ECE 8101: Nonconvex Optimization for Machine Learning

Lecture Note 3-1: Federated Learning (feat. Distributed Learning)

Jia (Kevin) Liu

Assistant Professor
Department of Electrical and Computer Engineering
The Ohio State University, Columbus, OH, USA

Autumn 2024

Outline

In this lecture:

- Key Idea of Distributed Optimization for Federated Learning
- Representative Algorithms
- Convergence Results

Revisit the General Expectation Minimization Problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) = \min_{\mathbf{x}\in\mathbb{R}^d} \mathbb{E}_{\xi\sim\mathcal{D}}[f(\mathbf{x},\xi)]$$

- The SGD method using mini-batch $\mathcal{B}_k$ with $|\mathcal{B}_k| = B_k$ is:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \frac{s_k}{B_k}\sum_{i=1}^{B_k} \nabla f(\mathbf{x}_k, \xi_i)$$

- Key Insight: The "summation" in the mini-batched version of SGD implies a decomposable structure that lends itself to distributed implementation!
  - Each stochastic gradient $\nabla f(\mathbf{x}_k, \xi_i)$ can be computed by a "worker" i
  - B_k workers can compute such stochastic gradients in parallel
  - A server collects the stochastic gradients returned by workers and aggregate

This insight is the foundation of Distributed Learning and Federated Learning

Distributed Learning in Data Center Setting

- Distributed ML Systems

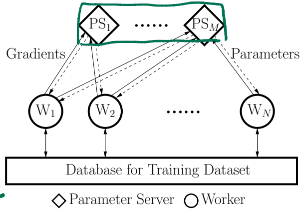
Keras TensorFlow Caffe PyTorch

- Time consuming
- Resource intensive

Model	ImageClassfication	DeepSpeech2
Dataset	ResNet50	LibriSpeech
System	8 GPUs	16 GPUs
Time	115 minutes[1]	3-5 days[2]

$$\sum_{i=1}^{4} g_i \equiv \begin{bmatrix} \sum_{i=1}^{4} g_i^{(1)} \\ \vdots \\ \sum_{i=1}^{4} g_i^{(4)} \end{bmatrix}$$

- Parameter Server-Worker (SW) Architecture

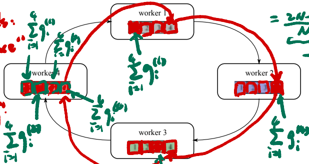


$$g_i = \begin{bmatrix} g_i^{(1)} \\ g_i^{(2)} \\ g_i^{(3)} \\ g_i^{(4)} \end{bmatrix}$$

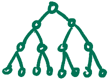
- Ring-All-Reduce (RAR) Architecture

1st N-1 rounds: "share-Reduce"

2nd N-1 rounds "share-only"

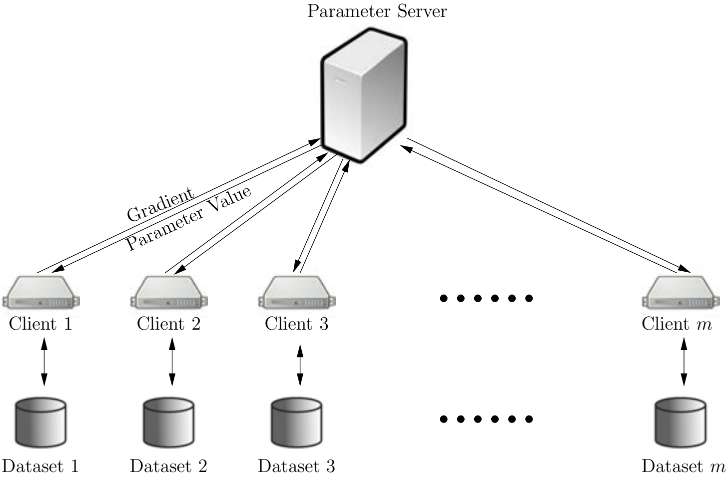


$$2(N-1)\frac{d}{N} = \frac{2N-2}{N}\cdot d \rightarrow 2 \text{ as } N\to\infty$$



[1] Mlperf training results, https://mlperf.org/training-results-0-6/
[2] E. B. Dario Amodei, Rishita Anubhai, C. Case, J. Casper, B. Catanzaro, J. Chen et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in Proc. of the 33th International Conference on Machine Learning (ICML), 2016.

