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)
- Di□erent (resp., identical) 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?

Di□erential Privacy (DP)

- 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.

Di□erential Privacy (DP) Background

Neighboring datasets

Two datasets $D \in D_c$ and $D' \in D_c$ are said to be neighboring if they di□er in exactly one sample, and we denote this by $jD \quad D'j = 1$.

(";)-DP [DMNS06]

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

$$P(M(h(D)) \in R) \quad e^{"}P(M(h(D')) \in R) + \; :$$

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

How Much Noise is Adequate?

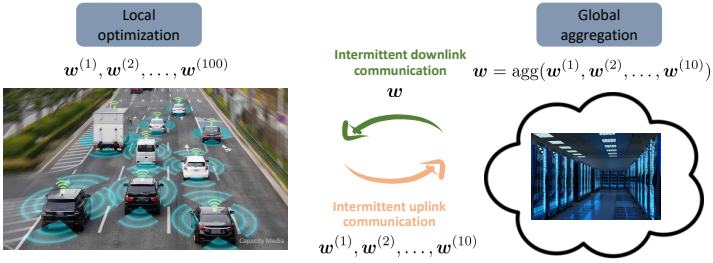
Gaussian mechanism [DR+14]

Let $:= \sup_{D;D^0 2 D_c : jD \quad D^0 j=1} kh(D) \quad h(D^0)k$. If we set $M(h(D)) = h(D) + Z;$ where $Z \quad N \quad 0_p; \frac{2\log(1:25= \ )}{"2}^2 I_p$, then the mechanism M is $(";\ )$-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 update: For E steps, do GD, i.e., $w_{k;\ +1}^{(i)} \quad w_{k;}^{(i)} \quad {}_k\, r\, f_i(w_{k;}^{(i)})$
- Each client communicates its update $u_k^{(i)} = \frac{w_k \quad w_{k;E}^{(i)}}{k}$ to server w.p. $\frac{r}{n}$ (total of K rounds)
- Global update: $w_{k+1} \quad w_k \quad \frac{k}{r} \quad \sum_{i\,2\,S_k} u_k^{(i)}$
- Output: w_k with $k \quad \mathrm{Unif}[0; K \quad 1]$.

$u_k^{(i)}$ contains local gradient information ! needs to be made private!

DP-FedAvg with Clipping

- Maximum sensitivity, , grows with norm of $u_k^{(i)} = \frac{w_k \quad w_{k;E}^{(i)}}{k}$
- Assuming G-Lipschitzness, i.e., $\sup_{\theta 2} \quad k r g(\ )k_2 \quad G$, this norm is at most $GE \quad !$ controlled additive noise
- In general, need to limit how large $u_k^{(i)}$ can get to remove unbounded impact of one client.

Clipped Updates: $u_k^{(i)} \min \ 1; \frac{C}{k u_k^{(i)} k} \ + \ {}_k^{(i)}; \quad {}_k^{(i)} \quad N(0_d; r \ {}^2 I_d)$

Theorem (Based on [ACG+16])

There exists an absolute constant $q > 0$ s.t. for $" = O(1)$, DP-FedAvg will be $(";\ )$-DP as long as

$$^2 = qKC^2 \frac{\log(1= \ )}{n^2 "^2}:$$

Convergence of DP-FedAvg with Clipping

Relevant Notations and Definitions

Convexity

A function $g : \ ! \ \mathbb{R}$ is convex if $g(\ + (1 \ ) \ ^{0}) \ \ g(\ ) + (1 \ )g(\ ^{0})$ for any $\ ; \ ^{0} 2 \ $ and $0 \ \ 1$.

