Hierarchical Federated Learning with Multi-Timescale Gradient Correction

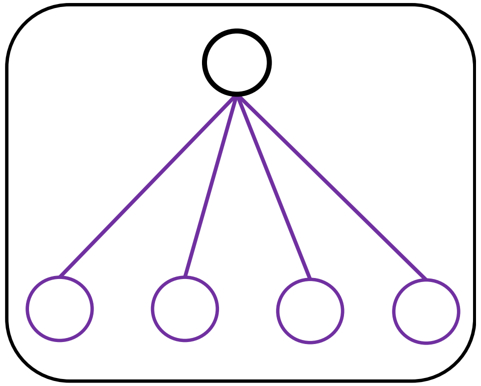
Wenzhi Fang (Purdue), Dong-Jun Han (Yonsei University), Evan Chen (Purdue),

Shiqiang Wang (IBM), Christopher G. Brinton (Purdue)

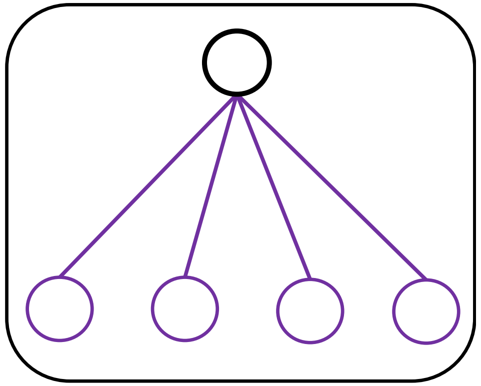
NEURAL INFORMATION PROCESSING SYSTEMS

Motivation of HFL

- Federated learning
  - Devices directly communicate with the cloud server



Motivation of HFL

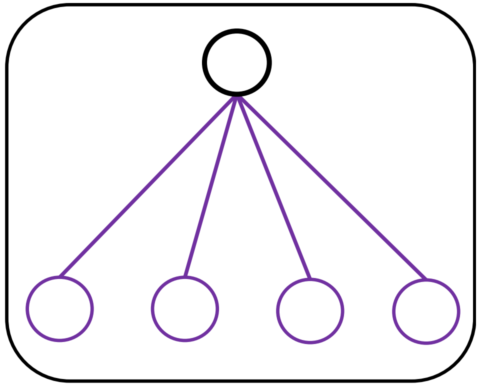


- Federated learning
  - Devices directly communicate with the cloud server

- Some potential problems
  - Topology of practical networks, e.g., fog learning system

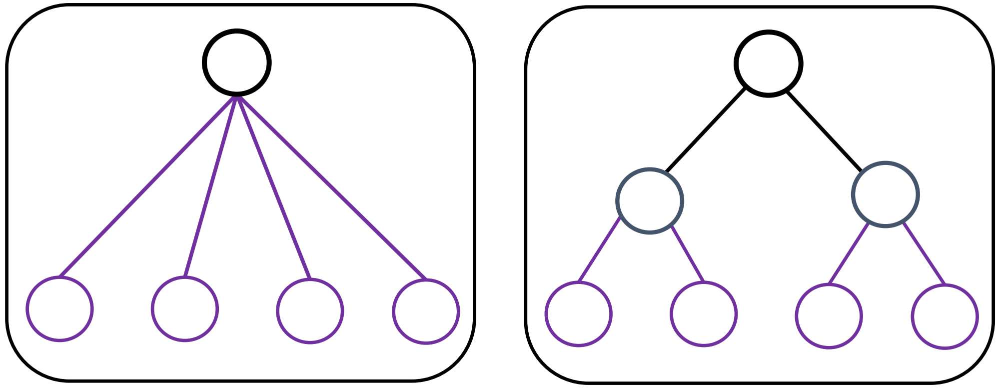
Motivation of HFL

- Federated learning
  - Devices directly communicate with the cloud server



- Some potential problems
  - Topology of practical networks, e.g., fog learning system
  - Large communication latency between devices and remote server

Motivation of HFL



- Federated learning
  - Devices directly communicate with the cloud server
- Some potential problems
  - Topology of practical networks, e.g., fog learning system
  - Large communication latency between devices and remote server
- Hierarchical Federated Learning
  - Reduce the communication frequency with the cloud server
  - Group devices into multiple cells and introduce edge servers to coordinate the training within each cell