Federated Learning System Architecture



Federated Learning (FL)

- The term "federated learning" was first coined in 2016 (arXiv):
  - *"We term our approach Federated Learning, since the learning task is solved by a loose federation of participating devices (which we refer to as clients) which are coordinated by a central server."* [McMahan et al. AISTATS'17]
- Key motivations of FL:
  - FL was first focused on mobile & edge devices collaborating to train a global model and later became a general learning paradigm
  - No need to transfer clients' data to the server to preserve privacy
- A very active ongoing research field with the following defining challenges:
  - Dataset sizes are unbalanced across clients in general
  - Datasets are non-i.i.d. across clients in general
  - Could involve a massive number of client devices
  - Limited communication bandwidth between server and clients
  - Limited device availability (e.g., powered-off, charging, no wifi...)
- Two widely studied FL settings:
  - Cross-device: Huge number of (unreliable) clients (e.g., mobile devices)
  - Cross-silo: Small number of (relatively) reliable clients (hospitals, banks, etc.)

Cross-Device Federated Learning

According to [Kairouz et al. arXiv-1912.04977]:

- Total population: 10^6–10^{10} devices
- Device selected per-round: 50–5000
- Total devices participated in training a model: 10^5–10^7
- Number of rounds for convergence: 500–10000
- Wall-clock training time: 1–10 days
- Data partition: By samples

Cross-Silo Federated Learning

- The number of clients is relatively small. Often reasonable to assume that clients are available at all times
- Relevant when a number of companies or organizations share incentive to training a model based on their data, but cannot share data directly
- Data partition: Could be either by samples or by features
  - Also referred to as "horizontal" and "vertical" FL in the literature, respectively
  - By examples: Relevant in cross-silo FL when a single organization cannot centralize their data
  - By features: Relevant in cross-silo FL if data security/privacy is of higher concerns (e.g., banks)
- Challenges:
  - Incentive mechanisms: participants might be competitors; utility fairness among clients (free-rider problem); dividing earning among participants, etc.
  - Preserving privacy on different levels (clients, users, etc.)

Applications of Federated Learning

- Cross-device FL:

Google Gboard

Apple QuickType

Apple "Hey Siri"

  - Google: Extensive use of cross-device FL in Gboard mobile keyboard, features on Pixel phones, and Android Messages
  - Apple: Use of cross-device FL in QuickType keyboard next word prediction and vocal classifier for "Hey Siri"
  - doc.ai uses cross-device FL for medical research, Snips uses cross-device FL for hotword detection, etc.

- Cross-silo FL:
  - Financial risk prediction for reinsurance, pharmaceutical discovery, electronic health record mining, medical data segmentation, smart manufacturing, etc.

Typical Federated Training Process

- Client selection:
  - Server samples from a set of available clients (idle, on wi-fi, plugged in...)
- Broadcast:
  - The selected clients download the current model weights
- Client computation:
  - Each selected client locally computes an update to the model by some algorithm (e.g., SGD or variants) on the local data
  - Potential additional processing: Privacy, compression, etc.
- Aggregation:
  - Server collects an aggregates of the updates from clients
  - Potential additional processing: filtering for security, etc.
- Model update:
  - The server updates the global model based on aggregated updates
  - Potential additional processing: additional scaling, momentum, extra data, etc.

Why Does Federated Learning Generate So Much Interest?

- FL is inherently inter-disciplinary:
  - Machine learning
  - Distributed optimization techniques
  - Cryptography
  - Security
  - Differential privacy
  - Fairness
  - Compressed sensing
  - Crowd-sensing
  - Wireless networking
  - Economics
  - Statistics
  - May play a role in emerging technologies (Blockchains, Metarverse, ...)
- Many of the hardest problems in FL are at the intersections of multiple areas

Optimization Algorithms for Federated Learning

- Key differences between distributed optimization and FL:
  - Non-i.i.d. and unbalanced datasets across clients
  - Limited communication bandwith
  - Unreliable and limited client device availability

