Federated Learning with Client Subsampling, Data Heterogeneity, and Unbounded Smoothness: A New Algorithm and Lower Bounds

Michael Crawshaw [1] Yajie Bao [2] Mingrui Liu [1]

[1]Department of Computer Science, George Mason University

[2]School of Mathematical Sciences, Shanghai Jiao Tong University

November 13, 2023

Problem Statement

Problem: Federated optimization with:

1. Heterogeneous data
2. Partial client participation
3. Relaxed smoothness.

Problem Statement (Federated Learning)

Federated learning McMahan et al. [2017] is a distributed learning framework emphasizing:

- Decentralized data to maintain privacy.
- Minimizing communication between clients.
- Heterogeneous data.

Example: Gmail next word prediction.

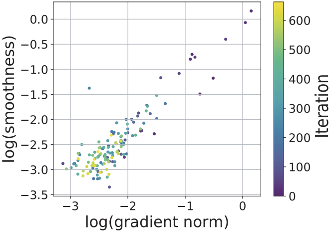
How to efficiently learn from heterogeneous user data (and leverage compute from user devices) while maintaining privacy and minimizing communication cost?

Problem Statement (Relaxed Smoothness)

Works in nonconvex optimization commonly assume smoothness of objective function Ghadimi and Lan [2013], Carmon et al. [2017], i.e. that the gradient is L-Lipschitz.

Zhang et al. [2020a] provide empirical evidence that some neural networks (e.g. RNNs) do not satisfy smoothness assumption.

They introduce a weaker assumption: "relaxed smoothness", where the smoothness constant may grow linearly with the gradient norm.



In this setting, gradient clipping significantly speeds up convergence Zhang et al. [2020a,b].

Problem Statement

Our goal: Design an optimization algorithm for federated learning with heterogeneous data, partial client participation, and relaxed smoothness.

Matches real-world: modern neural networks (relaxed smoothness) with real user data (heterogeneous) and user availability (partial participation).

Previous work:

- SCAFFOLD Karimireddy et al. [2020]: Heterogeneous data with **smoothness**.
- CELGC Liu et al. [2022]: Relaxed smoothness with **homogeneous data**.
- EPISODE Crawshaw et al. [2022]: Relaxed smoothness and heterogeneous data with **full participation**.

Our algorithm, EPISODE++, solves this optimization problem under **heterogeneous data**, **partial participation** and **relaxed smoothness**.

EPISODE++ Algorithm

Algorithm 1 EPISODE++

Initialize $\bar{\boldsymbol{x}}_0$, $\boldsymbol{G}_0^i \leftarrow \nabla F_i(\bar{\boldsymbol{x}}_0, \tilde{\xi}_i)$, $\boldsymbol{G}_0 \leftarrow \frac{1}{N}\sum_{i=1}^N \boldsymbol{G}_0^i$
for $r = 0, 1, \ldots, R-1$ **do**
  Sample $\mathcal{S}_r \subset [N]$ uniformly at random such that $|\mathcal{S}_r| = S$
  for $i \in \mathcal{S}_r$ **do**
    $\boldsymbol{x}_{r,0}^i \leftarrow \bar{\boldsymbol{x}}_r$
    for $k = 0, \ldots, I-1$ **do**
      Sample $\nabla F_i(\boldsymbol{x}_{r,k}^i; \xi_{r,k}^i)$, where $\xi_{r,k}^i \sim \mathcal{D}_i$
      $\boldsymbol{g}_{r,k}^i \leftarrow \nabla F_i(\boldsymbol{x}_{r,k}^i; \xi_{r,k}^i) - \boldsymbol{G}_r^i + \boldsymbol{G}_r$
      $\boldsymbol{x}_{r,k+1}^i \leftarrow \boldsymbol{x}_{r,k}^i - \eta \boldsymbol{g}_{r,k}^i \mathbb{1}\{\|\boldsymbol{G}_r\| \leq \gamma/\eta\} - \gamma \frac{\boldsymbol{g}_{r,k}^i}{\|\boldsymbol{g}_{r,k}^i\|} \mathbb{1}\{\|\boldsymbol{G}_r\| \geq \gamma/\eta\}$
    end for
    $\boldsymbol{G}_{r+1}^i \leftarrow \frac{1}{I}\sum_{k=0}^{I-1} \nabla F_i(\boldsymbol{x}_{r,k}^i; \xi_{r,k}^i)$
    $\Delta \boldsymbol{G}_r^i \leftarrow \boldsymbol{G}_{r+1}^i - \boldsymbol{G}_r^i$
  end for
  Update $\bar{\boldsymbol{x}}_{r+1} \leftarrow \frac{1}{S}\sum_{i \in \mathcal{S}_r} \boldsymbol{x}_{r,I}^i$
  Update $\boldsymbol{G}_{r+1} \leftarrow \boldsymbol{G}_r + \frac{1}{N}\sum_{i \in \mathcal{S}_r} \Delta \boldsymbol{G}_r^i$
  Denote $\boldsymbol{G}_{r+1}^i \leftarrow \boldsymbol{G}_r^i$ for all $i \notin \mathcal{S}_r$
end for

