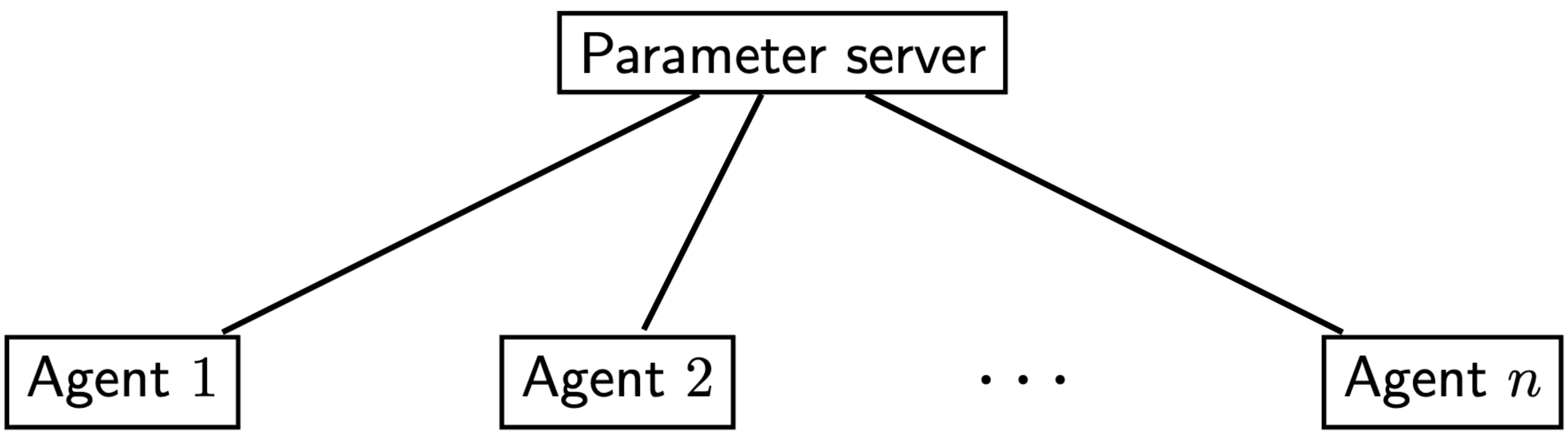
On the Convergence of FedAvg on Non-iid Data

Xiang Li*, Kaixuan Huang*, Wenhao Yang*, Shusen Wang and Zhihua Zhang
Accepted as an oral presentation in 2020 ICLR