- FedAvg Algorithm (aka Local SGD/parallel SGD): basic template of FL
  - N: Num. of clients; M: Clients per round;
  - T: Total communication round; K: Num. of local steps per round
  - At Server:
    1. Initialize $\mathbf{x}_0$
    2. **for** each round $t = 1, 2, \ldots, T$ **do**
       $S_t \leftarrow$ (random set of M clients)
       for each client $i \in S_t$ **in parallel do**
       $\mathbf{x}_i^{t+1} \leftarrow \text{ClientUpdate}(i, \bar{\mathbf{x}}^t)$
       $\bar{\mathbf{x}}^{t+1} \leftarrow (1/M)\sum_{i=1}^{M} \mathbf{x}_i^{t+1}$ $\rightarrow$: $\bar{x}^{t+1} = \bar{x}^t + \frac{1}{M}\sum_{i=1}^{M} \Delta_i^t$

Generalized FedAvg w/ 2-sided LR [Yang, Fang, Lin, ICLR'21]

  - ClientUpdate($i, \mathbf{x}$):
    1. $\mathbf{x}_0 \leftarrow \mathbf{x}$
    2. **for** local step $k = 0, \ldots, K-1$ **do**
       $\mathbf{x}_{k+1} \leftarrow \mathbf{x}_k - \mathbf{s}_k \nabla f(\mathbf{x}_k, \xi)$ for $\xi \sim \mathcal{P}_i$
    3. Return $\mathbf{x}_K$ to server

$\Delta_i^t = -\sum_{k=0}^{K-1} s_k \nabla f(x_k, \xi_k)$.

Convergence Results: FedAvg with I.I.D. Datasets

- Mini-batch of data used for a client's local update is statistically identical to a uniform sampling (with replacement) from the union of all clients' datasets
- Although unlikely in practice, i.i.d. case provides basic understanding for FL
- For simplicity, assume for now $M = N$. Consider the problem:

$$\min_{\mathbf{x} \in \mathbb{R}^m} f(\mathbf{x}) \triangleq \min_{\mathbf{x} \in \mathbb{R}^m} \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{x}),$$

  where $f_i(\mathbf{x}) \triangleq \mathbb{E}_{\xi_i \sim \mathcal{D}_i}[F_i(\mathbf{x}, \xi_i)]$ is nonconvex
- Assumptions:
  - L-smooth: $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y}$.
  - Bounded gradient variance and gradient second moments: $\mathbb{E}_{\xi_i \sim \mathcal{P}_i}[\|\nabla F(\mathbf{x}, \xi_i) - \nabla f_i(\mathbf{x})\|^2] \leq \sigma^2$, $\mathbb{E}_{\xi_i \in \mathbb{D}_i}[\|\nabla F_i(\mathbf{x}, \xi_i)\|^2] \leq G^2$, $\forall \mathbf{x}, i$
  - Unbiased stochastic gradient: $\mathbf{G}_i^t = \nabla F_i(\mathbf{x}_i^{t-1}, \xi_i^t)$ with $\mathbb{E}_{\xi_i^t \sim \mathcal{D}_i}[\mathbf{G}_i^t | \boldsymbol{\xi}^{[t-1]}] = \nabla f_i(\mathbf{x}_i^{t-1})$, $\forall i$, where $\boldsymbol{\xi}^{[t-1]} \triangleq [\xi_i^\tau]_{i \in [N], \tau \in [t-1]}$

Convergence Results: FedAvg with I.I.D. Datasets

To fix notation, we use the following equivalent code for FedAvg (also referred to as Parallel Restarted SGD in [Yu et al. AAAI'19]):

1. Initialize $\mathbf{x}_i^0 = \bar{\mathbf{y}} \in \mathbb{R}^m$. Choose constant step-size $s > 0$ and synchronization interval $K > 0$
2. **for** $t = 1, \ldots, T$ **do**
   Each client i observes stochastic gradient $\mathbf{G}_t^t$ of $f_i(\cdot)$ at $\mathbf{x}_i^{t-1}$
   if $t \mod K = 0$ **then**
   Compute node average $\mathbf{y} \triangleq \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i^{t-1}$
   Each client i in parallel updates its local solution

$$\mathbf{x}_i^t = \bar{\mathbf{y}} - s\mathbf{G}_i^t, \quad \forall i$$

   else
   Each client i in parallel updates its local solution:

$$\mathbf{x}_i^t = \mathbf{x}_i^{t-1} - s\mathbf{G}_i^t, \quad \forall i$$

   end if
   end for

Convergence Results: FedAvg with I.I.D. Datasets

Theorem 1 ([Yu et al. AAAI'19])

