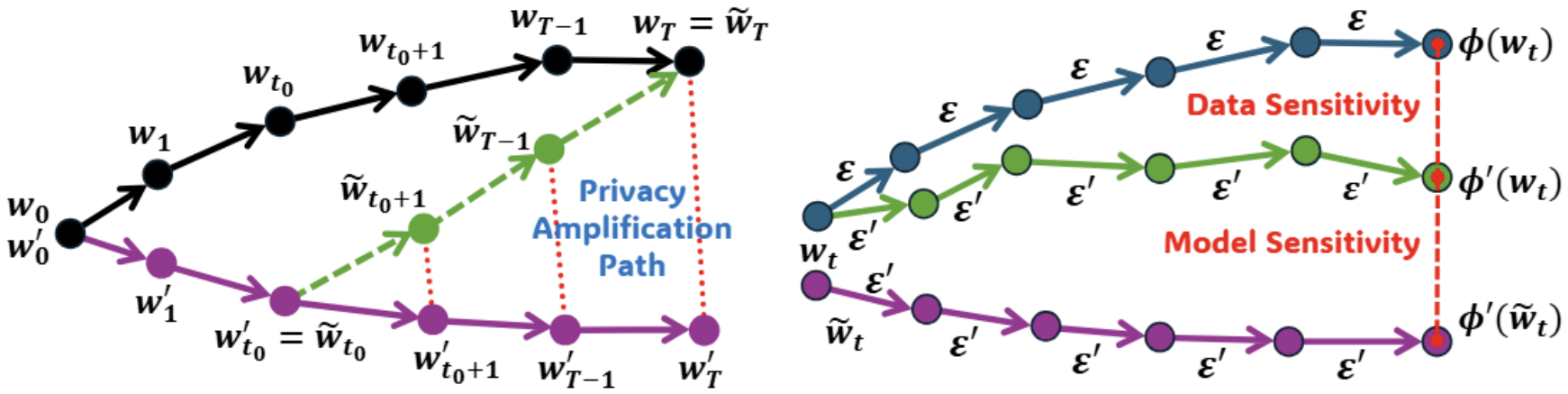
$$\delta(\epsilon) = \Phi\left(-\frac{\epsilon}{\mu} + \frac{\mu}{2}\right) - e^{\epsilon}\Phi\left(-\frac{\epsilon}{\mu} - \frac{\mu}{2}\right). \tag{20}$$

Lemma 4 (GDP $\to$ RDP) *A μ-GDP mechanism with a trade-off function $T_G(\mu)$ is also $\left(\zeta, \frac{1}{2}\mu^2\zeta\right)$-RDP for any $\zeta > 1$.*

We state the transition and conversion calculations from f-DP (we specifically consider the GDP) to other DP relaxations, e.g. for the (ε, δ)-DP and RDP. These lemmas can effectively compare our theoretical results with existing ones. Our comparison primarily aims to demonstrate that the convergent privacy obtained in our analysis would directly derive bounded privacy budgets in other DP relaxations. Moreover, we will illustrate how the convergent f-DP further addresses conclusions that current FL-DP work cannot cover theoretically, which provides solid support for understanding its reliability of privacy protection.

Lemma 5 (Accumulation in GDP.) *For GDP, $T_G(\mu_1) \otimes \cdots \otimes T_G(\mu_n) = T_G(\sqrt{\mu_1^2 + \cdots + \mu_n^2})$.*

4. Convergent DP



To simplify presentations, we denote global updates at round t on the adjacent datasets $\mathcal{C}$ and $\mathcal{C}'$ as:

$$\mathcal{C} : w_{t+1} = \phi(w_t) + \overline{n}_t, \quad \mathcal{C}' : w'_{t+1} = \phi'(w'_t) + \overline{n}'_t. \tag{8}$$