Xiang Li, Upenn, 2024-08-21

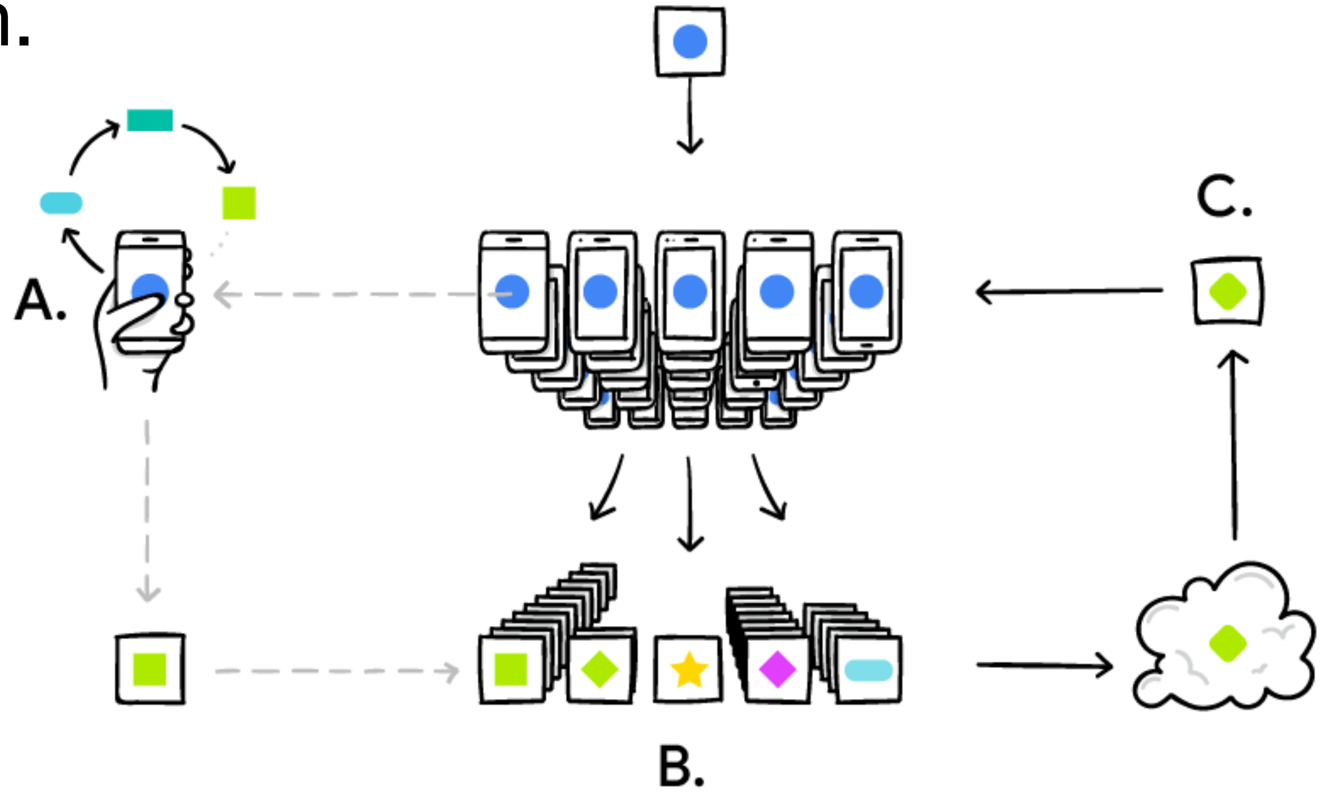
Distributed Learning

- Centralize data and then fit models.
- Full control of data.



Federated Learning

- Fit model collaboratively **without** data sharing.
- Many benefits:
  - (1) use massive available computing resources;
  - (2) protect users' privacy;
  - (3) alleviate storage burden.



Federated Learning

Some characters

1. FL has training data is **massively distributed.**

   Communication efficiency.

2. **Unable to control** over users' devices.

   Partial device participation.

3. The training data are **non-iid.**

   Data heterogeneity.

4. The training data are **unbalanced.**

   Fairness.

Problem Setup

- Training a global shared model:

$$\min_{\mathbf{w}} F(\mathbf{w}) \triangleq \sum_{k=1}^{N} p_k F_k(\mathbf{w})$$

- For the k-th device:

$$F_k(\mathbf{w}) \triangleq \frac{1}{n_k} \sum_{j=1}^{n_k} \ell(\mathbf{w}; x_{k,j})$$

- Local data: n_k points with $x_{k,1}, x_{k,2}, \cdots, x_{k,n_k} \sim \mathscr{D}_k$.
- Note that (i) N could be very large; (ii) $\mathscr{D}_i \neq \mathscr{D}_j$ with $i \neq j$ due to data heterogeneity; (iii) $p_k = \dfrac{n_k}{n}$.

Some algorithms

- Fully Synchronous SGD:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \left[\frac{1}{N} \sum_{i=1}^{N} g\left(\mathbf{w}_t; \xi_t^{(i)}\right) \right]$$

- Local SGD:

$$\mathbf{w}_{t+1}^{(i)} = \begin{cases} \frac{1}{N} \sum_{j=1}^{N} \left[\mathbf{w}_t^{(j)} - \eta g\left(\mathbf{w}_t^{(j)}\right)\right], & t \bmod E = 0 \\ \mathbf{w}_t^{(i)} - \eta g\left(\mathbf{w}_t^{(i)}\right), & \text{otherwise} \end{cases}$$

- FedAvg: A random portion of devices perform updates.

FedAvg

Algorithm 1 `FederatedAveraging`. The K clients are indexed by k; B is the local minibatch size, E is the number of local epochs, and η is the learning rate.

Server executes:
  initialize w_0
  for each round $t = 1, 2, \dots$ **do**
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of m clients)
    for each client $k \in S_t$ **in parallel do**
      $w^k_{t+1} \leftarrow \text{ClientUpdate}(k, w_t)$
    $w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w^k_{t+1}$

ClientUpdate(k, w)**:** *// Run on client* k
  $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of size B)
  for each local epoch i from 1 to E **do**
    for batch $b \in \mathcal{B}$ **do**
      $w \leftarrow w - \eta \nabla \ell(w; b)$
  return w to server

McMahan B, Moore E, Ramage D, et al. Communication-efficient learning of deep networks from decentralized data[C]//Artificial intelligence and statistics. PMLR, 2017: 1273-1282.