Under the stated assumptions and if $s \in (0, \frac{1}{L}]$*, then for all* $T \geq 1$*, then the iterates* $\{\mathbf{x}_t\}$ *generated by FedAvg satisfies:*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2] \leq \frac{2}{sT}(f(\bar{\mathbf{x}}^0) - f^*) + 4s^2K^2G^2L^2 + \frac{L}{N}s\sigma^2,$$

where f^* *is the optimal value of the FL problem.*

Theorem 1 ([Yu et al. AAAI'19])

Under the stated assumptions and if $s \in (0, \frac{1}{L}]$, then for all $T \geq 1$, then the iterates $\{\mathbf{x}_t\}$ generated by FedAvg satisfies:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2] \leq \frac{2}{sT}(f(\bar{\mathbf{x}}^0) - f^*) + 4s^2K^2G^2L^2 + \frac{L}{N}s\sigma^2,$$

$= O(\frac{1}{T})$

where f^* is the optimal value of the FL problem.

Proof: From L-smoothness and descent lemma:

$$\mathbb{E}[f(\bar{x}^t)] \leq \mathbb{E}[f(\bar{x}^{t-1})] + \mathbb{E}[\nabla f(\bar{x}^{t-1})^T(\bar{x}^t - \bar{x}^{t-1})] + \frac{L}{2}\mathbb{E}[\|\bar{x}^t - \bar{x}^{t-1}\|^2] \quad (1)$$

We first bnd the quadratic term:

$$\bar{x}^t \triangleq \frac{1}{N}\sum_{i=1}^{N} x_i^t = \frac{1}{N}\sum_{i=1}^{N}(x_i^{t-1} - sG_i^t) = \bar{x}^{t-1} - \frac{1}{N}s\sum_{i=1}^{N}G_i^t \quad (2).$$

Therefore, $\mathbb{E}[\|\bar{x}^t - \bar{x}^{t-1}\|_2^2] = \mathbb{E}[\|s\frac{1}{N}\sum_{i=1}^{N}G_i^t\|_2^2]$

$$= s^2\,\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}G_i^t\|^2]$$

add & subtract mean of G_i

$$\stackrel{}{=} s^2\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}(G_i^t - \nabla f_i(x_i^{t-1}))\|^2] + s^2\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^{t-1})\|^2]$$

$\mathbb{E}[\|z\|^2] = \mathbb{E}[\|z - \mathbb{E}[z]\|^2] + \|\mathbb{E}[z]\|^2$

($\mathbb{E}[\|z_1 + \cdots z_n\|^2] \leq \mathbb{E}[\|z_1\|^2 + \cdots + \|z_n\|^2] = \sum_{i=1}^{n}\mathbb{E}[\|z_i\|^2]$ for r.v. $z_1 \cdots z_n$ indep and 0-mean).

$$= \frac{s^2}{N^2}\sum_{i=1}^{N}\mathbb{E}[\underbrace{\|G_i^t - \nabla f_i(x_i^{t-1})\|^2}_{\leq \sigma^2}] + s^2\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^{t-1})\|^2]$$

$$\leq \frac{1}{N}s^2\sigma^2 + s^2\mathbb{E}[\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^{t-1})\|^2] \quad (3)$$

