A Unified Analysis of Federated Learning with Arbitrary Client Participation

Shiqiang Wang[1] and Mingyue Ji[2]

[1] IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
[2] Department of Electrical & Computer Engineering, University of Utah, Salt Lake City, UT, USA

Federated Learning

- Collaborative model training while not sharing raw data
  - Local objective at client *n*:

$$F_n(\mathbf{x}) := \mathbb{E}_{\xi_n \sim \mathcal{D}_n}\left[\ell_n(\mathbf{x}, \xi_n)\right]$$

Loss function of model with parameter $\mathbf{x}$ for data sample ξ_n: $\ell_n(\mathbf{x}, \xi_n)$

  - Global objective (not directly observable):

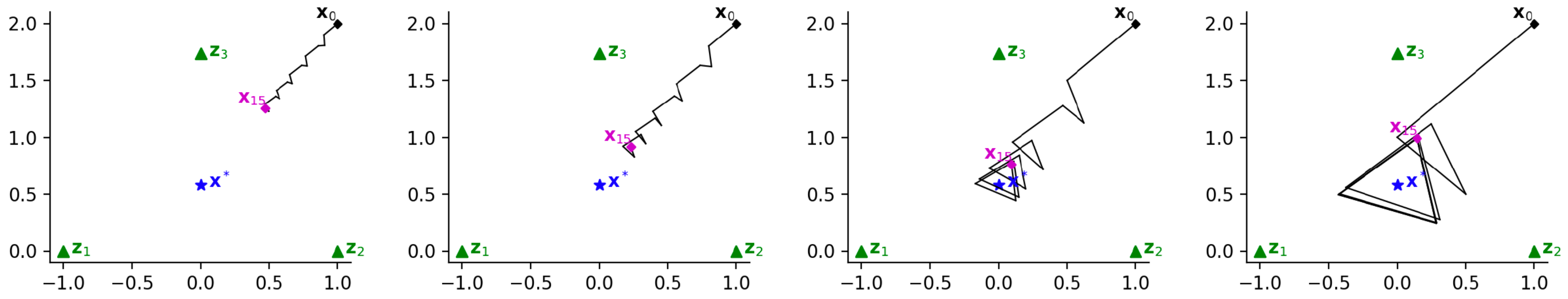
$$f(\mathbf{x}) := \frac{1}{N}\sum_{n=1}^{N} F_n(\mathbf{x})$$

Find $\mathbf{x}^*$ to minimize $f(\mathbf{x})$

- *Federated averaging (FedAvg):* local SGD at clients + parameter aggregation via the server
- **Our focus**
  - Clients may be only intermittently available to participate in training
    - For example: mobile devices during charging, edge servers when idle
  - Questions
    - How to effectively train models when clients participate arbitrarily?
    - How do unavailable clients affect the performance of model training?

Problem with Intermittent Participation

- Motivating example with $F_n(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{x} - \mathbf{z}_n\right\|^2$
- Three clients participating cyclically ($P = 3$), one in each round



(a) $\gamma = 0.05, \eta = 1$ (b) $\gamma = 0.1, \eta = 1$ (c) $\gamma = 0.25, \eta = 1$ (d) $\gamma = 0.5, \eta = 1$

γ: local learning rate

- Observation
  - Moves slowly to $\mathbf{x}^*$ when γ is small
  - Circles around $\mathbf{x}^*$ when γ is large

Apparent gap between $\mathbf{x}_{15}$ and $\mathbf{x}^*$

Generalized FedAvg

Input: $\gamma, \eta, \mathbf{x}_0, I, P, T$; **Output:** $\{\mathbf{x}_t : \forall t\}$
Initialize $t_0 \leftarrow 0$, $\mathbf{u} \leftarrow \mathbf{0}$;
for $t \leftarrow 0, \ldots, T-1$ **do**
  for $n \leftarrow 1, \ldots, N$ *in parallel* **do**
    $\mathbf{y}^n_{t,0} \leftarrow \mathbf{x}_t$;
    for $i \leftarrow 0, \ldots, I-1$ **do**
      $\mathbf{y}^n_{t,i+1} \leftarrow \mathbf{y}^n_{t,i} - \gamma \mathbf{g}_n(\mathbf{y}^n_{t,i})$
    $\Delta^n_t \leftarrow \mathbf{y}^n_{t,I} - \mathbf{x}_t$;
  Participation weight
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \sum_{n=1}^{N} \boxed{q^n_t} \Delta^n_t$; //update
  $\mathbf{u} \leftarrow \mathbf{u} + \sum_{n=1}^{N} q^n_t \Delta^n_t$; //accumulate
  if $t + 1 - t_0 = \boxed{P}$ **then** Amplification interval
    $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_{t+1} + (\boxed{\eta} - 1)\mathbf{u}$; //amplify
    $t_0 \leftarrow t + 1$; Amplification factor
    $\mathbf{u} \leftarrow \mathbf{0}$;