Problem Formulation

- Training Objective

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) := \frac{1}{N}\sum_{j=1}^{N} f_j(\boldsymbol{x}), \text{ where } f_j(\boldsymbol{x}) := \frac{1}{n_j}\sum_{i\in\mathcal{C}_j} F_i(\boldsymbol{x}) \text{ and } F_i(\boldsymbol{x}) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[F_i(\boldsymbol{x},\xi_i)]$$

  - $f(\boldsymbol{x})$ denotes the global loss, $f_j(\boldsymbol{x})$ is the group loss, $F_i(\boldsymbol{x})$ represents the client loss

Problem Formulation

- Training Objective

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) := \frac{1}{N}\sum_{j=1}^{N} f_j(\boldsymbol{x}), \text{ where } f_j(\boldsymbol{x}) := \frac{1}{n_j}\sum_{i\in\mathcal{C}_j} F_i(\boldsymbol{x}) \text{ and } F_i(\boldsymbol{x}) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[F_i(\boldsymbol{x},\xi_i)]$$

  - $f(\boldsymbol{x})$ denotes the global loss, $f_j(\boldsymbol{x})$ is the group loss, $F_i(\boldsymbol{x})$ represents the client loss

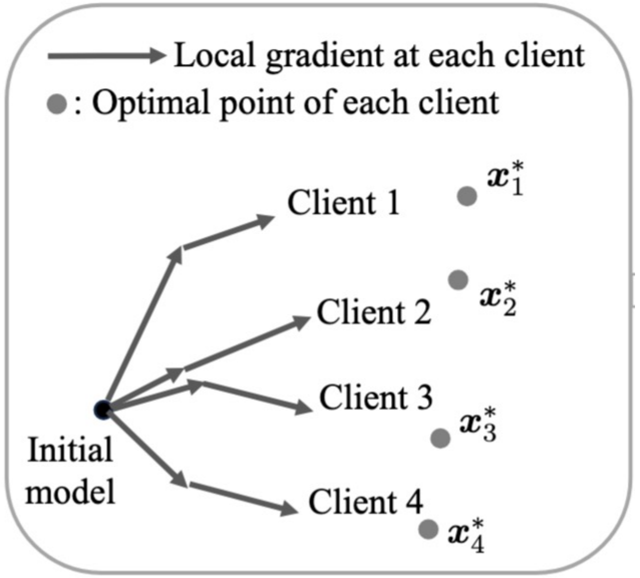
- HFedAvg algorithm: main procedures

1. Local model updates at devices
   - Conducting SGD iterations
2. Edge server aggregates local models from clients within its coverage
   - Communication period aggregation happens every *H* local iterations
3. Global model aggregations
   - Communication period Aggregation happens every *E* edge aggregations

Challenge within HFL

- Challenges
  - Data heterogeneity across clients and groups
  - Local models deviate from the global optimum

$$\boldsymbol{x}^* \neq \boldsymbol{x}^* - \gamma \nabla F_i(\boldsymbol{x}^*)$$



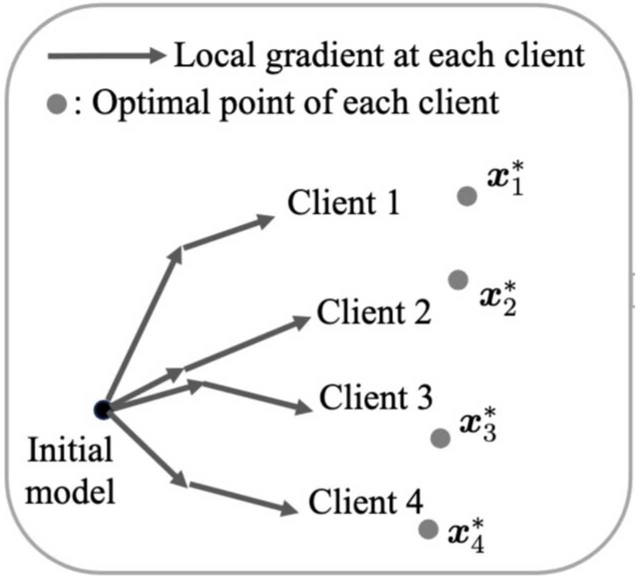
(a) No gradient correction

Challenge within HFL

- Challenges
  - Data heterogeneity across clients and groups
  - Local models deviate from the global optimum

$$\boldsymbol{x}^* \neq \boldsymbol{x}^* - \gamma \nabla F_i(\boldsymbol{x}^*)$$

- Existing schemes
  - No correction: converging to client-level local minimum as shown in fig. (a)