$\phi(w_t)$ denotes the accumulation of total K steps from the initialization state $w_{i,0,t} = w_t$ at round t. $\overline{n}_t$ could be considered as the averaged noise, i.e. $\overline{n}_t \sim \mathcal{N}(0, \sigma^2 I_d/m)$. Traditional methods require performing privacy amplification T times based on the relationship between w and w', yielding non-convergent privacy as T. To avoid loose privacy amplification, we follow Bok et al. (2024) to adopt the *shifted interpolation* technique. Specifically, we define the following sequence:

$$\widetilde{w}_{t+1} = \lambda_{t+1}\phi(w_t) + (1 - \lambda_{t+1})\phi'(\widetilde{w}_t) + \overline{n}_t, \tag{9}$$

where $t = t_0, \cdots, T-1$. By setting $\lambda_T = 1$, then $\widetilde{w}_T = w_T$, and we add the definition of $\widetilde{w}_{t_0} = w'_{t_0}$ as the beginning of interpolations. $0 \leq \lambda_t \leq 1$ are interpolation coefficients to be optimized. As shown in Figure 1 (left), the interpolation sequence path enables a privacy amplification analysis over $T - t_0$ times where t_0 is an optimizable coefficient. Therefore, we can establish the following theorem along this new privacy amplification path.

4. Convergent DP

Theorem 1 *Under Assumption 1 and corresponding updates in Eq.(8), After T training rounds on the adjacent datasets $\mathcal{C}$ and $\mathcal{C}'$, we can bound the trade-off function between w_T and w'_T as:*

$$T(w_T; w'_T) = T(\widetilde{w}_T; w'_T) \geq T_G\left(\frac{\sqrt{m}}{\sigma}\sqrt{\sum_{t=t_0}^{T-1}\lambda_{t+1}^2\|\phi(w_t) - \phi'(\widetilde{w}_t)\|^2}\right). \tag{10}$$

Theorem 2 *Under K local updates by Eq.(2) and Eq.(4), the global sensitivity in* `Noisy-FedAvg` *and* `Noisy-FedProx` *methods can be shown as:*

$$\|\phi(w_t) - \phi'(\widetilde{w}_t)\| \leq \underbrace{\rho_t\|w_t - \widetilde{w}_t\|}_{\text{from model sensitivity}} + \underbrace{\gamma_t}_{\text{from data sensitivity}}, \tag{11}$$

where ρ_t and γ_t are shown in Table 2.

Table 2: Specific formulation of ρ_t and γ_t in Theorem 2.

	Learning rate	ρ_t	γ_t
`Noisy-FedAvg`	μ	$(1+\mu L)^K$	$\frac{2\mu V}{m}K$
	$\frac{\mu}{k+1}$	$(1+K)^{c\mu L}$	$\frac{2cV}{m}\ln(K+1)$
	$\frac{\mu}{t+1}$	$\left(1+\frac{\mu L}{t+1}\right)^K$	$\frac{2\mu V}{m}\frac{K}{t+1}$
	$\frac{\mu}{tK+k+1}$	$\left(\frac{t+2}{t+1}\right)^{z\mu L}$	$\frac{2zV}{m}\ln\left(\frac{t+2}{t+1}\right)$
`Noisy-FedProx`	non-increase	$\frac{\alpha}{\alpha-L}$	$\frac{2V}{m\alpha}$

4. Convergent DP

Theorem 3 *Let $f_i(w)$ be a L-smooth and non-convex local objective and local updates be performed as shown in Eq.(2). Under perturbations of isotropic noises $n_i \sim \mathcal{N}\left(0, \sigma^2 I_d\right)$, the worst privacy of the* `Noisy-FedAvg` *method achieves:*

(a) under constant learning rates $\eta_{k,t} = \mu$:

$$T(w_T; w_T') \geq T_G\left(\frac{2\mu V K}{\sqrt{m}\sigma}\sqrt{\frac{(1+\mu L)^K+1}{(1+\mu L)^K-1}\frac{(1+\mu L)^{KT}-1}{(1+\mu L)^{KT}+1}}\right). \tag{15}$$

(b) under cyclically decaying $\eta_{k,t} = \frac{\mu}{k+1}$:

$$T(w_T; w_T') \geq T_G\left(\frac{2cV\ln(K+1)}{\sqrt{m}\sigma}\sqrt{\frac{(1+K)^{c\mu L}+1}{(1+K)^{c\mu L}-1}\frac{(1+K)^{c\mu LT}-1}{(1+K)^{c\mu LT}+1}}\right). \tag{16}$$

(c) under stage-wise decaying $\eta_{k,t} = \frac{\mu}{t+1}$:

$$T(w_T; w_T') > T_G\left(\frac{2\mu V K}{\sqrt{m}\sigma}\sqrt{2-\frac{1}{T}}\right). \tag{17}$$

