On the Convergence of Differentially Private Federated Learning on Non-Lipschitz Objectives via Clipping and Normalized Client Updates

Abolfazl Hashemi, Purdue ECE

FLOW Seminar, April 20th, 2022

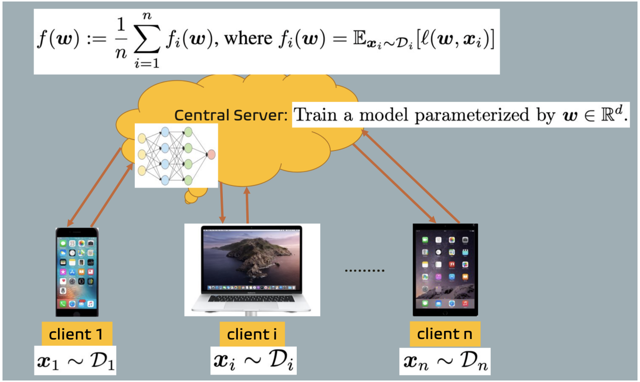
https://arxiv.org/abs/2106.07094

Rudrajit Das (UT CS)

Sujay Sanghavi (UT ECE)

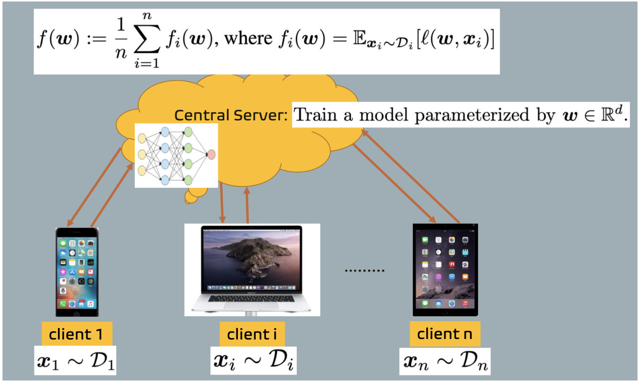
Inderjit S. Dhillon (UT CS)

Collaborative ML via Federated Learning (FL)



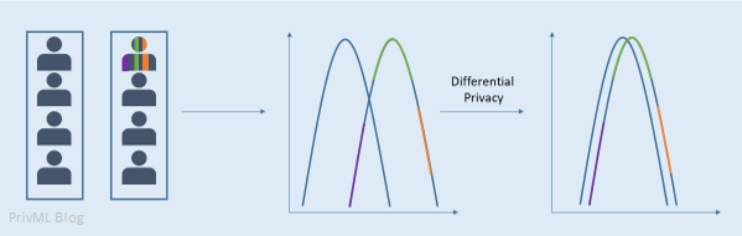
- Decentralized data (as opposed to traditional distributed learning)
- Different (resp., identical) $\mathcal{D}_i$'s: **heterogeneous** (resp., **homogeneous**) setting.

Privacy-Preserving FL



- Despite the locality of data storage in FL, information-sharing opens the door to the possibility of sabotaging the security of personal data through communication.
- Can the server optimize while preserving a strong notion of privacy of clients' data?

- DP is a popular privacy-quantifying framework for training of ML models.
- Goal: Learning nothing about an individual while learning useful information about a whole population
- Reducing learning algorithm's sensitivity to an individual's data
- To protect the individuals' privacy, one adds a controlled amount of random noise to the results of our analysis.



Neighboring datasets

Two datasets $\mathcal{D} \in D_c$ and $\mathcal{D}' \in D_c$ are said to be neighboring if they differ in exactly one sample, and we denote this by $|\mathcal{D} - \mathcal{D}'| = 1$.