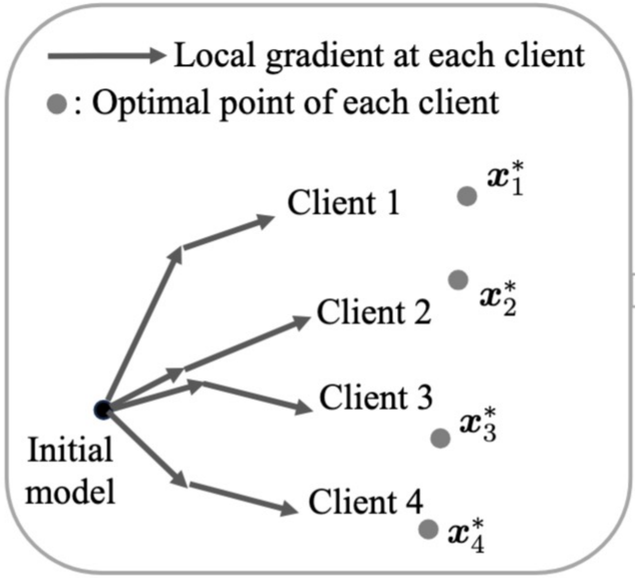
(a) No gradient correction

Challenge within HFL

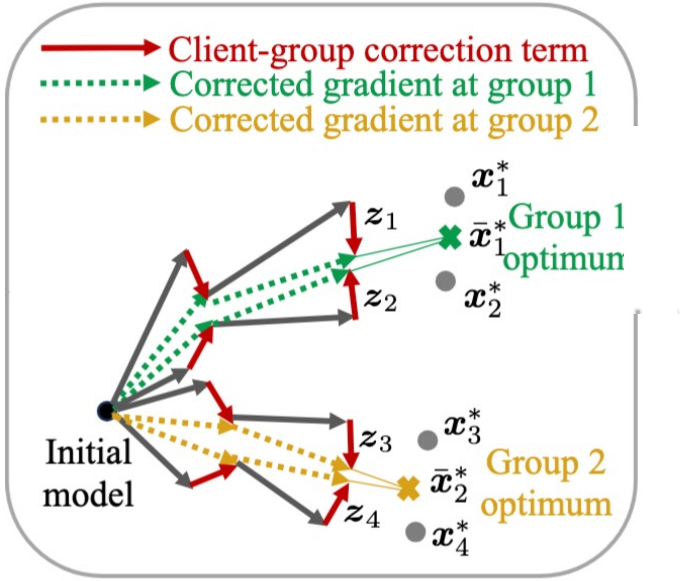
- Challenges
  - Data heterogeneity across clients and groups
  - Local models deviate from the global optimum

$$\boldsymbol{x}^* \neq \boldsymbol{x}^* - \gamma \nabla F_i(\boldsymbol{x}^*)$$

- Existing schemes
  - No correction: converging to client-level local minimum as shown in fig. (a)
  - Only client-group correction: converging to the group-level optimum as shown in fig. (b)



(a) No gradient correction



(b) Only client-group gradient correction

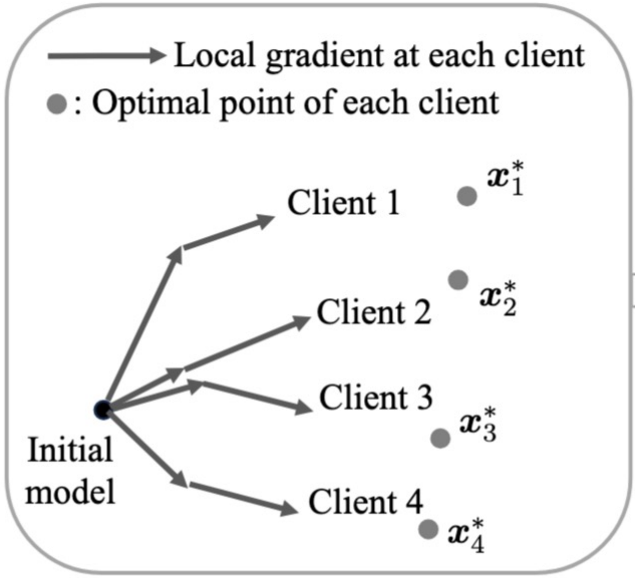
Challenge within HFL

- Challenges
  - Data heterogeneity across clients and groups
  - Local models deviate from the global optimum

