Understanding Clipped FedAvg: Convergence and Client-Level Differential Privacy

Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Jinfeng Yi

University of Minnesota

July 11, 2022

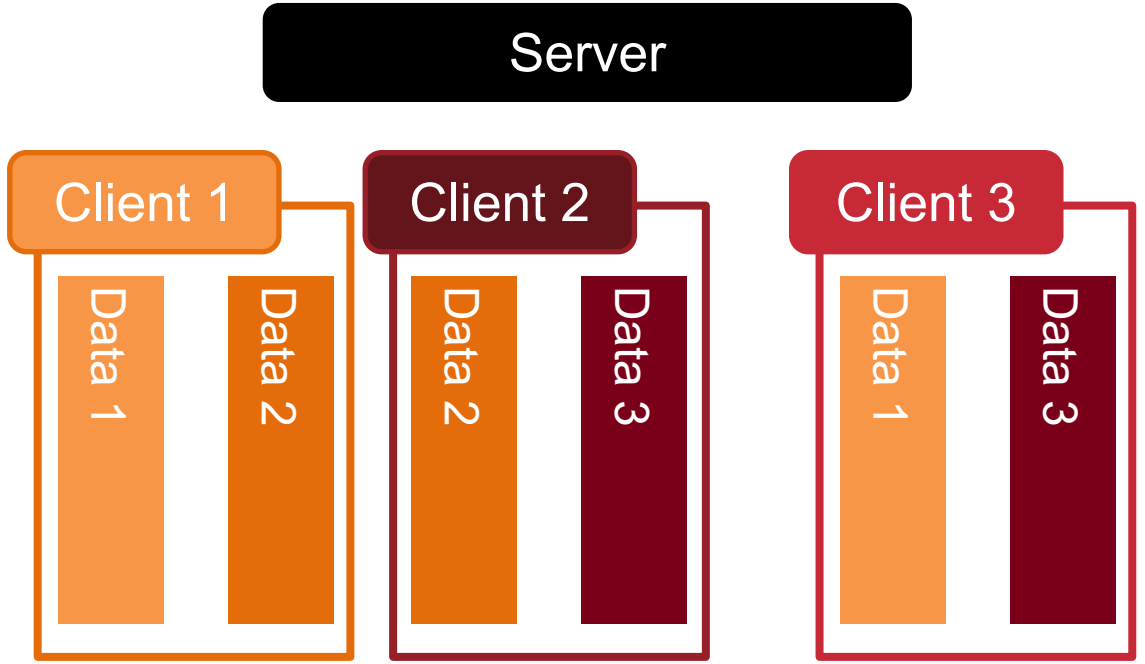
❑ Introduction

- Federated Learning
- Differential Privacy
- Gaussian Mechanism
- Clipping Operation