Now, for the cross term:

$$\mathbb{E}\left[\nabla f(\bar{x}^{t-1})^T(\bar{x}^{t}-\bar{x}^{t-1})\right] = -s\,\mathbb{E}\left[\nabla f(\bar{x}^{t-1})^T \cdot \frac{1}{N}\sum_{i=1}^{N} G_i^t\right]$$

$$\overset{\text{iter. law of } \mathbb{E}[\cdot]}{=} -s\,\mathbb{E}\left[\mathbb{E}\left[\nabla f(\bar{x}^{t-1})^T \frac{1}{N}\sum_{i=1}^{N} G_i^t \,\middle|\, \xi^{[t-1]}\right]\right]$$

$$= -s\,\mathbb{E}\left[\nabla f(\bar{x}^{t-1})^T \cdot \frac{1}{N}\sum_{i=1}^{N} \underbrace{\mathbb{E}\left[G_i^t \,\middle|\, \xi^{[t-1]}\right]}_{\text{unbiasedness}}\right]$$

$$= -s\,\mathbb{E}\left[\underbrace{\nabla f(\bar{x}^{t-1})}_{a}{}^T \underbrace{\frac{1}{N}\sum_{i=1}^{N} \nabla f_i(x_i^{t-1})}_{b}\right] \qquad a^T b = \frac{1}{2}\left(\|a\|^2+\|b\|^2-\|a-b\|^2\right)$$

$$= -\frac{s}{2}\,\mathbb{E}\left[\|\nabla f(\bar{x}^{t-1})\|^2 + \left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^{t-1})\right\|^2 - \left\|\nabla f(\bar{x}^{t-1}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^{t-1})\right\|^2\right] \tag{4}$$

Plugging (3) and (4) into (1):

$$\mathbb{E}[f(\bar{x}^t)] \le \mathbb{E}[f(\bar{x}^{t-1})] - \frac{s-s^2L}{2}\,\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^{t-1})\right\|^2\right]$$

$$-\frac{s}{2}\,\mathbb{E}\left[\|\nabla f(\bar{x}^{t-1})\|^2\right] + \frac{s}{2}\underbrace{\mathbb{E}\left[\left\|\nabla f(\bar{x}^{t-1}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^{t-1})\right\|^2\right]}_{(a)} + \frac{Ls^2\sigma^2}{2N}. \tag{7}$$

$$(a) \overset{\text{def. of } \bar{x}^{t-1}}{=} \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\bar{x}^{t-1}) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^{t-1})\right\|^2\right]$$

$$= \frac{1}{N^2}\,\mathbb{E}\left[\left\|\sum_{i=1}^{N}\left(\nabla f_i(\bar{x}^{t-1}) - \nabla f(x_i^{t-1})\right)\right\|^2\right]$$

$$\le \frac{1}{N}\,\mathbb{E}\left[\sum_{i=1}^{N}\underbrace{\left\|\nabla f_i(\bar{x}^{t-1}) - \nabla f_i(x_i^{t-1})\right\|^2}_{\le L^2\|\bar{x}^{t-1}-x_i^{t-1}\|^2}\right]$$

$$\left(\begin{aligned}&\mathbb{E}\left[\|z_1+\cdots+z_n\|^2\right] \\ &\le n\,\mathbb{E}\left[\|z_1\|^2+\cdots+\|z_n\|^2\right]\end{aligned}\right)$$

for r.v. $z_1 \cdots z_n$ not necc. indep.

with $n=N$ (★).

$$\leq \frac{L^2}{N}\sum_{i=1}^{N}\underbrace{\mathbb{E}\left[\|\bar{x}^{t-1}-x_i^{t-1}\|^2\right]}_{\text{client drift.}} \qquad (\Delta)$$

Lemma 1: (Client Drift): Under FedAvg, it holds that

$$\mathbb{E}\left[\|\bar{x}^{t}-x_i^{t}\|^2\right] \leq 4s^2K^2G^2.$$

Proof. For $t>1$ and $i\in[N]$. Recall FedAvg calculates client average $\bar{y} \triangleq \frac{1}{N}\sum_{i=1}^{N} x_i^{t-1}$ every k iters.