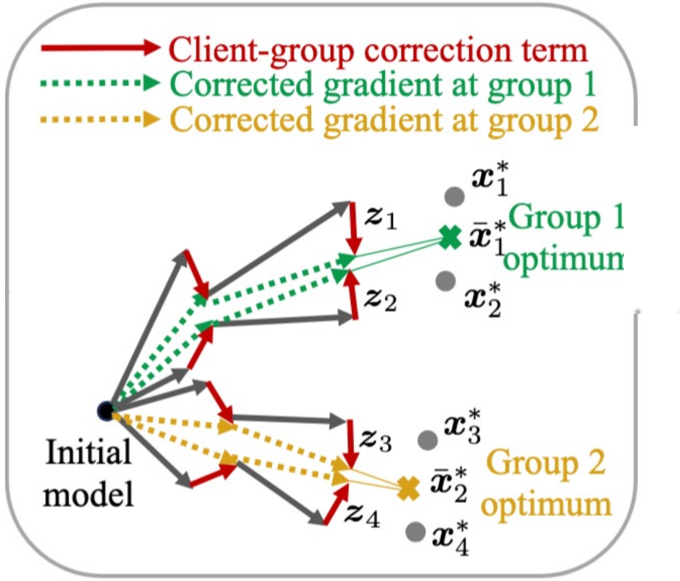
$$\boldsymbol{x}^* \neq \boldsymbol{x}^* - \gamma \nabla F_i(\boldsymbol{x}^*)$$

- Existing schemes
  - No correction: converging to client-level local minimum as shown in fig. (a)
  - Only client-group correction: converging to the group-level optimum as shown in fig. (b)
- Observation

$$\boldsymbol{x}_{\text{new}} = \boldsymbol{x}^* - \gamma\{\nabla F_i(\boldsymbol{x}^*) + \underbrace{(\nabla f_j(\boldsymbol{x}^*) - \nabla F_i(\boldsymbol{x}^*))}_{\text{client-group correction}} + \underbrace{(\nabla f(\boldsymbol{x}^*) - \nabla f_j(\boldsymbol{x}^*))}_{\text{group-global correction}}\}$$