Previous work

- If data are **iid** and **all** devices are active, FedAvg = LocalSGD, while the latter has been analyzed by many work [Coppola (2015); Zhou and Cong (2017); Stich (2018); Lin et al (2018); Wang and Joshi (2018); Yu et al. (2019); Khaled et al. (2019)].
- FedProx [Sahu (2018)] doesn't require the two assumptions. It incorporates FedAvg as a special cases. But their theory couldn't to cover FedAvg.
- We focus the theoretical understanding on FedAvg under more realistic (i.e., non-iid and partial participation) settings.

Assumptions

1. All F_k's are L-smooth: $\|\nabla F_k(\mathbf{w}) - \nabla F_k(\mathbf{v})\|_2 \leq L\|\mathbf{v} - \mathbf{w}\|_2$.
2. All F_k's are μ-smooth: $\nabla^2 F_k(\mathbf{w}) \geq \mu I$.
3. Bounded variance: $\mathbb{E}\|\nabla F_k(\mathbf{w}_t^k, \xi_t^k) - \nabla F_k(\mathbf{w}_t^k)\|^2 \leq \sigma_k^2$.
4. Bounded moments: $\mathbb{E}\|\nabla F_k(\mathbf{w}_t^k, \xi_t^k)\|^2 \leq G^2$.

Non-iid quantification: $\Gamma = F^* - \sum_{k=1}^{N} p_k F_k^*$.

Convergence result
Full participation

- Under previous Assumption 1-4, we show that

$$\mathbb{E}[F(\mathbf{w}_T)] - F^* \leq \frac{2\kappa}{\gamma + T}\left(\frac{B}{\mu} + 2L \parallel \mathbf{w}_0 - \mathbf{w}^* \parallel^2\right)$$

where $\kappa = L/\mu$, $\gamma = \max\{8\kappa, E\}$, and

$$B = \sum_{k=1}^{N} p_k^2 \sigma_k^2 + 6L\Gamma + 8(E-1)^2 G^2 \,.$$

Convergence result

Partial participation

- Two models:

  sample K devices randomly and average $\mathbf{w}_t \longleftarrow \frac{1}{K}\sum_{k\in\mathcal{S}_t} \mathbf{w}_t^k$.

  Scheme I: with replacement according to probabilities $p_1, \cdots, p_N$.
  Scheme II: without replacement $(p_1 = \cdots = p_N = 1/N)$.

- Under previous Assumption 1-4, we show that

$$\mathbb{E}[F(\mathbf{w}_T)] - F^* \leq \frac{2\kappa}{\gamma + T}\left(\frac{B+C}{\mu} + 2L \parallel \mathbf{w}_0 - \mathbf{w}^* \parallel^2\right)$$

  For Scheme I, $C = 4E^2G^2/K$. For Scheme II, $C = \frac{N-K}{N-1}\frac{4}{K}E^2G^2$.

Convergence result

Discussion

- Data heterogeneity slows down the convergence via Γ.
- If $\mathcal{S}_t \equiv [N]$, $C = 0$ for Scheme II not for Scheme I.
- How to choose E ? $$\frac{T_\epsilon}{E} \propto \left(1 + \frac{1}{K}\right) EG^2 + \frac{\sum_k p_k^2 \sigma_k^2 + L\Gamma + \kappa G^2}{E}$$
- How to choose K ? A milder dependence.
- How to choose sampling schemes?

Learning rate decay

- We artificially construct a strongly convex and smooth distributed optimization problem. With full batch size (no randomness), $E > 1$, and any fixed and constant step size η, FedAvg will converge to sub-optimal points:

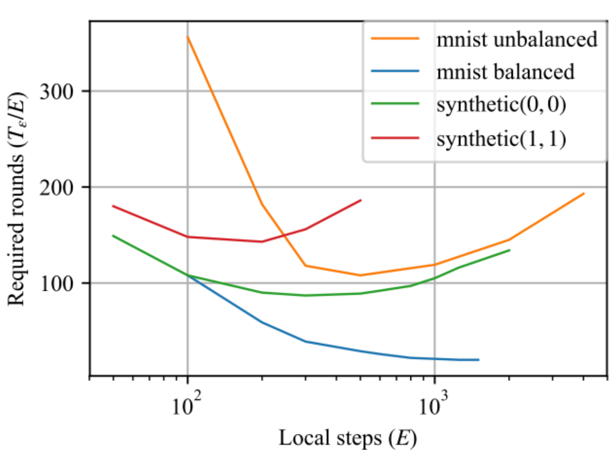
$$\| \tilde{\mathbf{w}}^* - \mathbf{w}^* \|_2 = \Omega((E-1)\eta) \cdot \| \mathbf{w}^* \|_2 .$$