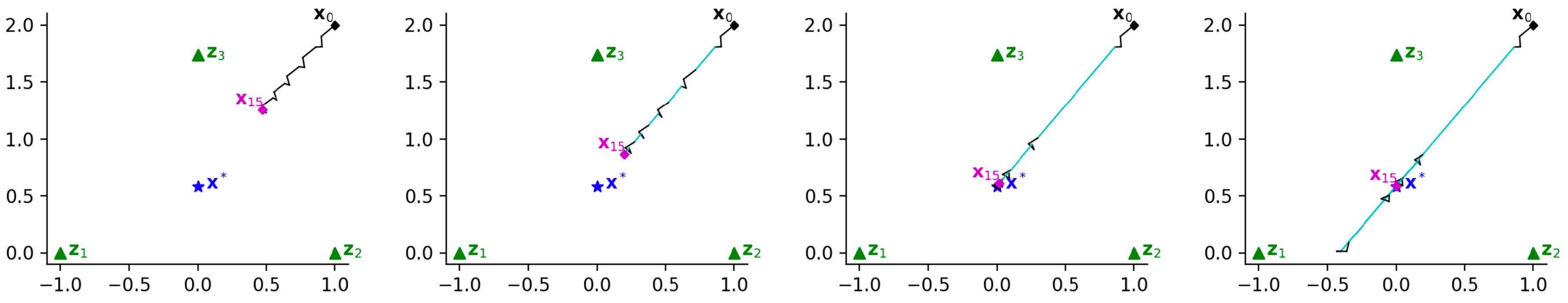
Same as standard FedAvg

Accumulate and amplify updates every *P* rounds
(no additional communication, minimal additional computation)

Amplification Helps!

- Motivating example with $F_n(\mathbf{x}) = \frac{1}{2}\left\|\mathbf{x} - \mathbf{z}_n\right\|^2$
- Three clients participating cyclically ($P = 3$), one in each round

Change in $\mathbf{x}$ due to amplification shown in cyan color



(a) $\gamma = 0.05, \eta = 1$ (b) $\gamma = 0.05, \eta = 2$ (c) $\gamma = 0.05, \eta = 5$ (d) $\gamma = 0.05, \eta = 10$

- Observation
  - By choosing a smaller γ and a larger η, we can get very close to $\mathbf{x}^*$ within only a few rounds

γ: local learning rate

η: amplification factor

Main Building Block of Unified Analysis

Assumption 1 (Lipschitz gradient).

$$\|\nabla F_n(\mathbf{x}) - \nabla F_n(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y}, n.$$

Assumption 2 (Unbiased stochastic gradient with bounded variance).

$$\mathbb{E}\left[\mathbf{g}_n(\mathbf{x})\right] = \nabla F_n(\mathbf{x}) \text{ and } \mathbb{E}\left[\|\mathbf{g}_n(\mathbf{x}) - \nabla F_n(\mathbf{x})\|^2\right] \leq \sigma^2, \forall \mathbf{x}, n.$$

Assumption 3 (Bounded gradient divergence).

$$\|\nabla F_n(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq d^2, \forall \mathbf{x}, n.$$

Decomposition of gradient divergence:

$$\left\|\sum_{n=1}^N q_t^n \left[\nabla F_n(\mathbf{x}) - \nabla f(\mathbf{x})\right]\right\|^2 \leq \tilde{\beta}^2, \forall \mathbf{x}, t,$$

$$\sum_{n=1}^N q_t^n \left\|\nabla F_n(\mathbf{x}) - \sum_{n'=1}^N q_t^{n'} \nabla F_{n'}(\mathbf{x})\right\|^2 \leq \tilde{\nu}^2, \forall \mathbf{x}, t,$$

$$\left\|\frac{1}{P}\sum_{t=t_0}^{t_0+P-1} \sum_{n=1}^N q_t^n \left(\nabla F_n(\mathbf{x}) - \nabla f(\mathbf{x})\right)\right\|^2 \leq \boxed{\tilde{\delta}^2(P)}, \forall \mathbf{x}, t_0.$$

Effect of partial participation

Choose depending on whether *P* scales in *T*

Corollary 3.2. *Choosing* $\gamma = \frac{1}{12LIP\sqrt{T}}$ *and* $\eta = \min\left\{\frac{12P\sqrt{LI\mathcal{F}}}{\sigma\rho}; 12\sqrt{T}\right\}$, *for* $P \leq \frac{T}{2}$, *we have*

$$\min_t \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2 \middle| \mathcal{Q}\right] \leq \mathcal{O}\left(\frac{\sigma\rho\sqrt{L\mathcal{F}}}{\sqrt{IT}} + \frac{LP\mathcal{F}}{T} + \frac{\tilde{\nu}^2}{P^2T} + \frac{\tilde{\beta}^2}{T} + \boxed{\tilde{\delta}^2(P)} + \frac{\sigma^2}{IPT}\right).$$