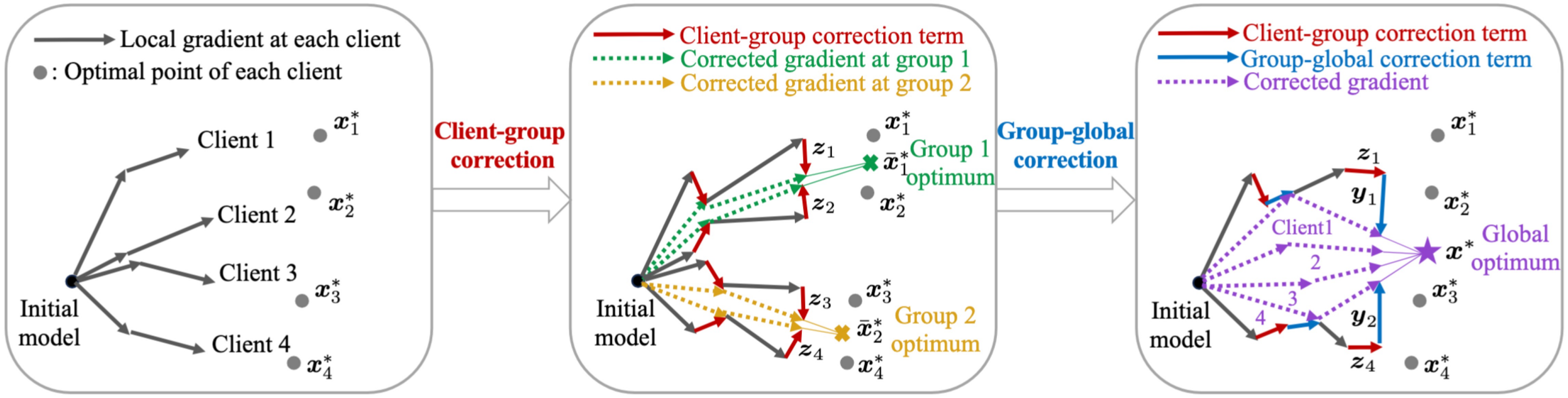


(a) No gradient correction



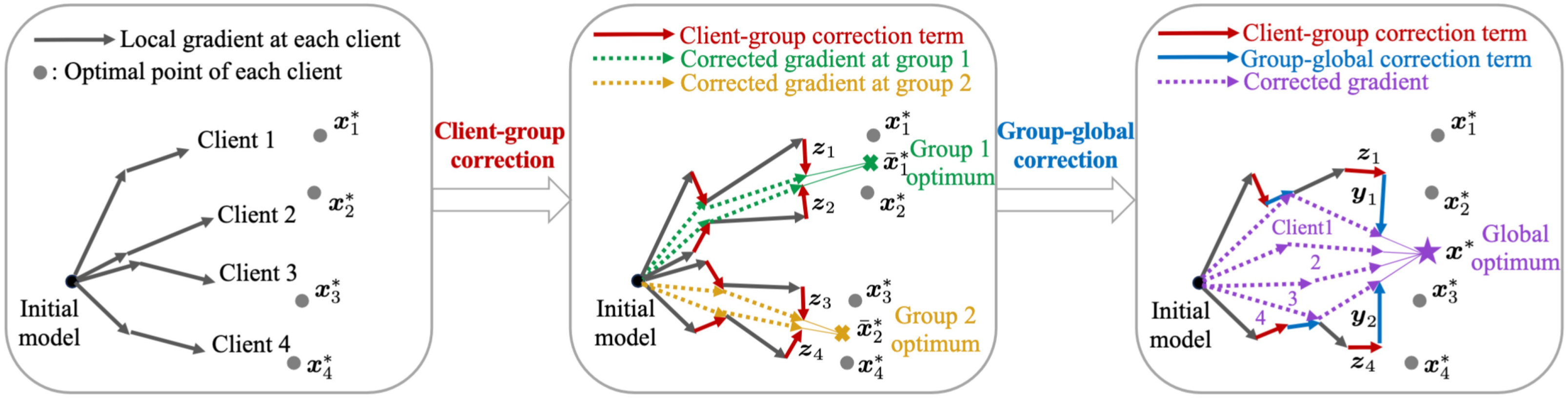
(b) Only client-group gradient correction

Multi-timescale gradient correction



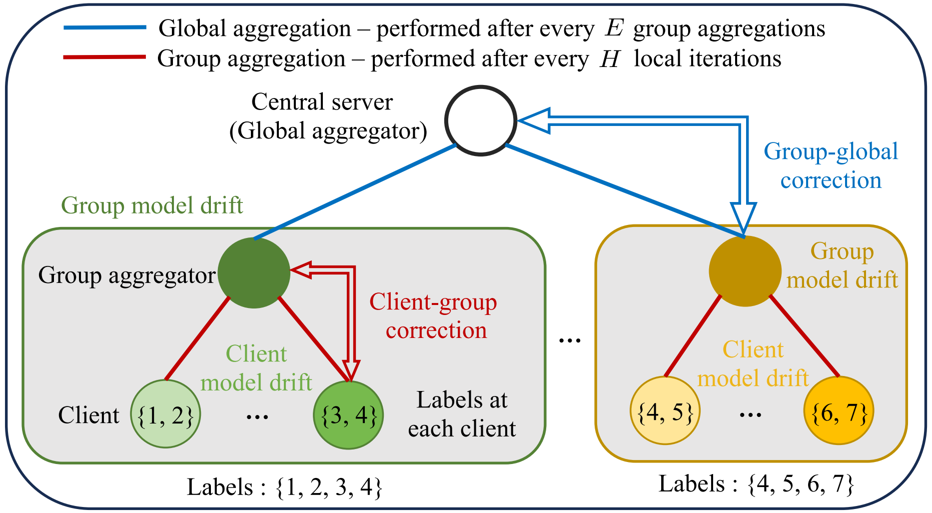
(a) No gradient correction (b) Only client-group gradient correction (c) Multi-timescale gradient correction (Ours)

Multi-timescale gradient correction



(a) No gradient correction (b) Only client-group gradient correction (c) Multi-timescale gradient correction (Ours)

- Proposed Algorithm: $\boldsymbol{x}_{i,h+1}^{t,e} = \boldsymbol{x}_{i,h}^{t,e} - \gamma\left(\nabla F_i\left(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}\right) + \boldsymbol{z}_i^{t,e} + \boldsymbol{y}_j^t\right)$