(d) under continuously decaying $\eta_{k,t} = \frac{\mu}{tK+k+1}$:

$$T(w_T; w_T') > T_G\left(\frac{2zV}{\sqrt{m}\sigma}\sqrt{2-\frac{1}{T}}\right). \tag{18}$$

Theorem 4 *Let $f_i(w)$ be a L-smooth and non-convex local objective and local updates be performed as shown in Eq.(4). Let the proximal coefficient $\alpha > L$ and $\eta < \frac{1}{\alpha - L}$, under perturbations of isotropic noises $n_i \sim \mathcal{N}\left(0, \sigma^2 I_d\right)$, the worst privacy of the* `Noisy-FedProx` *method achieves:*

$$T(w_T; w_T') \geq T_G\left(\frac{2V}{\sqrt{m}\alpha\sigma}\sqrt{\frac{2\alpha - L}{L}\left(1 - \frac{2}{\left(\frac{\alpha}{\alpha - L}\right)^T + 1}\right)}\right), \tag{19}$$

Table 3: Comparisons with the existing theoretical results in FL-DP. We losslessly transfer our results into (ϵ, δ)-DP and RDP results. In (ϵ, δ)-DP, we compare the requirement of noise variance corresponding to achieving (ϵ, δ)-DP. In (ζ, ϵ)-RDP, we directly compare the privacy budget term $\delta(\zeta)$. We mainly focus on the privacy changes on T and K. $\Omega(\cdot)$, $\mathcal{O}(\cdot)$, and $o(\cdot)$ correspond to the lower, upper bound, and not tight upper bound of the complexity, respectively.

	(ϵ, δ)-DP	(ζ, ϵ)-RDP	when $T, K \to \infty$
Wei et al. (2020)	$\sigma = \mathcal{O}\left(\frac{V}{\epsilon m}\sqrt{T^2 - mL^2}\right)$	-	$\sigma \to \infty$ on non-convex
Shi et al. (2021)	$\sigma = \mathcal{O}\left(\frac{V\sqrt{\log\left(\frac{1}{\delta}\right)}}{\epsilon}T\sqrt{K}\right)$	-	
Zhang et al. (2021b)	$\sigma = \mathcal{O}\left(\frac{V\sqrt{\log\left(\frac{1}{\delta}\right)}}{\epsilon m}\sqrt{T + mK}\right)$	-	
Noble et al. (2022)	$\sigma = \Omega\left(\frac{V\sqrt{\log\left(\frac{2T}{\delta}\right)}}{\epsilon\sqrt{m}}\sqrt{TK}\right)$	-	
Cheng et al. (2022)	$\sigma = \Omega\left(\frac{V\sqrt{\log\left(\frac{1}{\delta}\right)}}{\epsilon}\sqrt{T}\right)$	-	
Zhang & Tang (2022)	-	$\epsilon = \Omega\left(\frac{\zeta V^2}{\sigma^2}TK\right)$	
Hu et al. (2023)	$\sigma = \Omega\left(\frac{V\sqrt{\epsilon + 2\log\left(\frac{1}{\delta}\right)}}{\epsilon}\sqrt{T}\right)$	-	
Fukami et al. (2024)	$\sigma = \Omega\left(\frac{V(1+\sqrt{1+\epsilon})\sqrt{\log\left(e + \frac{\epsilon}{\delta}\right)}}{\epsilon}\sqrt{T}\right)$	-	
Bastianello et al. (2024)	-	$\epsilon = \mathcal{O}\left(\frac{\zeta L V^2}{\beta^2\sigma^2}\left(1 - e^{-\beta T}\right)\right)$	convergent on β-strongly convex
Ours (`Noisy-FedAvg`)	$\sigma = o\left(\frac{V\sqrt{(\Phi^{-1}(\delta))^2 + 4\epsilon}}{\epsilon\sqrt{m}}\sqrt{2 - \frac{1}{T}}\right)$	$\epsilon = \mathcal{O}\left(\frac{\zeta V^2}{m\sigma^2}\left(2 - \frac{1}{T}\right)\right)$	convergent on non-convex

4. Convergent DP

Table 4: Comparison of the accuracy under different experimental settings. We select the scale $\imath$ from $[50, 100]$. Each client holds 600 heterogeneous data samples of MNIST or 500 heterogeneou data samples of CIFAR-10. For each scale, we test two settings of the local interval $K = 50$, 10 and 200, respectively. Throughout the entire process, we fix $TK = 30000$. "-" means the trainin loss diverges. Each result is repeated 5 times to compute its mean and variance.