Corollary 3.3. *If* $\frac{\sqrt{\mathcal{F}}}{\rho\sqrt{LIT}} \leq \frac{1}{LIP}$, *choosing* $\gamma = \frac{1}{12LIP\sqrt{T}}$ *and* $\eta = \frac{12P\sqrt{LI\mathcal{F}}}{\rho}$, *for* $P \leq \frac{T}{2}$, *we have*

$$\min_t \mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2 \middle| \mathcal{Q}\right] \leq \mathcal{O}\left(\frac{(1+\sigma^2)\rho\sqrt{L\mathcal{F}}}{\sqrt{IT}} + \frac{\tilde{\nu}^2}{P^2T} + \frac{\tilde{\beta}^2}{T} + \boxed{\tilde{\delta}^2(P)} + \frac{\sigma^2}{IPT}\right).$$

$$\sum_{n=1}^N q_t^n = 1$$

$$\sum_{n=1}^N (q_t^n)^2 \leq \rho^2$$

Results for Different Participation Patterns

- Assuming S clients participate in each round with equal weight, we have

$$q_t^n = {}^1\!/_S \qquad \rho = \left[\sum_{n=1}^{N} (q_t^n)^2\right]^{1/2} = {}^1\!/\sqrt{S}$$

"Linear speedup"

Participation	Convergence error upper bound	Remark
Regularized	$\mathcal{O}\left(\frac{\sigma}{\sqrt{\boxed{S}IT}}\right)$	Matching centralized SGD lower bound
Ergodic	Approaches zero as $T \to \infty$	
Mixing	$\mathcal{O}\left(\frac{1+\sigma^2}{\sqrt{\boxed{S}IT}}\right)$ i.E.; $\mathcal{O}\left(\frac{1+\sigma^2}{c\sqrt{\boxed{S}IT}}\right)$ w.p. $1-c$	Matching known bound with idealized participation
Independent	$\mathcal{O}\left(\frac{1+\sigma^2}{\sqrt{\boxed{S}IT}}\right)$ i.E.; $\mathcal{O}\left(\frac{(1+\sigma^2)\ln(2/c)}{\sqrt{\boxed{S}IT}}\right)$ w.p. $1-c$	Matching known bound with idealized participation

From Corollary 3.2

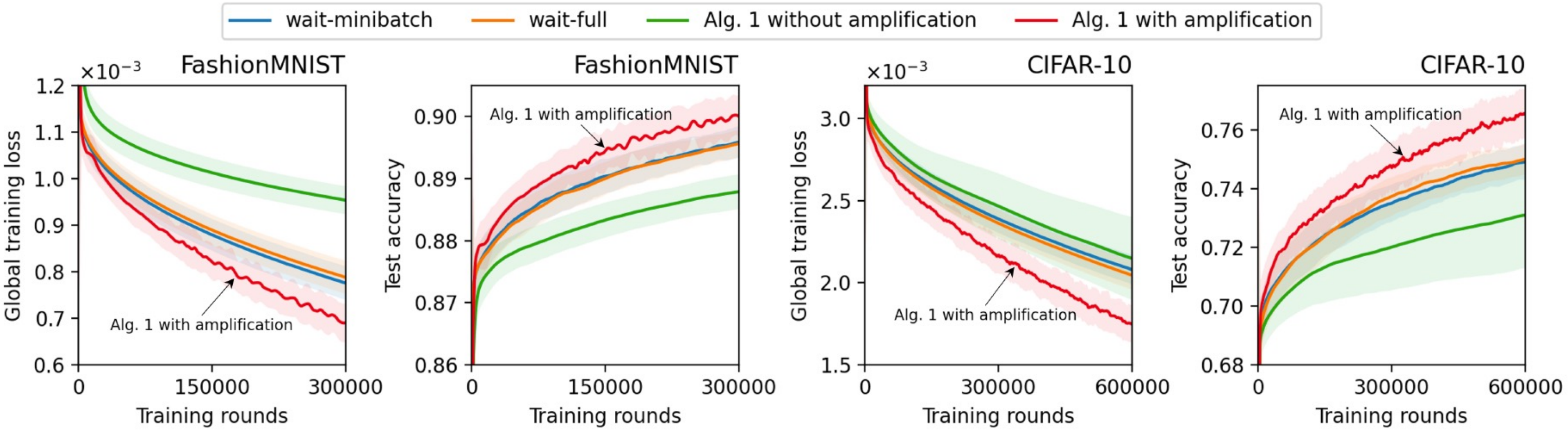
From Corollary 3.3

The dominant term does not depend on P

i.E. = in expectation
w.p. = with probability

Experiments

- Cyclic participation of clients with heterogeneous data, where each full cycle includes 500 rounds
- Optimized learning rates from grid search for each method
- $\eta = 10$ and $P = 500$ for the generalized FedAvg algorithm with amplification



Recap

- Generalized FedAvg with amplification
- A unified framework for convergence analysis with arbitrary participation
- Theoretical convergence bounds for different participation patterns
- Experiment confirming improvement compared to baselines

Thank You!

Email: wangshiq@us.ibm.com
Homepage: https://shiqiang.wang/