Multi-timescale gradient correction

- Update correction variables via accumulated gradients

Algorithm 1: HFL with Multi-Timescale Gradient Correction (MTGC)

Input: Initial model $\bar{\boldsymbol{x}}^0$, global aggregation period E, group aggregation period H, learning rate γ, and group-global correction $\boldsymbol{y}_j^0 = -\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\boldsymbol{x}_{i,0}^{t,0}, \xi_{i,0}^{0,0}) + \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\boldsymbol{x}_{i,0}^{0,0}, \xi_{i,0}^{0,0}), \forall j$

each global round $t = 0, 1, \ldots, T-1$ **do**

- Group model initialization: $\boldsymbol{x}_j^{t,0} = \bar{\boldsymbol{x}}^t, \forall j$
- Client-group correction initialization:
  $\boldsymbol{z}_i^{t,0} = -\nabla F_i(\boldsymbol{x}_{i,0}^{t,0}, \xi_{i,0}^{t,0}) + \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\boldsymbol{x}_{i,0}^{t,0}, \xi_{i,0}^{t,0}), \forall i \in \mathcal{C}_j, \forall j$
- **each group communication round** $e = 0, 1, \ldots, E-1$ **do**
  - Local model initialization: $\boldsymbol{x}_{i,0}^{t,e} = \bar{\boldsymbol{x}}_j^{t,e}, \forall i, j$
  - **each local iteration** $h = 0, 1, \ldots, H-1$ **do**
    - $\boldsymbol{x}_{i,h+1}^{t,e} = \boldsymbol{x}_{i,h}^{t,e} - \gamma\left(\nabla F_i(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}) + \boldsymbol{z}_i^{t,e} + \boldsymbol{y}_j^t\right), \forall i \in \mathcal{C}_j, \forall j$ ◇ Clients do in parallel
  - Group aggregation: $\bar{\boldsymbol{x}}_j^{t,e+1} = \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\boldsymbol{x}_{i,H}^{t,e}$
  - Client-group corr. update: $\boldsymbol{z}_i^{t,e+1} = \boldsymbol{z}_i^{t,e} + \frac{1}{H\gamma}\left(\boldsymbol{x}_{i,H}^{t,e} - \bar{\boldsymbol{x}}_j^{t,e+1}\right), \forall i \in \mathcal{C}_j, \forall j$ ◇ Clients do in parallel
- Global aggregation: $\bar{\boldsymbol{x}}^{t+1} = \frac{1}{N}\sum_{j=1}^{N}\bar{\boldsymbol{x}}_j^{t,E}$
- Group-global corr. update: $\boldsymbol{y}_j^{t+1} = \boldsymbol{y}_j^t + \frac{1}{HE\gamma}\left(\bar{\boldsymbol{x}}_j^{t,E} - \bar{\boldsymbol{x}}^{t+1}\right), \forall j$ ◇ Group aggregators in parallel

Multi-timescale gradient correction

- Update correction variables via accumulated gradients

Algorithm 1: HFL with Multi-Timescale Gradient Correction (MTGC)

Input: Initial model $\bar{\boldsymbol{x}}^0$, global aggregation period E, group aggregation period H, learning rate γ, and group-global correction $\boldsymbol{y}_j^0 = -\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\boldsymbol{x}_{i,0}^{t,0},\xi_{i,0}^{0,0}) + \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\boldsymbol{x}_{i,0}^{0,0},\xi_{i,0}^{0,0}), \forall j$

each global round $t = 0, 1, \dots, T-1$ **do**

  Group model initialization: $\boldsymbol{x}_j^{t,0} = \bar{\boldsymbol{x}}^t, \forall j$

  Client-group correction initialization:

  $\boldsymbol{z}_i^{t,0} = -\nabla F_i(\boldsymbol{x}_{i,0}^{t,0},\xi_{i,0}^{t,0}) + \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\boldsymbol{x}_{i,0}^{t,0},\xi_{i,0}^{t,0}), \forall i\in\mathcal{C}_j, \forall j$

  each group communication round $e = 0, 1, \dots, E-1$ **do**

    Local model initialization: $\boldsymbol{x}_{i,0}^{t,e} = \bar{\boldsymbol{x}}_j^{t,e}, \forall i, j$

    each local iteration $h = 0, 1, \dots, H-1$ **do**

      $\boldsymbol{x}_{i,h+1}^{t,e} = \boldsymbol{x}_{i,h}^{t,e} - \gamma\left(\nabla F_i(\boldsymbol{x}_{i,h}^{t,e},\xi_{i,h}^{t,e}) + \boldsymbol{z}_i^{t,e} + \boldsymbol{y}_j^t\right), \forall i\in\mathcal{C}_j, \forall j$ ◇ Clients do in parallel