Two main features:

- Local update corrections.
- Episodic gradient clipping.

EPISODE++ Algorithm

Algorithm 1 EPISODE++

Initialize $\bar{\boldsymbol{x}}_0$, $\boldsymbol{G}_0^i \leftarrow \nabla F_i(\bar{\boldsymbol{x}}_0, \tilde{\xi}_i)$, $\boldsymbol{G}_0 \leftarrow \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{G}_0^i$
for $r = 0, 1, \ldots, R-1$ **do**
  Sample $\mathcal{S}_r \subset [N]$ uniformly at random such that $|\mathcal{S}_r| = S$
  for $i \in \mathcal{S}_r$ **do**
    $\boldsymbol{x}_{r,0}^i \leftarrow \bar{\boldsymbol{x}}_r$
    for $k = 0, \ldots, I-1$ **do**
      Sample $\nabla F_i(\boldsymbol{x}_{r,k}^i; \xi_{r,k}^i)$, where $\xi_{r,k}^i \sim \mathcal{D}_i$
      $\boldsymbol{g}_{r,k}^i \leftarrow \nabla F_i(\boldsymbol{x}_{r,k}^i; \xi_{r,k}^i) - \boldsymbol{G}_r^i + \boldsymbol{G}_r$
      $\boldsymbol{x}_{r,k+1}^i \leftarrow \boldsymbol{x}_{r,k}^i - \eta \boldsymbol{g}_{r,k}^i \mathbb{1}\{\|\boldsymbol{G}_r\| \leq \gamma/\eta\} - \gamma \frac{\boldsymbol{g}_{r,k}^i}{\|\boldsymbol{g}_{r,k}^i\|} \mathbb{1}\{\|\boldsymbol{G}_r\| \geq \gamma/\eta\}$
    end for
    $\boldsymbol{G}_{r+1}^i \leftarrow \frac{1}{I}\sum_{k=0}^{I-1} \nabla F_i(\boldsymbol{x}_{r,k}^i; \xi_{r,k}^i)$
    $\Delta \boldsymbol{G}_r^i \leftarrow \boldsymbol{G}_{r+1}^i - \boldsymbol{G}_r^i$
  end for
  Update $\bar{\boldsymbol{x}}_{r+1} \leftarrow \frac{1}{S}\sum_{i \in \mathcal{S}_r} \boldsymbol{x}_{r,I}^i$
  Update $\boldsymbol{G}_{r+1} \leftarrow \boldsymbol{G}_r + \frac{1}{N}\sum_{i \in \mathcal{S}_r} \Delta \boldsymbol{G}_r^i$
  Denote $\boldsymbol{G}_{r+1}^i \leftarrow \boldsymbol{G}_r^i$ for all $i \notin \mathcal{S}_r$
end for

Two main features:

- **Local update corrections**.
- Episodic gradient clipping.