Smoothness

A function $g : \ ! \ \mathbb{R}$ is to said to be L-smooth if for all $\ ; \ ^{0} 2 \ $, $k r g(\ ) \ \ r g(\ ^{0})k_2 \ \ Lk \ \ ^{0}k_2$. If g is twice di□erentiable, then for all $\ ; \ ^{0} 2 \ $:

$$g(\ ^{0}) \ \ g(\ ) + h r g(\ ); \ ^{0} \ \ i + \frac{L}{2} k \ ^{0} \ \ k_2^2:$$

Heterogeneity

Let $w \ \ 2 \arg\min_{w^0 2 \mathbb{R}^d} f(w^0)$ and $\ _i := f_i(w \ ) \ \ \min_{w^0 2 \mathbb{R}^d} f_i(w^0) \ \ 0$. Then the heterogeneity of the system is quantified by some increasing function of the $\ _i$'s.

Convergence Without Assuming Lipschitzness

Theorem 1

Suppose the f_i's are convex and L-smooth over $\mathbb{R}^d$. Define $0 < \quad := \frac{p}{\frac{qd\log(1=\ )}{n''}} < 1$. There exists constant local and global learning rates and , and a lower bound on the clipping threshold C_{low}, such that for any $C \quad C_{\text{low}}$ and in $K = O\left(\frac{Lkw_0 \quad w \ k}{C^{\ 2}}\right)$ rounds

$$E\left[\frac{1}{n}\sum_{i=1}^{n} \min \quad f_i(w_k) \quad f_i(w\ ); O\left(\frac{C}{LE}\right) k\, r\, f_i(w_k) k\right] \quad ! \# \quad O\left(\frac{C}{E}\right) kw_0 \quad w \ k + E\left[\frac{1}{n}\sum_{i=1}^{n} \quad _i\right] \quad !\ :$$

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

Friendlier Results Under Lipschitzness

- Under G-Lipschitzness we can set $C = GE$ since $\|u_k^{(i)}\| \quad GE$
- This means no clipping occurs and we can get rid of $O\left(\frac{C}{LE}\|r f_i(w_k)\| \quad \right)$ from the convergence criterion to obtain

$$\mathrm{E}[f(w_k)] \quad f(w^\star) \quad \frac{8}{5} \quad G\|w_0 \quad w^\star\| + \frac{3}{4}E \quad \frac{1}{n} \quad \sum_{i=1}^{N} \quad i \quad \Bigg) \quad :$$

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

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

E□ect of Multiple Local Steps

- In a nutshell, increasing E mitigates the e□ect of initialization at the cost of increasing the e□ect 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 $w \in \mathbb{R}^d$ and each $i \in [n]$, we have $k \nabla f_i(w \quad \nabla f_i(w)) \quad \nabla f_i(w)k \quad k \nabla f_i(w)k$, for some $0 < \quad L$ and $\quad \frac{}{2L}$. (ii) Additionally, each f_i is G-Lipschitz over $\mathbb{R}^d$.

- For small enough $, k \nabla f_i(w \quad \nabla f_i(w)) \quad \nabla f_i(w)k = \quad (\quad k \nabla^2 f_i(w) \nabla f_i(w)k)$; so, we are basically assuming $k \nabla^2 f_i(w) \nabla f_i(w)k \quad (\quad k \nabla f_i(w)k)$ which is weaker than strong convexity.

E□ect of Multiple Local Steps

- Recall $= O\left(\frac{p}{d\log(1=\ )}\right)$ 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:

$$\mathrm{E}[f(w_k)] \quad f(w\ ) \quad 2Gkw_0 \quad w\ k + \frac{6}{5}E\left(\frac{1}{n}\sum_{i=1}^{n} \quad i\right) \quad \frac{11Gkw_0 \quad w\ k}{48}\left(\frac{2}{L^2}\right) \quad o \quad :$$

- So if $\frac{1}{n}\sum_{i=1}^{n} \quad i < O\left(G\ \frac{2}{L^2} \quad kw_0 \quad w\ k\right)$, then having a large value of E is beneficial; in particular, setting the maximum permissible value of E, which is $\frac{1}{2}$, is the best (in terms of smallest suboptimality gap). Otherwise, having a small value of E is better; specifically, $E = 1$ is the best.