Consider the largest $t_0 \leq t$ s.t. $\bar{y}=\bar{x}^{t_0}$

From the updates in FedAvg:

$$x_i^t = \bar{y} - s\sum_{\tau=t_0+1}^{t} G_i^\tau \qquad (5)$$

Recall: $\bar{x}^t \triangleq \frac{1}{N}\sum_{i=1}^{N} x_i^t = \frac{1}{N}\sum_{i=1}^{N}\left(\bar{y}-s\sum_{\tau=t_0+1}^{t} G_i^\tau\right)$

$$= \bar{y} - s\sum_{\tau=t_0+1}^{t}\frac{1}{N}\sum_{i=1}^{N} G_i^\tau \qquad (6)$$

Using (5) and (6) in (Δ):

$$\mathbb{E}\left[\|x_i^t-\bar{x}^t\|^2\right] = \mathbb{E}\left[\Big\|\underbrace{s\sum_{\tau=t_0+1}^{t}\frac{1}{N}\sum_{i=1}^{N} G_i^\tau}_{z_1} - \underbrace{s\sum_{\tau=t_0+1}^{t} G_i^\tau}_{z_2}\Big\|^2\right]$$

Using (☆) with $n=2$.

$$\leq 2s^2\,\mathbb{E}\left[\Big\|\sum_{\tau=t_0+1}^{t}\frac{1}{N}\sum_{i=1}^{N} G_i^\tau\Big\|^2 + \Big\|\sum_{\tau=t_0+1}^{t} G_i^\tau\Big\|^2\right]$$

$$\leq 2s^2(t-t_0)\left[\sum_{\tau=t_0+1}^{t}\mathbb{E}\left[\Big\|\frac{1}{N}\sum_{i=1}^{N} G_i^\tau\Big\|^2\right] + \sum_{\tau=t_0+1}^{t}\mathbb{E}\left[\|G_i^\tau\|^2\right]\right]$$

$$\le 2s^2 \underbrace{(t-t_0)}_{\le K} \mathbb{E}\Big[\underbrace{\sum_{\tau=t_0+1}^{t}}_{\le K}\Big(\frac{1}{N}\sum_{i=1}^{N}\underbrace{\|G_i^\tau\|^2}_{\le G^2}\Big) + \underbrace{\sum_{\tau=t_0+1}^{t}}_{\le K}\underbrace{\|G_i^\tau\|^2}_{G^2}\Big]$$

$$\le 4s^2K^2G^2$$

Proof of Thm 1:

With Lemma 1, the descent lemma implies:

$$\mathbb{E}[f(\bar{x}^t)] \le \mathbb{E}[f(\bar{x}^{t-1})] - \frac{s-s^2L}{2}\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i^{t-1})\Big\|^2\Big]$$
$$- \frac{s}{2}\mathbb{E}[\|\nabla f(\bar{x}^{t-1})\|^2] + 2s^3K^2G^2L^2 + \frac{Ls^2\sigma^2}{2N}. \quad (8)$$

Note that by picking $s \le \frac{1}{L}$,

$$(8) \le \mathbb{E}[f(\bar{x}^{t-1})] - \frac{s}{2}\mathbb{E}[\|\nabla f(\bar{x}^{t-1})\|^2] + \underbrace{2s^3K^2G^2L^2}_{\text{client drift.}} + \frac{Ls^2\sigma^2}{2N}. \quad (9)$$

Dividing both sides of (9) by $\frac{s}{2}$ and rearranging:

$$\mathbb{E}[\|\nabla f(\bar{x}^{t-1})\|^2] \le \frac{2}{s}\Big(\mathbb{E}[f(\bar{x}^{t-1})] - \mathbb{E}[f(\bar{x}^t)]\Big) + 4s^2K^2G^2L^2 + \frac{Ls\sigma^2}{2N}.$$

Summing over $t \in [1, T]$, dividing both sides by T, and using $\mathbb{E}[f(\bar{x}^T)] \ge f^*$, we complete the proof. ∎

Convergence Results: FedAvg with I.I.D. Datasets

Corollary 2 ([Yu et al. AAAI'19])

- *If we let* $s = \frac{\sqrt{N}}{L\sqrt{T}}$*:*

$O\left(\frac{1}{\sqrt{T}}\right) \le \varepsilon \Rightarrow T \approx \frac{1}{\varepsilon^2}$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2] \le \underbrace{\frac{2L}{\sqrt{NT}}(f(\bar{\mathbf{x}}^0) - f^*) + 4\frac{N}{T}K^2G^2 + \frac{1}{\sqrt{NT}}\sigma^2}_{= O\left(\frac{1}{\sqrt{NT}}\right) \le \varepsilon \Rightarrow T = \frac{1}{N\varepsilon^2}}$$