EPISODE++ Algorithm

Algorithm 1 EPISODE++

Initialize $\bar{\boldsymbol{x}}_0$, $\boldsymbol{G}_0^i \leftarrow \nabla F_i(\bar{\boldsymbol{x}}_0, \tilde{\xi}_i)$, $\boldsymbol{G}_0 \leftarrow \frac{1}{N}\sum_{i=1}^N \boldsymbol{G}_0^i$
for $r = 0, 1, \ldots, R-1$ **do**
  Sample $\mathcal{S}_r \subset [N]$ uniformly at random such that $|\mathcal{S}_r| = S$
  for $i \in \mathcal{S}_r$ **do**
    $\boldsymbol{x}_{r,0}^i \leftarrow \bar{\boldsymbol{x}}_r$
    for $k = 0, \ldots, I-1$ **do**
      Sample $\nabla F_i(\boldsymbol{x}_{r,k}^i; \xi_{r,k}^i)$, where $\xi_{r,k}^i \sim \mathcal{D}_i$
      $\boldsymbol{g}_{r,k}^i \leftarrow \nabla F_i(\boldsymbol{x}_{r,k}^i; \xi_{r,k}^i) - \boldsymbol{G}_r^i + \boldsymbol{G}_r$
      $\boldsymbol{x}_{r,k+1}^i \leftarrow \boldsymbol{x}_{r,k}^i - \eta \boldsymbol{g}_{r,k}^i \mathbb{1}\{\|\boldsymbol{G}_r\| \leq \gamma/\eta\} - \gamma \frac{\boldsymbol{g}_{r,k}^i}{\|\boldsymbol{g}_{r,k}^i\|} \mathbb{1}\{\|\boldsymbol{G}_r\| \geq \gamma/\eta\}$
    end for
    $\boldsymbol{G}_{r+1}^i \leftarrow \frac{1}{I}\sum_{k=0}^{I-1} \nabla F_i(\boldsymbol{x}_{r,k}^i; \xi_{r,k}^i)$
    $\Delta \boldsymbol{G}_r^i \leftarrow \boldsymbol{G}_{r+1}^i - \boldsymbol{G}_r^i$
  end for
  Update $\bar{\boldsymbol{x}}_{r+1} \leftarrow \frac{1}{S}\sum_{i \in \mathcal{S}_r} \boldsymbol{x}_{r,I}^i$
  Update $\boldsymbol{G}_{r+1} \leftarrow \boldsymbol{G}_r + \frac{1}{N}\sum_{i \in \mathcal{S}_r} \Delta \boldsymbol{G}_r^i$
  Denote $\boldsymbol{G}_{r+1}^i \leftarrow \boldsymbol{G}_r^i$ for all $i \notin \mathcal{S}_r$
end for

Two main features:

- Local update corrections.
- **Episodic gradient clipping**.