❑ Horizontal FL

Horizontal FL

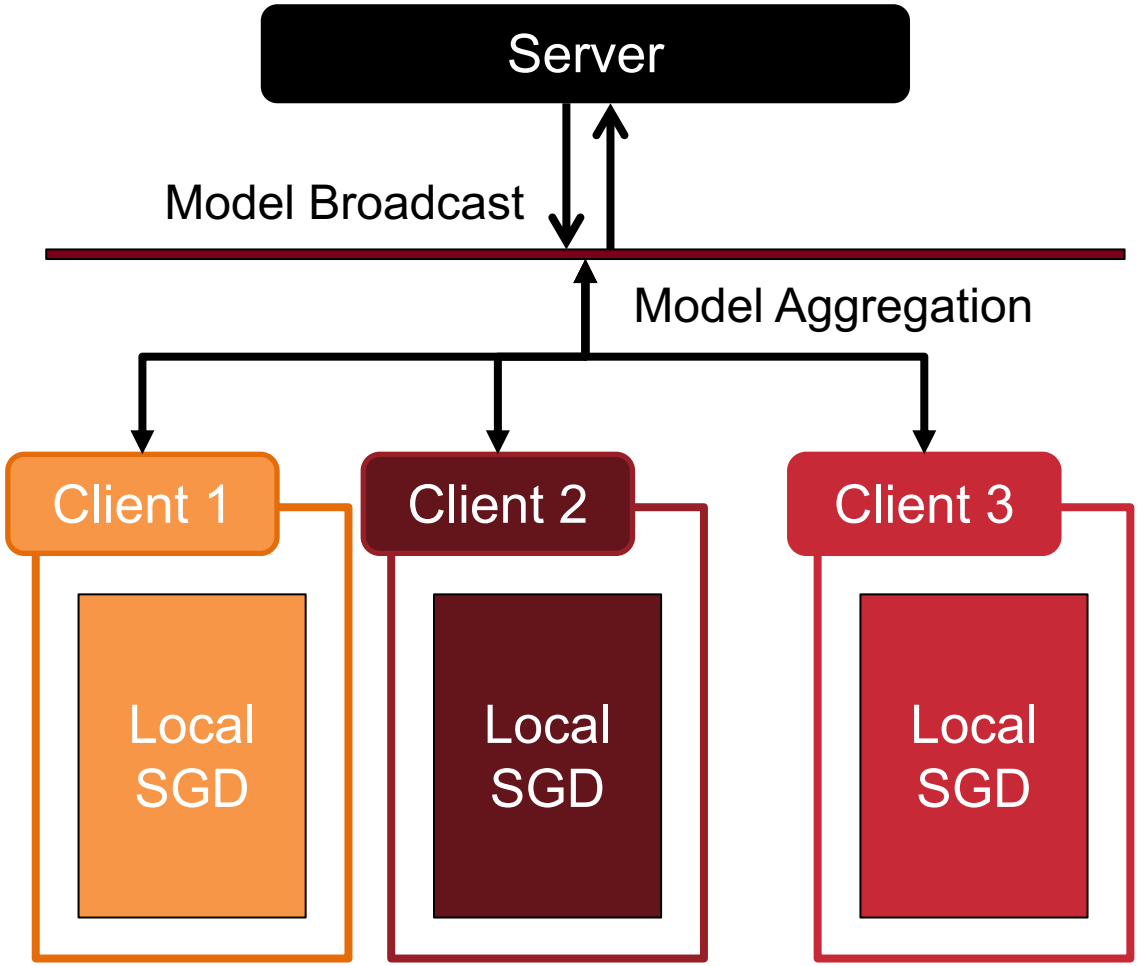
- Partial samples, same model
- Application
  - Cross-device
    - Google Keyboard [1]
  - Cross-silo
    - Medical image classification [2]



[1] Bonawitz K, Eichner H, Grieskamp W, et al. Towards federated learning at scale: System design. arXiv 2019.
[2] Li W, Milletarì F, Xu D, et al. Privacy-preserving federated brain tumour segmentation, MLMI 2019 .

❑ Privacy in FL

- FedAvg Algorithm:
  - Global Averaging
  - Local SGD
- Cross-Device [1]
  - Protect privacy of each client (Application user)
  - $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^{N}$
- Cross-Silo [2]
  - Protect privacy of each sample (patients' record)
  - $\mathcal{D} = \bigcup \mathcal{D}_i$



[1] Bonawitz K, Eichner H, Grieskamp W, et al. Towards federated learning at scale: System design. arXiv 2019.
[2] Li W, Milletarì F, Xu D, et al. Privacy-preserving federated brain tumour segmentation, MLMI 2019 .
[3] Agarwal, et al. cpSGD: communication-efficient and differentially-private distributed SGD. NurIPS 2018.

❑ Differential Privacy

Gaussian DP

A randomized mechanism $\mathcal{A}$ satisfies (ϵ, δ)-DP, if for all measurable sets $\mathcal{S} \subseteq Range(\mathcal{A})$ and for any two adjacent data sets $\mathcal{D}, \mathcal{D}'$,

$$\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon} \Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{S}] + \delta\,.$$

- Intuition: the output of the algorithm should not change too much by changing one input data
- Sample privacy: $\mathcal{D} = \bigcup \mathcal{D}_i$, $\mathcal{D}$ and $\mathcal{D}'$ vary by one sample ξ.
- Client privacy: $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^{N}$, $\mathcal{D}$ and $\mathcal{D}'$ vary by one client $\mathcal{D}_i$.