Roadmap

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!

$$\text{clip}(z; c) := z \min\left(1; \frac{c}{\|z\|}\right) :$$

- As our FL algorithm converges, norm of model update $u_k^{(i)} = \frac{w_k - w_{k;E}^{(i)}}{\eta}$ becomes small ! no clipping occurs
- But / 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

$$\text{norm}(z; c) := \frac{cz}{\|z\|} \quad \text{vs.} \quad \text{clip}(z; c) := z \min\left(1; \frac{c}{\|z\|}\right)$$

in DP-FedAvg, i.e.,

$$\text{Normalized Updates:} \quad \frac{C u_k^{(i)}}{\|u_k^{(i)}\|} + \zeta_k^{(i)}; \quad \zeta_k^{(i)} \sim \mathcal{N}(0_d; r^2 \mathrm{I}_d)$$

- 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 e□ect of initialization on convergence.
- Not easy to characterize whether the e□ect of heterogeneity is smaller for normalization or clipping.
- But recall, the e□ect 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 a□ord training for large K.

Experiments: Synthetic Convex Quadratic Problems

- We set $(\varepsilon; \delta) = (5; 10^{-6})$, $K = 500$, $E = 20$, and $n = 100$.
- $f_i(w) = \frac{1}{2}(w - w_i)^T Q_i (w - w_i)$
- w_i and $Q_i = A_i A_i^T$, where A_i is 200×20, are formed randomly.
- Two different initialization
  - **I1**: $w_0 = w + z$, and
  - **I2**: $w_0 = w + \frac{z}{5}$,
- Finally, we consider full-device participation and vary σ and C.

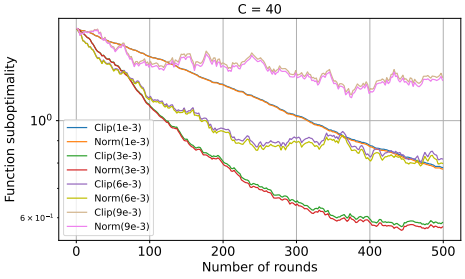
Synthetic Problems: 2D Trajectories

(a) I1: $C = 50$ and $\eta = 0.003$

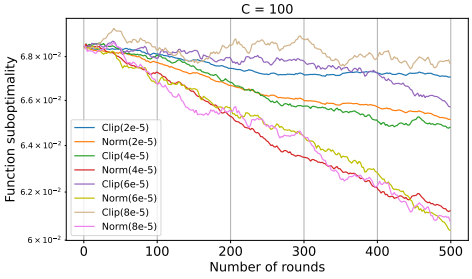
(b) I1: $C = 100$ and $\eta = 0.001$

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

Synthetic Problems: Convergence Curves



(a) I1: $C = 40$

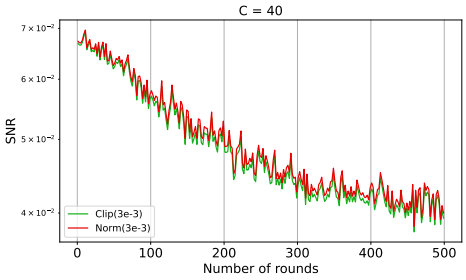


(b) I2: $C = 100$

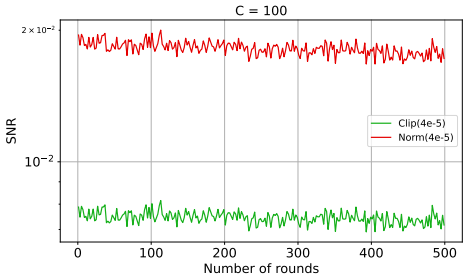
- Normalization does significantly better than clipping for large C.
- For smaller C normalization and clipping are nearly equivalent (as expected).

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
- E□ect of heterogeneity can be controlled while e□ect of initialization remains (cannot hope to do better)
- Role of local steps E: If $\frac{1}{n}\sum_{i=1}^{n} \quad _i < O\ G\ \frac{2}{L^2}\quad kw_0 \quad w\ k$ 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 e□ect 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)