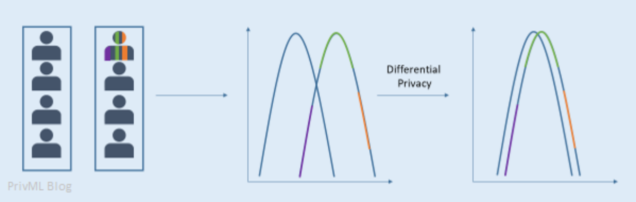
(ε, δ)-DP [DMNS06]

Given a collection of datasets D_c and a query function $h : D_c \to \mathcal{X}$, a randomized mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is said to be (ε, δ)-DP, if for any two neighboring datasets

$$\mathbb{P}(\mathcal{M}(h(\mathcal{D})) \in \mathcal{R}) \leq e^{\varepsilon}\mathbb{P}(\mathcal{M}(h(\mathcal{D}')) \in \mathcal{R}) + \delta.$$

- When $\delta = 0$, it is commonly known as pure DP. Otherwise, it is known as approximate DP.
- Setting $\mathcal{M}$ to additive random Gaussian noise – known as the Gaussian mechanism – is a customary approach to provide DP.

How Much Noise is Adequate?



Gaussian mechanism [DR+14]

Let $\Delta := \sup_{\mathcal{D},\mathcal{D}' \in D_c : |\mathcal{D}-\mathcal{D}'|=1} \|h(\mathcal{D}) - h(\mathcal{D}')\|$. If we set $\mathcal{M}(h(\mathcal{D})) = h(\mathcal{D}) + \boldsymbol{Z}$, where $\boldsymbol{Z} \sim \mathcal{N}\left(\vec{0}_p, \frac{2\log(1.25/\delta)\Delta^2}{\varepsilon^2} \mathrm{I}_p\right)$, then the mechanism $\mathcal{M}$ is (ε, δ)-DP.

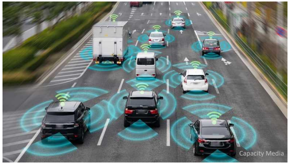
- Noise power increases with sensitivity and desired privacy guarantees.
- Similar results exists for Laplace mechanism and discretized/truncated distributions

Going back to FL: How to apply DP in FL?

DP in FL: FedAvg Revisited

Local optimization

$\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(100)}$



Intermittent downlink communication

$\boldsymbol{w}$

Intermittent uplink communication

$\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(10)}$

Global aggregation

$\boldsymbol{w} = \mathrm{agg}(\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots, \boldsymbol{w}^{(10)})$

- Local update: For E steps, do GD, i.e., $\boldsymbol{w}_{k,\tau+1}^{(i)} \leftarrow \boldsymbol{w}_{k,\tau}^{(i)} - \eta_k \nabla f_i(\boldsymbol{w}_{k,\tau}^{(i)})$
- Each client communicates its update $\boldsymbol{u}_k^{(i)} = \frac{\boldsymbol{w}_k - \boldsymbol{w}_{k,E}^{(i)}}{\eta_k}$ to server w.p. $\frac{r}{n}$ (total of K rounds)
- Global update: $\boldsymbol{w}_{k+1} \leftarrow \boldsymbol{w}_k - \frac{\beta_k}{r} \sum_{i \in \mathcal{S}_k} \boldsymbol{u}_k^{(i)}$
- Output: $\boldsymbol{w}_{\tilde{k}}$ with $\tilde{k} \sim \mathrm{Unif}[0, K-1]$.

$\boldsymbol{u}_k^{(i)}$ contains local gradient information $\rightarrow$ needs to be made private!

DP-FedAvg with Clipping

- Maximum sensitivity, Δ, grows with norm of $\boldsymbol{u}_k^{(i)} = \frac{\boldsymbol{w}_k - \boldsymbol{w}_{k,E}^{(i)}}{\eta_k}$
- Assuming G-Lipschitzness, i.e., $\sup_{\boldsymbol{\theta}\in\Theta} \|\nabla g(\boldsymbol{\theta})\|_2 \leq G$, this norm is at most $GE \rightarrow$ controlled additive noise
- In general, need to limit how large $\boldsymbol{u}_k^{(i)}$ can get to remove unbounded impact of one client.