Complexity Results

Method	Communication Complexity (R)	Best Iteration Complexity	Largest I to guarantee linear speedup	Setting
Local SGD [48]	$O\left(\frac{\Delta L\sigma^2}{NI\epsilon^4} + \frac{\Delta L\kappa^2 NI}{\sigma^2\epsilon^2} + \frac{\Delta LN}{\epsilon^2}\right)$	$O\left(\frac{L\sigma^2}{N\epsilon^4}\right)$	$O\left(\frac{\sigma^2}{\kappa N\epsilon}\right)$	(H)
SCAFFOLD [24]	$O\left(\frac{\Delta L\sigma^2}{SI\epsilon^4} + \frac{\Delta L}{\epsilon^2}\right)$	$O\left(\frac{\Delta L\sigma^2}{S\epsilon^4}\right)$	$O\left(\frac{\sigma^2}{N\epsilon^2}\right)$	(H), (S)
CELGC [32]	$O\left(\frac{\Delta L_0\sigma^2}{NI\epsilon^4}\right)$	$O\left(\frac{\Delta L_0\sigma^2}{N\epsilon^4}\right)$	$O\left(\frac{\sigma}{N\epsilon}\right)$	(Re)
EPISODE [9]	$O\left(\frac{\Delta L_0\sigma^2}{NI\epsilon^4} + \frac{\Delta(L_0+L_1(\kappa+\sigma))}{\epsilon^2}\left(1+\frac{\sigma}{\epsilon}\right)\right)$	$O\left(\frac{\Delta L_0\sigma^2}{N\epsilon^4}\right)$	$O\left(\frac{L_0\sigma^2}{(L_0+L_1(\kappa+\sigma)(1+\frac{\sigma}{\epsilon}))N\epsilon^2}\right)$	(Re), (H)
EPISODE++ (Theorem 1)†	$\tilde{O}\left(\frac{\Delta L_0\sigma^2}{SI\epsilon^4} + \frac{\Delta(L_0+L_1(\kappa+\rho\sigma))}{I\epsilon^3}\frac{L_0}{L_1\rho}\right)$	$\tilde{O}\left(\frac{\Delta L_0\sigma^2}{S\epsilon^4}\right)$	$\tilde{O}\left(\frac{L_0\sigma^2}{(L_0+L_1(\kappa+\rho\sigma))\left(\sigma+\frac{L_0}{L_1\rho}\right)S\epsilon}\right)$	(Re), (H), (S)

- EPISODE++ is the only algorithm with guarantees in our setting.
- Achieves linear speedup, reduced communication, and resilience to heterogeneity.
- Recovers iteration complexity of previous work for case of full participation.

Lower Bound for Baseline

- Minibatch SGD: Classical baseline for distributed optimization Cotter et al. [2011].
- Clipped Minibatch SGD: Extend to relaxed smooth setting by limiting length of each update (i.e. apply gradient clipping).

In the centralized setting, gradient clipping avoides exploding gradients Zhang et al. [2020a,b], i.e. the convergence rate does not depend on

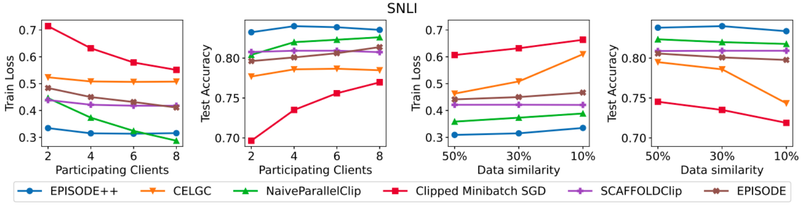
$$M := \sup \{ \|\nabla f(\boldsymbol{x})\| \mid f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0) \} .$$

Lower bound for communication steps for Clipped Minibatch SGD (Theorem 2):

$$R \geq \tilde{\Omega} \left(\frac{\Delta L_1 M}{\epsilon^2} \right)$$

The dependence on M shows that, in our setting, adding **gradient clipping** to Minibatch SGD **does not eliminate exploding gradients**.

Experimental Results



(a) Training loss and testing accuracy for SNLI dataset.

Bidirectional RNN for text classification on SNLI dataset Bowman et al. [2015].

EPISODE++ maintains superior performance as participation decreases, and as data heterogeneity increases.

Further experiments in the paper.

References

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *arXiv preprint arXiv:1710.11606*, 2017.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2020a.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020b.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Mingrui Liu, Zhenxun Zhuang, Yunwen Lei, and Chunyang Liao. A communication-efficient distributed gradient clipping algorithm for training deep neural networks. *Advances in Neural Information Processing Systems*, 35:26204–26217, 2022.

Michael Crawshaw, Yajie Bao, and Mingrui Liu. Episode: Episodic gradient clipping with periodic resampled corrections for federated learning with heterogeneous data. In *The Eleventh International Conference on Learning Representations*, 2022.

Andrew Cotter, Ohad Shamir, Nati Srebro, and Karthik Sridharan. Better mini-batch algorithms via accelerated gradient methods. *Advances in neural information processing*