	Noisy Intensity	$m = 50$			$m = 100$		
		$K = 50$	$K = 100$	$K = 200$	$K = 50$	$K = 100$	$K = 200$
MNIST LeNet-5	$\sigma = 1.0$	-	-	-	-	-	-
	$\sigma = 10^{-1}$	$95.40_{\pm 0.18}$	$95.42_{\pm 0.15}$	$95.21_{\pm 0.11}$	$97.32_{\pm 0.14}$	$97.50_{\pm 0.11}$	$97.42_{\pm 0.18}$
	$\sigma = 10^{-2}$	$98.33_{\pm 0.12}$	$98.02_{\pm 0.15}$	$97.88_{\pm 0.12}$	$98.71_{\pm 0.10}$	$97.97_{\pm 0.08}$	$97.72_{\pm 0.12}$
	$\sigma = 10^{-3}$	$98.41_{\pm 0.07}$	$98.23_{\pm 0.03}$	$98.00_{\pm 0.07}$	$98.94_{\pm 0.04}$	$98.50_{\pm 0.06}$	$98.01_{\pm 0.10}$
CIFAR-10 ResNet-18	$\sigma = 1.0$	-	-	-	-	-	-
	$\sigma = 10^{-1}$	$53.76_{\pm 0.25}$	$53.38_{\pm 0.23}$	$53.49_{\pm 0.21}$	$62.02_{\pm 0.28}$	$61.33_{\pm 0.25}$	$61.11_{\pm 0.17}$
	$\sigma = 10^{-2}$	$70.11_{\pm 0.22}$	$69.08_{\pm 0.12}$	$66.63_{\pm 0.16}$	$74.34_{\pm 0.29}$	$72.87_{\pm 0.19}$	$70.74_{\pm 0.15}$
	$\sigma = 10^{-3}$	$70.98_{\pm 0.11}$	$69.81_{\pm 0.20}$	$67.98_{\pm 0.03}$	$75.38_{\pm 0.19}$	$74.44_{\pm 0.12}$	$72.11_{\pm 0.06}$

Table 5: Performance and sensitivity ($T = 600$).

	Accuracy	Sensitivity
Noisy-FedAvg	60.67	31.33
Noisy-FedProx $\alpha = 0.01$	60.69	30.97
Noisy-FedProx $\alpha = 0.1$	**60.94**	**18.52**
Noisy-FedProx $\alpha = 1$	56.33	6.34

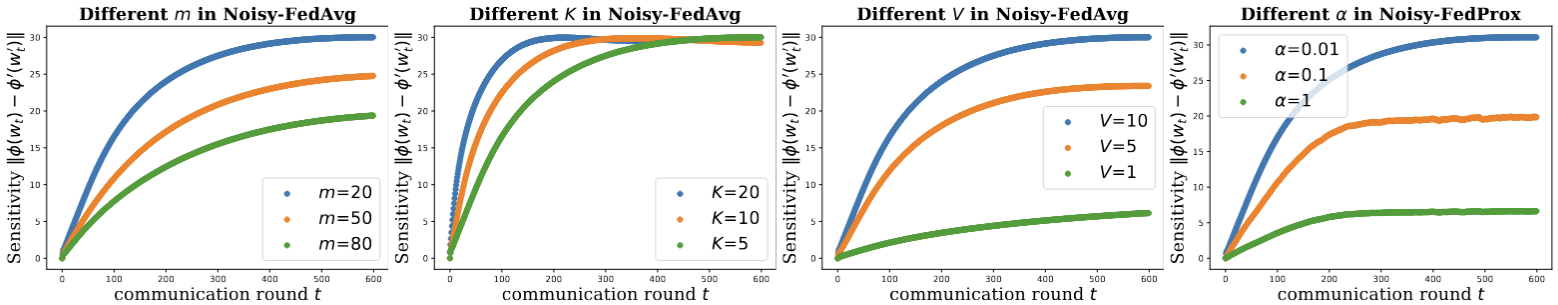


Figure 2: Sensitivity studies on Noisy-FedAvg and Noisy-FedProx. The general setups are $m = 20$, $K = 5$, and $V = 10$. In each group, we keep all other parameters fixed to ensure fairness.

Thank You!

Thanks for attention!

The University of Sydney