Clipped Updates: $$\boldsymbol{u}_k^{(i)} \min\left(1, \frac{C}{\|\boldsymbol{u}_k^{(i)}\|}\right) + \zeta_k^{(i)}, \quad \zeta_k^{(i)} \sim \mathcal{N}(0_d, r\sigma^2 \mathrm{I}_d)$$

Theorem (Based on [ACG+16])

There exists an absolute constant $q > 0$ s.t. for $\varepsilon = \mathcal{O}(1)$, DP-FedAvg will be (ε, δ)-DP as long as

$$\sigma^2 = qKC^2 \frac{\log(1/\delta)}{n^2 \varepsilon^2}.$$

Convergence of DP-FedAvg with Clipping

Convexity

A function $g : \Theta \rightarrow \mathbb{R}$ is convex if $g(\lambda\boldsymbol{\theta} + (1-\lambda)\boldsymbol{\theta}') \leq \lambda g(\boldsymbol{\theta}) + (1-\lambda)g(\boldsymbol{\theta}')$ for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ and $0 \leq \lambda \leq 1$.

Smoothness

A function $g : \Theta \rightarrow \mathbb{R}$ is to said to be L-smooth if for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, $\|\nabla g(\boldsymbol{\theta}) - \nabla g(\boldsymbol{\theta}')\|_2 \leq L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$. If g is twice differentiable, then for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$:

$$g(\boldsymbol{\theta}') \leq g(\boldsymbol{\theta}) + \langle \nabla g(\boldsymbol{\theta}), \boldsymbol{\theta}' - \boldsymbol{\theta} \rangle + \frac{L}{2}\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|_2^2.$$

Heterogeneity

Let $\boldsymbol{w}^* \in \arg\min_{\boldsymbol{w}' \in \mathbb{R}^d} f(\boldsymbol{w}')$ and $\Delta_i^* := f_i(\boldsymbol{w}^*) - \min_{\boldsymbol{w}' \in \mathbb{R}^d} f_i(\boldsymbol{w}') \geq 0$. Then the heterogeneity of the system is quantified by some increasing function of the Δ_i^*'s.

Machine Intelligence & Networked Data Science

Theorem 1

Suppose the f_i's are convex and L-smooth over $\mathbb{R}^d$. Define $0 < \rho := \frac{\sqrt{qd\log(1/\delta)}}{n\varepsilon} < 1$. There exists constant local and global learning rates η and β, and a lower bound on the clipping threshold C_{low}, such that for any $C \geq C_{\text{low}}$ and in $K = \mathcal{O}(\frac{L\|\mathbf{w}_0 - \mathbf{w}^*\|}{C\rho^2})$ rounds

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\min\left(f_i(\mathbf{w}_{\tilde{k}}) - f_i(\mathbf{w}^*), \mathcal{O}\left(\frac{C}{LE}\|\nabla f_i(\mathbf{w}_{\tilde{k}})\|\right)\right)\right] \leq \mathcal{O}\left(\frac{C}{E}\|\mathbf{w}_0 - \mathbf{w}^*\| + E\left(\frac{1}{n}\sum_{i=1}^{n}\Delta_i^*\right)\right)\rho.$$

- ρ: the price of privacy – increases as the level of privacy increases (i.e., ε and δ decrease).
- Non-vanishing convergence error in Private FL
- The second term, effect of heterogeneity, can be reduced arbitrarily by increasing K, but the first term, effect of initialization remains.

Friendlier Results Under Lipschitzness

- Under G-Lipschitzness we can set $C = GE$ since $\|\boldsymbol{u}_k^{(i)}\| \leq GE$
- This means no clipping occurs and we can get rid of $\mathcal{O}\left(\frac{C}{LE}\|\nabla f_i(\boldsymbol{w}_{\tilde{k}})\|\right)$ from the convergence criterion to obtain

$$\mathbb{E}[f(\boldsymbol{w}_{\tilde{k}})] - f(\boldsymbol{w}^*) \leq \frac{8}{5}\left(G\|\boldsymbol{w}_0 - \boldsymbol{w}^*\| + \frac{3}{4}E\left(\frac{1}{n}\sum_{i=1}^{n}\Delta_i^*\right)\right)\rho.$$

- With $E = \mathcal{O}(1)$ our bound matches the lower bound for the centralized convex and Lipschitz case with respect to the dependence on ρ

Are multiple local steps ($E > 1$) beneficial or detrimental?

Effect of Multiple Local Steps

- In a nutshell, increasing E mitigates the effect of initialization at the cost of increasing the effect of heterogeneity; the "best" value of E depends on which one is more dominant, and also the privacy level.
- Let us quantify this with an additional assumption.