    Group aggregation: $\bar{\boldsymbol{x}}_j^{t,e+1} = \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\boldsymbol{x}_{i,H}^{t,e}$

    Client-group corr. update: $\boldsymbol{z}_i^{t,e+1} = \boldsymbol{z}_i^{t,e} + \frac{1}{H\gamma}\left(\boldsymbol{x}_{i,H}^{t,e} - \bar{\boldsymbol{x}}_j^{t,e+1}\right), \forall i\in\mathcal{C}_j, \forall j$ ◇ Clients do in parallel

  Global aggregation: $\bar{\boldsymbol{x}}^{t+1} = \frac{1}{N}\sum_{j=1}^{N}\bar{\boldsymbol{x}}_j^{t,E}$

  Group-global corr. update: $\boldsymbol{y}_j^{t+1} = \boldsymbol{y}_j^t + \frac{1}{HE\gamma}\left(\bar{\boldsymbol{x}}_j^{t,E} - \bar{\boldsymbol{x}}^{t+1}\right), \forall j$ ◇ Group aggregators in parallel

Client-group correction
Update after each edge aggregation

Multi-timescale gradient correction

- Update correction variables via accumulated gradients

Algorithm 1: HFL with Multi-Timescale Gradient Correction (MTGC)

Input: Initial model $\bar{\boldsymbol{x}}^0$, global aggregation period E, group aggregation period H, learning rate γ, and group-global correction $\boldsymbol{y}_j^0 = -\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\boldsymbol{x}_{i,0}^{t,0}, \xi_{i,0}^{0,0}) + \frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\boldsymbol{x}_{i,0}^{0,0}, \xi_{i,0}^{0,0}), \forall j$

1. **each global round** $t = 0, 1, \ldots, T-1$ **do**
2. Group model initialization: $\boldsymbol{x}_j^{t,0} = \bar{\boldsymbol{x}}^t, \forall j$
3. Client-group correction initialization: $\boldsymbol{z}_i^{t,0} = -\nabla F_i(\boldsymbol{x}_{i,0}^{t,0}, \xi_{i,0}^{t,0}) + \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\nabla F_i(\boldsymbol{x}_{i,0}^{t,0}, \xi_{i,0}^{t,0}), \forall i \in \mathcal{C}_j, \forall j$
4. **each group communication round** $e = 0, 1, \ldots, E-1$ **do**
5. Local model initialization: $\boldsymbol{x}_{i,0}^{t,e} = \bar{\boldsymbol{x}}_j^{t,e}, \forall i, j$
6. **each local iteration** $h = 0, 1, \ldots, H-1$ **do**
7. $\boldsymbol{x}_{i,h+1}^{t,e} = \boldsymbol{x}_{i,h}^{t,e} - \gamma\left(\nabla F_i(\boldsymbol{x}_{i,h}^{t,e}, \xi_{i,h}^{t,e}) + \boldsymbol{z}_i^{t,e} + \boldsymbol{y}_j^t\right), \forall i \in \mathcal{C}_j, \forall j$ ◇ Clients do in parallel
8. Group aggregation: $\bar{\boldsymbol{x}}_j^{t,e+1} = \frac{1}{n_j}\sum_{i\in\mathcal{C}_j}\boldsymbol{x}_{i,H}^{t,e}$
9. Client-group corr. update: $\boldsymbol{z}_i^{t,e+1} = \boldsymbol{z}_i^{t,e} + \frac{1}{H\gamma}\left(\boldsymbol{x}_{i,H}^{t,e} - \bar{\boldsymbol{x}}_j^{t,e+1}\right), \forall i \in \mathcal{C}_j, \forall j$ ◇ Clients do in parallel
10. Global aggregation: $\bar{\boldsymbol{x}}^{t+1} = \frac{1}{N}\sum_{j=1}^{N}\bar{\boldsymbol{x}}_j^{t,E}$
11. Group-global corr. update: $\boldsymbol{y}_j^{t+1} = \boldsymbol{y}_j^t + \frac{1}{HE\gamma}\left(\bar{\boldsymbol{x}}_j^{t,E} - \bar{\boldsymbol{x}}^{t+1}\right), \forall j$ ◇ Group aggregators in parallel

