On The Convergence of FedAvg on Non-iid Data

Xiang Li*, Kaixuan Huang*, Wenhao Yang*
Shusen Wang and Zhihua Zhang

Federated Learning

Federated Learning

- Standard Distributed Learning = **centralize** data and then fit models

Federated Learning

- Standard Distributed Learning = **centralize** data and then fit models
- Federated Learning (FL) = fit model collaboratively **without** data sharing

Federated Learning

- Standard Distributed Learning = **centralize** data and then fit models
- Federated Learning (FL) = fit model collaboratively **without** data sharing
- FL has three unique characters:

Federated Learning

- Standard Distributed Learning = **centralize** data and then fit models
- Federated Learning (FL) = fit model collaboratively **without** data sharing
- FL has three unique characters:
    - training data is **massively distributed;**

Federated Learning

- Standard Distributed Learning = **centralize** data and then fit models
- Federated Learning (FL) = fit model collaboratively **without** data sharing
- FL has three unique characters:
  - training data is **massively distributed;**

Communication efficiency.

Federated Learning

- Standard Distributed Learning = **centralize** data and then fit models
- Federated Learning (FL) = fit model collaboratively **without** data sharing
- FL has three unique characters:
  - training data is **massively distributed;**
  - **unable to control** over users' devices;

Communication efficiency.

Federated Learning

- Standard Distributed Learning = **centralize** data and then fit models
- Federated Learning (FL) = fit model collaboratively **without** data sharing
- FL has three unique characters:
  - training data is **massively distributed;** — **Communication efficiency.**
  - **unable to control** over users' devices; — **Partial device participation.**

Federated Learning

- Standard Distributed Learning = **centralize** data and then fit models
- Federated Learning (FL) = fit model collaboratively **without** data sharing
- FL has three unique characters:
  - training data is **massively distributed;** — Communication efficiency.
  - **unable to control** over users' devices; — Partial device participation.
  - the training data are **non-iid.**

Federated Learning

- Standard Distributed Learning = **centralize** data and then fit models
- Federated Learning (FL) = fit model collaboratively **without** data sharing
- FL has three unique characters:
  - training data is **massively distributed;** — **Communication efficiency.**
  - **unable to control** over users' devices; — **Partial device participation.**
  - the training data are **non-iid.** — **Heterogeneity.**

Problem Setup

Problem Setup

- Consider the distributed optimization: $\min_{w} F(w) \triangleq \sum_{k=1}^{N} p_k F_k(w)$ where N is # of devices and p_k is the weight of the k-th device.

Problem Setup

- Consider the distributed optimization: $\min_{w} F(w) \triangleq \sum_{k=1}^{N} p_k F_k(w)$ where N is # of devices and p_k is the weight of the k-th device.
- The k-th device holds n_k training data: $x_{k,1}, x_{k,2}, \cdots, x_{k,n_k} \sim \mathcal{D}_k$.

Problem Setup

- Consider the distributed optimization: $\min_{w} F(w) \triangleq \sum_{k=1}^{N} p_k F_k(w)$ where N is # of devices and p_k is the weight of the k-th device.
- The k-th device holds n_k training data: $x_{k,1}, x_{k,2}, \cdots, x_{k,n_k} \sim \mathscr{D}_k$.
- The local objective is defined by $F_k(w) \triangleq \frac{1}{n_k} \sum_{j=1}^{n_k} \ell(w; x_{k,j})$ where $\ell(\cdot\,;\cdot)$ is a loss function.

Problem Setup

- Consider the distributed optimization: $\min_{w} F(w) \triangleq \sum_{k=1}^{N} p_k F_k(w)$ where N is # of devices and p_k is the weight of the k-th device.
- The k-th device holds n_k training data: $x_{k,1}, x_{k,2}, \cdots, x_{k,n_k} \sim \mathscr{D}_k$.
- The local objective is defined by $F_k(w) \triangleq \frac{1}{n_k} \sum_{j=1}^{n_k} \ell(w; x_{k,j})$ where $\ell(\cdot\,;\cdot)$ is a loss function.
- Note that (i) N could be very large; (ii) $\mathscr{D}_i \neq \mathscr{D}_j$ with $i \neq j$ due to heterogeneity; (iii) $p_k = \frac{n_k}{n}$.

Problem Setup

Problem Setup

- Consider the distributed optimization: $\min_w F(w) \triangleq \sum_{k=1}^{N} p_k F_k(w)$ where N is # of devices and p_k is the weight of the k-th device.
- The k-th device holds n_k training data: $x_{k,1}, x_{k,2}, \cdots, x_{k,n_k} \sim \mathcal{D}_k$.
- The local objective is defined by $F_k(w) \triangleq \frac{1}{n_k} \sum_{j=1}^{n_k} \ell(w; x_{k,j})$ where $\ell(\cdot\,;\cdot)$ is a loss function.
- Note that (i) N could be very large; (ii) $\mathcal{D}_i \neq \mathcal{D}_j$ with $i \neq j$ due to heterogeneity; (iii) $p_k = \frac{n_k}{n}$.