1. Abadi et al, Deep learning with differential privacy. Conference on Computer and Communications Security 2016.

❑ Gaussian Mechanism & Clipping

Gaussian Mechanism

Given an algorithm $\mathcal{A}$, by adding Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ to the output of $\mathcal{A}$. With $\sigma = \Delta_2(\mathcal{A})\sqrt{2log(1.25/\delta)}/\epsilon$ the mechanism is (ϵ, δ)-DP for any $\epsilon, \delta \in (0,1)$.

$$\ell_2 \text{ sensitivity: } \Delta_2(\mathcal{A}) = \max_{|\mathcal{D}-\mathcal{D}'|=1} \|\mathcal{A}(\mathcal{D}) - \mathcal{A}(\mathcal{D}')\|^2$$

- To bound ℓ_2 sensitivity, we need to clip the output of the local updates in FL:

$$\text{clip}(\Delta \mathbf{x}_i^t, c) = \Delta \mathbf{x}_i^t \cdot \min\left\{1, \frac{c}{\|\Delta \mathbf{x}_i^t\|}\right\}$$

❑ Contribution

- The first convergence result for DP-FedAvg with clipping
  - Provide error decomposition
- Numerical results show:
  - How clipping affects the performance of FedAvg in different settings
  - How privacy noise affects the performance on FedAvg

❑ Algorithm Design

- DP-FedAvg
- Clipping Bias
- Understanding Error Terms

❑ DP-FedAvg

Algorithm 4 DP-FedAvg Algorithm

Initialize: $\mathbf{x}_i^0 \triangleq \mathbf{x}^0, i = 1, \ldots, N$
for $t = 0, \ldots, T-1$ *(stage)* **do**
  for $i \in \mathcal{P}_t \subseteq [N]$ in parallel **do** [Client Subsampling]
    Update agents' $\mathbf{x}_i^{t,0} = \mathbf{x}^t$
    for $q = 0, \ldots, Q-1$ *(iteration)* **do**
      Compute stochastic gradient $g_i^{t,q}$ with $\mathbb{E}[g_i^{t,q}] = \nabla f_i(x_i^{t,q})$
      Local update: $\mathbf{x}_i^{t,q+1} = \mathbf{x}_i^{t,q} - \eta_l g_i^{t,q}$
    Compute update difference: $\Delta \mathbf{x}_i^t = \mathbf{x}_i^{t,Q} - \mathbf{x}_i^{t,0}$
    Clip and perturb: $\tilde{\Delta}\mathbf{x}_i^t = \text{clip}(\Delta \mathbf{x}_i^t, c) + \mathbf{z}_i^t$, [Apply Gaussian Mechanism]
  Global averaging: $\mathbf{x}^{t+1} = \mathbf{x}^t + \eta_g \frac{1}{|\mathcal{P}_t|} \sum_{i \in \mathcal{P}_t} \tilde{\Delta}\mathbf{x}_i^t$

❑ Clipping Bias

- Clipping introduces bias to FedAvg;
- Bias is related to
  - Model (AlexNet, ResNet, etc.)
  - Data set (EMNIST, Cifar-10, etc.)
  - Data distribution (IID, Non-IID)

Model	Data set	IID (%)	IID Clipping (% drop)	Non-IID (%)	Non-IID Clipping (% drop)
AlexNet	EMNIST	98.2	0.19	95.6	3.60
	Cifar-10	66.01	4.83	57.14	7.30
ResNet-18	EMNIST	99.61	0.02	95.43	0.10
	Cifar-10	76.36	0.53	59.46	1.55

❑ Clipping Bias & Update Distribution

- Update magnitude:

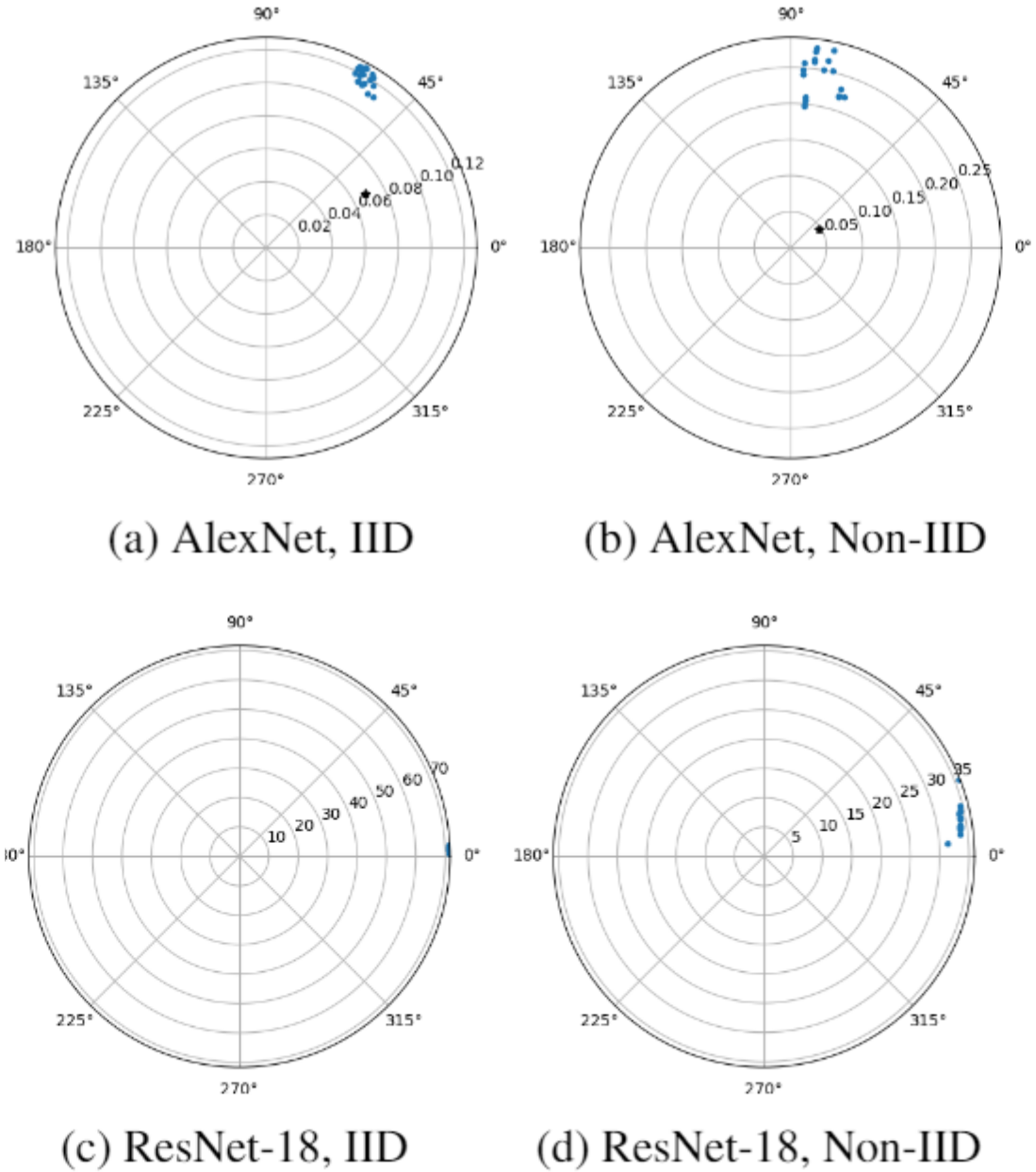
$$\left\|\Delta x_i^t\right\|^2 = \left\|x_i^{t,Q} - x_i^{t,0}\right\|^2$$

- Update direction:

$$\cos^{-1}\left(\Delta x_i^t \cdot \overline{\Delta x^{t-1}}\right)$$

- **More concentrated update**

  → **smaller accuracy drop**.