Assumption 1

(i) For any $\boldsymbol{w} \in \mathbb{R}^d$ and each $i \in [n]$, we have $\|\nabla f_i(\boldsymbol{w} - \eta \nabla f_i(\boldsymbol{w})) - \nabla f_i(\boldsymbol{w})\| \geq \eta\lambda \|\nabla f_i(\boldsymbol{w})\|$, for some $0 < \lambda \leq L$ and $\eta \leq \frac{\rho}{2L}$. (ii) Additionally, each f_i is G-Lipschitz over $\mathbb{R}^d$.

- For small enough η, $\|\nabla f_i(\boldsymbol{w} - \eta \nabla f_i(\boldsymbol{w})) - \nabla f_i(\boldsymbol{w})\| = \Theta(\eta \|\nabla^2 f_i(\boldsymbol{w}) \nabla f_i(\boldsymbol{w})\|)$; so, we are basically assuming $\|\nabla^2 f_i(\boldsymbol{w}) \nabla f_i(\boldsymbol{w})\| \geq \Omega(\lambda \|\nabla f_i(\boldsymbol{w})\|)$ which is weaker than strong convexity.

Effect of Multiple Local Steps

- Recall $\rho = \mathcal{O}\left(\frac{\sqrt{d\log(1/\delta)}}{n\varepsilon}\right)$ is the privacy cost.

Proposition 1

Under Assumption 1, there exists a choice of C (depending on E), s.t. we get the following convergence guarantee:

$$\mathbb{E}[f(\boldsymbol{w}_{\tilde{k}})] - f(\boldsymbol{w}^*) \leq \left(2G\|\boldsymbol{w}_0 - \boldsymbol{w}^*\| + \frac{6}{5}E\left\{\frac{1}{n}\sum_{i=1}^{n}\Delta_i^* - \frac{11G\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|\rho}{48}\left(\frac{\lambda^2}{L^2}\right)\right\}\right)\rho.$$

- So if $\frac{1}{n}\sum_{i=1}^{n}\Delta_i^* < \mathcal{O}\left(G\left(\frac{\lambda^2}{L^2}\right)\right)\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|\rho$, then having a large value of E is beneficial; in particular, setting the maximum permissible value of E, which is $\frac{1}{2\rho}$, is the best (in terms of smallest suboptimality gap). Otherwise, having a small value of E is better; specifically, $E = 1$ is the best.

So far we discussed

- Convergence of DP-FedAvg with clipping with and without Lipschitzness
- Role of E in DP-FedAvg with clipping

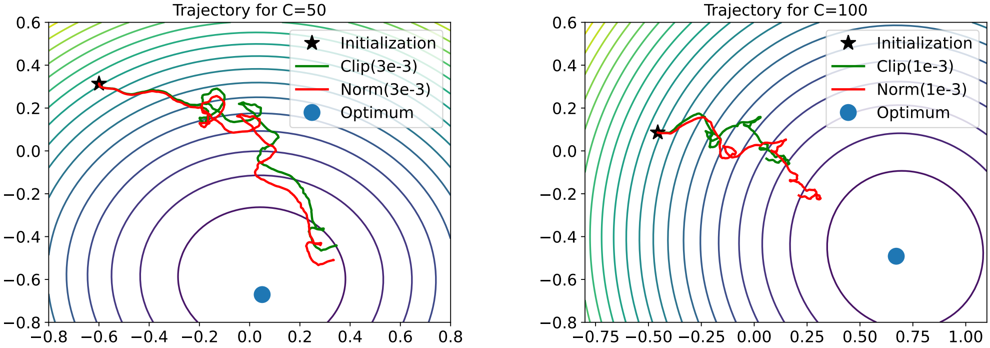
But clipping has a potential issue!

$$\mathrm{clip}(\boldsymbol{z}, c) := \boldsymbol{z} \min\left(1, \frac{c}{\|\boldsymbol{z}\|}\right).$$

- As our FL algorithm converges, norm of model update $\boldsymbol{u}_k^{(i)} = \frac{\boldsymbol{w}_k - \boldsymbol{w}_{k,E}^{(i)}}{\eta}$ becomes small $\rightarrow$ no clipping occurs
- But $\sigma \propto C$ regardless (can we adaptively reduce C?)
- Hence, we enter a low SNR regime where added noise is dominant and hurts the convergence

Low SNR near Solution

- DP-FL on synthetic convex quadratic functions



- How can we ensure clients' update have higher SNR values?