- If $E = 1$, constant step sizes ensure exact convergence.
- $E > 1$ possible bias in FedAvg.
- The result motivates more efficient alternatives to FedAvg.

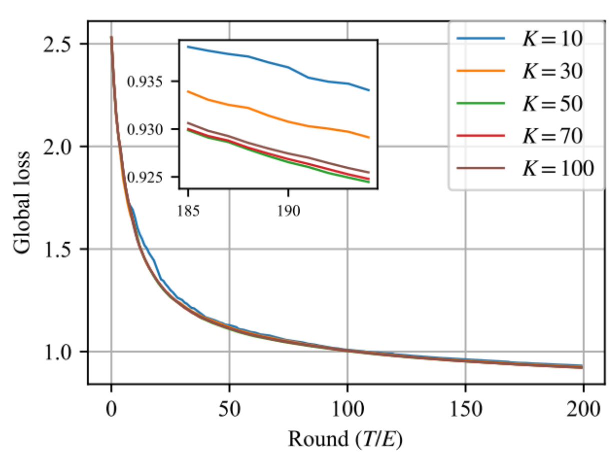
Numerical experiments

- $N = 100$, ridge penalty with $\lambda = 10^{-4}$.
- Transformed Scheme II: Scale the loss so that

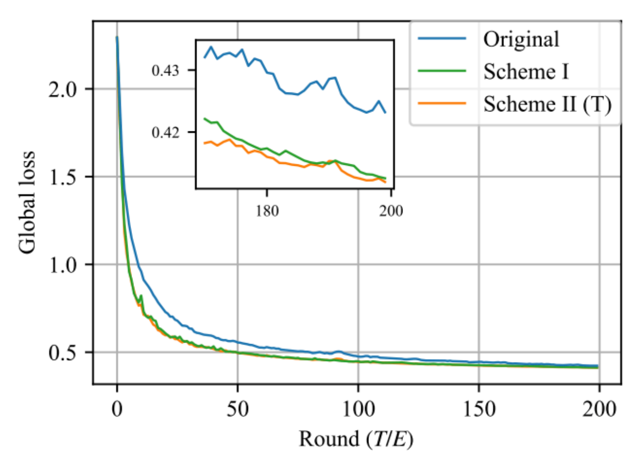
$$L = N \cdot \max_k p_k, \mu = N \cdot \min_k p_k .$$



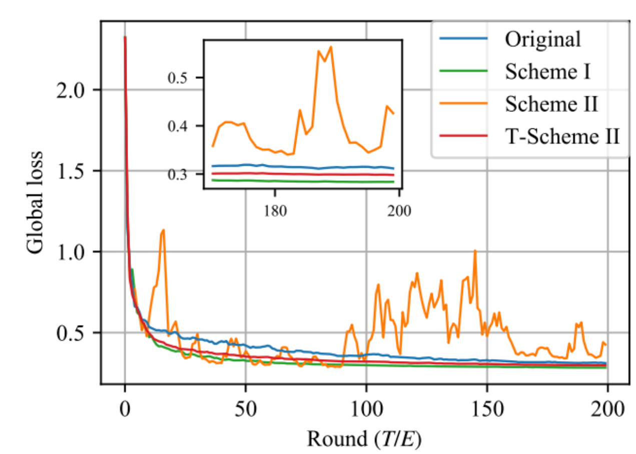
(a) The impact of E



(b) The impact of K



(c) Different schemes



(d) Different schemes

Take-away

- FedAvg converges when data are non-iid. (Assume convexity, smoothness, etc.)
- Convergence rate is affected by the degree of non-iid and partial participation.
- The decay of learning rate is necessary.

Other topics…

- How to overcome data heterogeneity.
- How to choose an optimal sampling scheme.
- Decentralized topology.
- Other federated tasks, e.g., minimax, nonconvex, transfer learning….
- Personalization.
- Other consideration (privacy, fairness, insensitive, ……)