(a) AlexNet, IID (b) AlexNet, Non-IID

(c) ResNet-18, IID (d) ResNet-18, Non-IID

❑ Algorithm Convergence

Theorem 3

Suppose A2–A3 hold, A1 hold with D=1 and finite G. Set $P_t = [N]$, $\eta_l \eta_g \leq \frac{1}{3QL}$, $\eta_l \leq \frac{1}{8QL}$, *then we have*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\bar{a}^t \|\nabla f(x^t)\|^2] \leq \mathcal{O}\left(\frac{1}{\eta_g \eta_l QT} + \eta_l^2 Q^2 + \frac{\eta_g \eta_l}{N}\right)$$

FedAvg standard terms

Clipping error

$$+\mathcal{O}\left(\mathbb{E}\left[\left|a_i^t - \tilde{a}_i^t\right| + \left|\tilde{a}_i^t - \bar{a}^t\right|\right] + \eta_g \eta_l Q\right) + \mathcal{O}\left(\frac{\eta_g d \sigma^2}{\eta_l QN}\right).$$

Privacy Noise

Where $a_i^t = \frac{1}{\max(c, \|\Delta x_i^t\|)}$, $\tilde{a}_i^t = \frac{1}{\max(c, \|\mathbb{E}\Delta x_i^t\|)}$, $\bar{a}^t = \frac{1}{|P_t|}\sum_{i \in P_t} \tilde{a}_i^t$.

- $\left|a_i^t - \tilde{a}_i^t\right|$: stochastic error, 0 when using local GD.
- $\left|\tilde{a}_i^t - \bar{a}^t\right|$: Heterogeneity update error, 0 when

  1) $c \geq \|\mathbb{E}\Delta x_i^t\|$ (no clipping) or 2) all $\|\mathbb{E}\Delta x_i^t\|$'s are the same (homogeneous data).

Error Decomposition

- Clipping Error $\mathcal{O}\left(\mathbb{E}\left[\left|a_i^t - \tilde{a}_i^t\right| + \left|\tilde{a}_i^t - \bar{a}^t\right|\right] + \eta_g \eta_l Q\right)$
  - $\left|a_i^t - \tilde{a}_i^t\right|$: stochastic error, 0 when using local GD.
  - $\left|\tilde{a}_i^t - \bar{a}^t\right|$: heterogeneity update error, 0 when
    - 1) $c \geq \left\|\mathbb{E}\Delta x_i^t\right\|$ (no clipping);
    - 2) all $\left\|\mathbb{E}\Delta x_i^t\right\|$'s are the same (homogeneous data)
  - $\eta_g \eta_l Q$: diminishing update error
- Privacy Noise $\mathcal{O}\left((\eta_g d \sigma^2)/(\eta_l Q N)\right)$
  - Scales with model size d
  - Inverse scaling with client number N

❑ Numerical Results

- Settings
- Results

❑ Numerical Result

- EMNIST
- 1920 clients, $|P_t| = 80$
- $(1.5, 10^{-5})$-DP for MLP, AlexNet, MobileNetV2
- $(5, 10^{-5})$-DP for ResNet-18

Model	FedAvg	Clipping drop	DP drop
MLP	94.0	1.84	0.29
AlexNet	96.4	1.47	**0.16**
MobileNetV2	97.8	0.35	1.62
ResNet-18	95.2	**-0.15**	3.76

- Cifar-10
- 1920 clients, $|P_t| = 80$
- $(1.5, 10^{-5})$-DP for MLP, AlexNet, ResNet-18

Model	FedAvg	Clipping drop	DP drop
MLP	51.9	7.39	0.9
AlexNet	66.0	4.83	**-0.18**
ResNet-18	76.4	**0.53**	5.15

- Thank you!