Client-group correction
Update after each edge aggregation

Group-global correction
Update after each global aggregation

Convergence Analysis

- The iterates generated by MTGC satisfy

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{F}_0 L\sigma^2}{\tilde{N}TEH}} + \left(\frac{\mathcal{F}_0 L\sigma}{T}\right)^{\frac{2}{3}} + \frac{L\mathcal{F}_0}{T}\right) \qquad \mathcal{F}_0 = f\left(\overline{\boldsymbol{x}}^0\right) - f^*$$

Convergence Analysis

- The iterates generated by MTGC satisfy

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{F}_0 L\sigma^2}{\tilde{N}TEH}} + \left(\frac{\mathcal{F}_0 L\sigma}{T}\right)^{\frac{2}{3}} + \frac{L\mathcal{F}_0}{T}\right) \qquad \mathcal{F}_0 = f\left(\overline{\boldsymbol{x}}^0\right) - f^*$$

  - This convergence rate is dominated by the first term as $T \to \infty$.

Convergence Analysis

- The iterates generated by MTGC satisfy

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 \le \mathcal{O}\left(\sqrt{\frac{\mathcal{F}_0 L\sigma^2}{\tilde{N}TEH}} + \left(\frac{\mathcal{F}_0 L\sigma}{T}\right)^{\frac{2}{3}} + \frac{L\mathcal{F}_0}{T}\right) \qquad \mathcal{F}_0 = f\left(\overline{\boldsymbol{x}}^0\right) - f^*$$

  - This convergence rate is dominated by the first term as $T \to \infty$.
  - Linear speedup in the number of group aggregations *E*

$\tilde{N}$

Convergence Analysis

- The iterates generated by MTGC satisfy

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{F}_0 L\sigma^2}{\tilde{N}TEH}}+\left(\frac{\mathcal{F}_0 L\sigma}{T}\right)^{\frac{2}{3}}+\frac{L\mathcal{F}_0}{T}\right) \qquad \mathcal{F}_0 = f\left(\overline{\boldsymbol{x}}^0\right)-f^*$$

  - This convergence rate is dominated by the first term as $T \to \infty$.
  - Linear speedup in the number of group aggregations *E*
  - Linear speedup in the number of local updates *H*

Convergence Analysis

- The iterates generated by MTGC satisfy

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 \le \mathcal{O}\left(\sqrt{\frac{\mathcal{F}_0 L\sigma^2}{\tilde{N}TEH}} + \left(\frac{\mathcal{F}_0 L\sigma}{T}\right)^{\frac{2}{3}} + \frac{L\mathcal{F}_0}{T}\right) \qquad \mathcal{F}_0 = f\left(\overline{\boldsymbol{x}}^0\right) - f^*$$

  - This convergence rate is dominated by the first term as $T \to \infty$.
  - Linear speedup in the number of group aggregations *E*
  - Linear speedup in the number of local updates *H*
  - Linear speedup in the number of clients $\tilde{N}$

Convergence Analysis

- The iterates generated by MTGC satisfy

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathbb{E}\left\|\nabla f\left(\hat{\boldsymbol{x}}^{t,e}\right)\right\|^2 \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{F}_0 L\sigma^2}{\tilde{N}TEH}} + \left(\frac{\mathcal{F}_0 L\sigma}{T}\right)^{\frac{2}{3}} + \frac{L\mathcal{F}_0}{T}\right) \qquad \mathcal{F}_0 = f\left(\bar{\boldsymbol{x}}^0\right) - f^*$$

  - This convergence rate is dominated by the first term as $T \to \infty$.
  - Linear speedup in the number of group aggregations *E*
  - Linear speedup in the number of local updates *H*
  - Linear speedup in the number of clients $\tilde{N}$

- ***For the first time*** attain these results for HFL without relying on data heterogeneity assumptions

Experimental Results

- Baselines
  - SCAFFOLD: with a single-level gradient correction
  - FedProx: prevent local overfitting with a proximal regularizer
  - FedDyn: based on a dynamic regularization term
  - H-FedAvg: no gradient correction

Experimental Results

- Baselines
  - SCAFFOLD: with a single-level gradient correction
  - FedProx: prevent local overfitting with a proximal regularizer
  - FedDyn: based on a dynamic regularization term
  - H-FedAvg: no gradient correction