- *If we further let* $K \le \frac{T^{1/4}}{N^{3/4}}$*:*

$K \le O\left(\frac{T^{\frac{1}{3}}}{N}\right) \to O\left(\frac{T}{N}\right)$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^{t-1})\|^2] \le \frac{2L}{\sqrt{NT}}(f(\bar{\mathbf{x}}^0) - f^*) + \frac{4}{\sqrt{NT}}G^2 + \frac{1}{\sqrt{NT}}\sigma^2$$

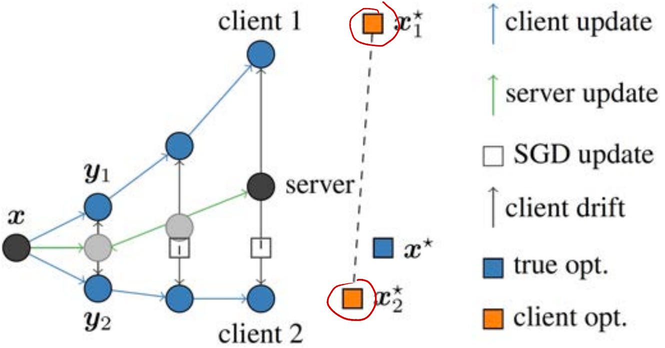
[Yang, Fang, Liu, ICLR'21]

G-FedAvg-TSLR.

s, s_L

Federated Learning with Non-I.I.D. Datasets

- "Client drift" problem with non-i.i.d. datasets (figure from [Karimireddy et al. ICML'20])



- Impose a limit on the number of local updates in FL with non-i.i.d. datasets (different algorithmic designs in FL lead to different limits)

What Do You Mean Exactly by Saying "Non-I.I.D" in FL?

- Bounded difference between client and global gradients (e.g., [Yu et al. ICML 2019] or [Yang et al. ICLR'21]):

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\right\|^2 \leq \sigma_G^2 \quad \text{or} \quad \left\|\nabla f_i(\mathbf{x}) - \nabla f(\mathbf{x})\right\|^2 \leq \sigma_G^2$$

- A unified bounded gradient dissimilarity (G, B)-BGD model [Karimireddy et al. ICML'20]:

$$\frac{1}{N}\sum_{i=1}^{N}\left\|\nabla f_i(\mathbf{x})\right\|^2 \leq G^2 + B^2\left\|\nabla f(\mathbf{x})\right\|^2$$

- Bounded difference between client and global optimal values (e.g., [Li et al., ICLR'20]):

$$f^* - \sum_{i=1}^{N} p_i f_i^* \triangleq \Gamma < \infty$$

Convergence Results: FedAvg with Non-I.I.D. Datasets

Theorem 3 ([Yu et al. ICML'19] Momentum-less Version)

Under the stated assumptions and if $s \in (0, \frac{1}{L}]$ *and* $K \leq \frac{1}{6Ls}$*, then for all* $T \geq 1$*, then the iterates* $\{\mathbf{x}_t\}$ *generated by FedAvg satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^t)\|^2] \leq \frac{2}{sT}(f(\bar{\mathbf{x}}^0) - f^*) + \frac{L}{N}s\sigma^2 + 4s^2KG^2L^2 \boxed{+ 9L^2s^2K^2\sigma_G^2,}$$

where f^* *is the optimal value of the FL problem.*

Convergence Results: FedAvg with Non-I.I.D. Datasets

Corollary 4 ([Yu et al. ICML'19])

- *If we let* $s = \frac{\sqrt{N}}{\sqrt{T}}$ *and* $K = 1$*, then for* $T \geq 36L^2N$

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^t)\|^2] = O\left(\frac{1}{\sqrt{NT}}\right) + O\left(\frac{N}{T}\right)$$

- *If we let* $s = \frac{\sqrt{N}}{\sqrt{T}}$ *and let* $K = O(\frac{T^{1/4}}{N^{3/4}})$*, then for* $T \geq L^2N$*:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}^t)\|^2] = O\left(\frac{1}{\sqrt{NT}}\right)$$

Next Class

Decentralized Consensus Optimization