Normalized Updates in DP-FL

Client-Update Normalization (Instead of Clipping)

- We propose to use

$$\mathrm{norm}(\boldsymbol{z}, c) := \frac{c\boldsymbol{z}}{\|\boldsymbol{z}\|} \qquad \text{vs.} \qquad \mathrm{clip}(\boldsymbol{z}, c) := \boldsymbol{z}\min\left(1, \frac{c}{\|\boldsymbol{z}\|}\right)$$

  in DP-FedAvg, i.e.,

$$\text{Normalized Updates:} \qquad \frac{C\boldsymbol{u}_k^{(i)}}{\|\boldsymbol{u}_k^{(i)}\|} + \zeta_k^{(i)}, \quad \zeta_k^{(i)} \sim \mathcal{N}(0_d, r\sigma^2 \mathrm{I}_d)$$

- This ensures the updates are uniformly bounded and at the same time noise will not overpower the update direction, leading to better convergence and accuracy.
- For smaller C, normalization and clipping become equivalent.

Theoretical Comparison of Clipping and Normalization

See Section 5.1 and Remark 1 in the paper for a precise comparison. In summary:

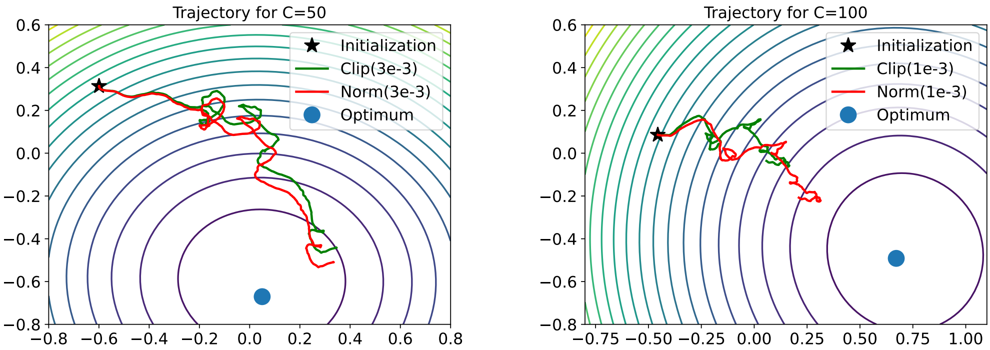
- Our theory shows normalization enjoys a smaller effect of initialization on convergence.
- Not easy to characterize whether the effect of heterogeneity is smaller for normalization or clipping.
- But recall, the effect of heterogeneity can be controlled by increasing K, the number of rounds.
- Hence, normalization has a better asymptotic convergence.

Theory-guided Recommendation

Do normalization if we can afford training for large K.

Experiments: Synthetic Convex Quadratic Problems

- We set $(\varepsilon, \delta) = (5, 10^{-6})$, $K = 500$, $E = 20$, and $n = 100$.
- $f_i(\boldsymbol{w}) = \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_i^*)^T \boldsymbol{Q}_i(\boldsymbol{w} - \boldsymbol{w}_i^*)$
- $\boldsymbol{w}_i^*$ and $\boldsymbol{Q}_i = \boldsymbol{A}_i \boldsymbol{A}_i^T$, where $\boldsymbol{A}_i$ is 200×20, are formed randomly.
- Two different initialization
  - **I1**: $\boldsymbol{w}_0 = \boldsymbol{w}^* + \boldsymbol{z}$, and
  - **I2**: $\boldsymbol{w}_0 = \boldsymbol{w}^* + \frac{\boldsymbol{z}}{5}$,
- Finally, we consider full-device participation and vary η and C.