FedAvg

FedAvg

- First, the central server **activates** a random small set (say $\mathcal{S}_t$) of devices and then **broadcasts** the latest model w_t to the **activated** devices;

FedAvg

- First, the central server **activates** a random small set (say $\mathcal{S}_t$) of devices and then **broadcasts** the latest model w_t to the **activated** devices;
- Second, every activated device (say the k-th and $k \in \mathcal{S}_t$) performs $E(\geq 1)$ **local updates**: $w_{t+i+1}^k \longleftarrow w_{t+i}^k - \eta_{t+i} \nabla F_k(w_{t+i}^k, \xi_{t+i}^k), i = 0, 1, \cdots, E-1$ where η_{t+i} is the learning rate and ξ_{t+i}^k is a sample uniformly chosen from the k-th local dataset.

FedAvg

- First, the central server **activates** a random small set (say $\mathcal{S}_t$) of devices and then **broadcasts** the latest model w_t to the **activated** devices;

- Second, every activated device (say the k-th and $k \in \mathcal{S}_t$) performs $E(\geq 1)$ **local updates**:$w_{t+i+1}^k \longleftarrow w_{t+i}^k - \eta_{t+i} \nabla F_k(w_{t+i}^k, \xi_{t+i}^k), i = 0,1,\cdots,E-1$ where η_{t+i} is the learning rate and ξ_{t+i}^k is a sample uniformly chosen from the k-th local dataset.

- Last, the server **aggregates** the local models, $\{w_{t+E}^k\}_{k\in\mathcal{S}_t}$ to produce the new global model, $w_{t+E} \longleftarrow$ Aggregate$(\{w_{t+E}^k\}_{k\in\mathcal{S}_t})$.

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 settings.

Convergence Result

Convergence Result

- Under some regularity conditions and decaying the learning rate, we have $\mathbb{E}\left[F(w_T) - F^*\right] \ \leq \ \mathcal{O}\left((B+C)/T\right)$, where $B = \Gamma + (E-1)^2$.

Convergence Result

- Under some regularity conditions and decaying the learning rate, we have $\mathbb{E}\left[F(w_T) - F^*\right] \ \le \ \mathcal{O}\left((B + C)/T\right)$, where $B = \Gamma + (E-1)^2$.

- The non-iid is measured by $\Gamma = F^* - \sum_{k=1}^{N} p_k F_k^*$.

Convergence Result

- Under some regularity conditions and decaying the learning rate, we have $\mathbb{E}\left[F(w_T) - F^*\right] \leq \mathcal{O}\left((B+C)/T\right)$, where $B = \Gamma + (E-1)^2$.

- The non-iid is measured by $\Gamma = F^* - \sum_{k=1}^{N} p_k F_k^*$.

- C is a term related with the way $\mathcal{S}_t$ is formed. If $\mathcal{S}_t = [N]$, $C = 0$.

Convergence Result

- Under some regularity conditions and decaying the learning rate, we have $\mathbb{E}\left[F(w_T) - F^*\right] \;\leq\; \mathcal{O}\left((B + C)/T\right)$, where $B = \Gamma + (E-1)^2$.

- The non-iid is measured by $\Gamma = F^* - \sum_{k=1}^{N} p_k F_k^*$.

- C is a term related with the way $\mathcal{S}_t$ is formed. If $\mathcal{S}_t = [N]$, $C = 0$.

- The number of required communication rounds is roughly $\left(1 + \frac{1}{K}\right) E + \frac{\Gamma}{E}$.

Learning Rate Decay

Learning Rate Decay

- If the learning rate doesn't decay, then $\tilde{w}^*$ (produced by FedAvg) is away from the optimal w^* (the optimal point): $\|\tilde{w}^* - w^*\|_2 = \Omega((E-1)\eta) \cdot \|w^*\|_2$.

Learning Rate Decay

- If the learning rate doesn't decay, then $\tilde{w}^*$ (produced by FedAvg) is away from the optimal w^* (the optimal point): $\|\tilde{w}^* - w^*\|_2 = \Omega((E-1)\eta) \cdot \|w^*\|_2$.

- The gradients are non-random and $\mathcal{S}_t = [N]$.

Learning Rate Decay

- If the learning rate doesn't decay, then $\tilde{w}^*$ (produced by FedAvg) is away from the optimal w^* (the optimal point): $\|\tilde{w}^* - w^*\|_2 = \Omega((E-1)\eta) \cdot \|w^*\|_2$.
- The gradients are non-random and $\mathcal{S}_t = [N]$.
- Diminishing learning rates is crucial.

Learning Rate Decay

- If the learning rate doesn't decay, then $\tilde{w}^*$ (produced by FedAvg) is away from the optimal w^* (the optimal point): $\|\tilde{w}^* - w^*\|_2 = \Omega((E-1)\eta) \cdot \|w^*\|_2$.
- The gradients are non-random and $\mathcal{S}_t = [N]$.
- Diminishing learning rates is crucial.
- Motivate alternatives.

Take-away

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

Thank You