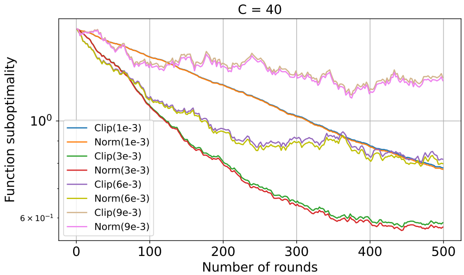


(a) l1: $C = 50$ and $\eta = 0.003$

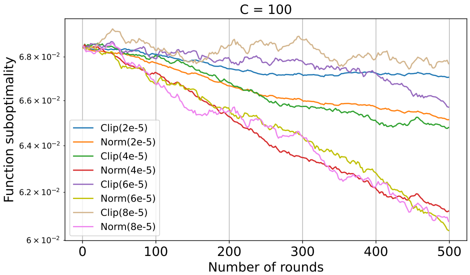
(b) l1: $C = 100$ and $\eta = 0.001$

- `DP-NormFedAvg` reaches closer to the optimum than `DP-FedAvg` with clipping.

Synthetic Problems: Convergence Curves



(a) I1: $C = 40$

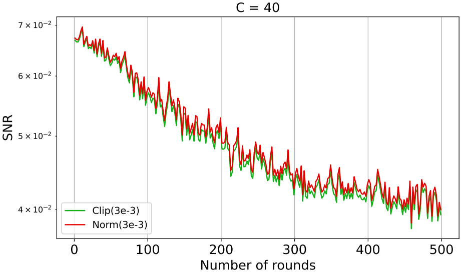


(b) I2: $C = 100$

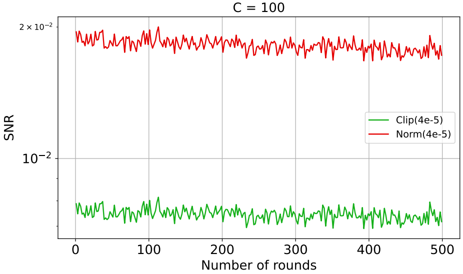
- Normalization does significantly better than clipping for large C.
- For smaller C normalization and clipping are nearly equivalent (as expected).

Synthetic Problems: SNR Comparison

- SNR: The ratio of average clipped/normalized per-client update and average per-client noise



(a) I1: $C = 40$



(b) I2: $C = 100$

- SNR of normalization is never lower than that of clipping, explaining the superiority of the former.

Logistic Regression on FMNIST, CIFAR-10 and CIFAR-100

- Average test accuracy over the last 5 rounds

FMNIST	$(5, 10^{-5})$-DP	$(1.5, 10^{-5})$-DP
Clipping	75.59%	56.90%
Normalization	**77.72%**	**57.80%**
FedAvg (w/o privacy)	83.43%	

CIFAR-10	$(5, 10^{-5})$-DP	$(1.5, 10^{-5})$-DP
Clipping	82.63%	81.53%
Normalization	**84.21%**	**82.42%**
FedAvg (w/o privacy)	85.64%	

CIFAR-100	$(5, 10^{-5})$-DP	$(1.5, 10^{-5})$-DP
Clipping	56.53%	41.33%
Normalization	**59.36%**	**42.76%**
FedAvg (w/o privacy)	64.61%	

Summary

Goal

Convergence analysis of private federated learning

- Established convergence of DP-FedAvg with clipping without Lipschitzness on smooth convex functions
- Effect of heterogeneity can be controlled while effect of initialization remains (cannot hope to do better)
- Role of local steps E: If $\frac{1}{n}\sum_{i=1}^{n}\Delta_i^* < \mathcal{O}\left(G\left(\frac{\lambda^2}{L^2}\right)\right)\|\boldsymbol{w}_0 - \boldsymbol{w}^*\|\rho$ having a large value of E is beneficial (under a suitable hessian assumption).

Theory-guided Recommendation

Normalized client updates instead of clipping for DP-FedAvg

- Ensures updates enjoy higher SNR
- Theoretical advantage over clipping in mitigating the effect of initialization

Thank you!

On the Convergence of Differentially Private Federated Learning on Non-Lipschitz Objectives via Clipping and Normalized Client Updates

https://arxiv.org/abs/2106.07094

Hiring Postdocs and PhD Students at Purdue ECE!

Abolfazl Hashemi (email: abolfazl@purdue.edu)