Carnegie Mellon University

Towards Optimal Heterogeneous Client Sampling in Multi-Model Federated Learning

SEP 27, 2024

Haoran Zhang, Zejun Gong, Zekai Li, Marie Siew,
Carlee Joe-Wong, Rachid El-Azouzi

Outline

- Introduction
    - Federated learning (FL)
    - Multi-model federated learning (MMFL)
- Variance-reduced client sampling in a simple MMFL system
- Modeling computational heterogeneity in MMFL
- Experiments

Federated Learning

Distributed learning with unshared local data

Server:

1 Receive updates from clients

2 Aggregate local updates for a better global model

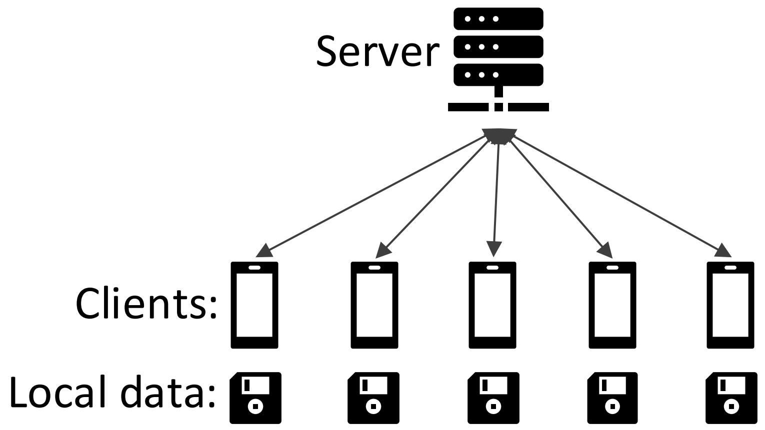
3 Broadcast new model parameters to clients

Local client (device):

1 Get global model parameters

2 Train model parameters with local data

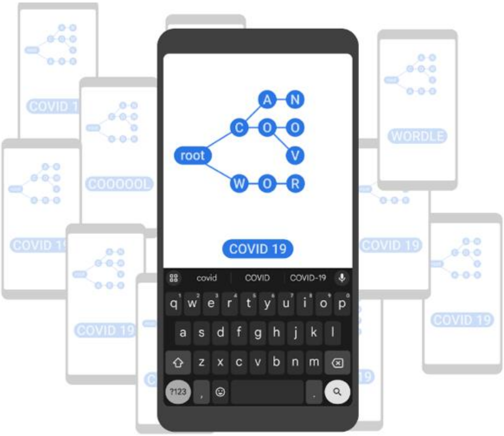
3 Send updated parameters to the server



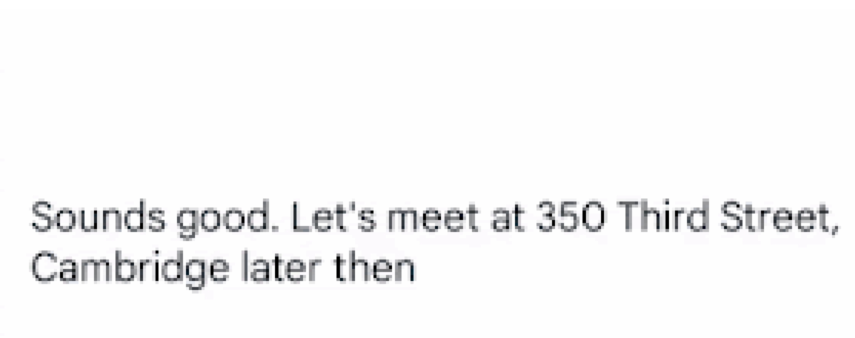
Multi-Model Federated Learning

Examples: Multiple FL applications on one device.

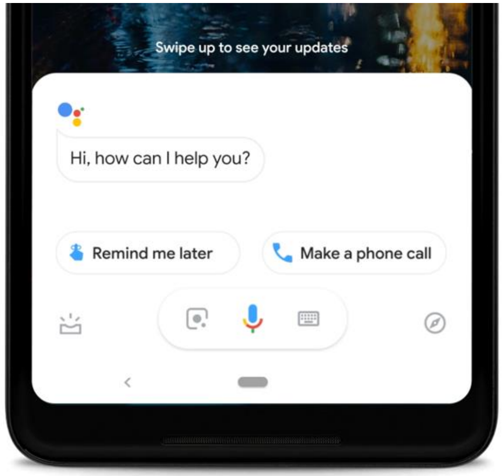
Keyboard prediction



Predicting text selection



Speech model



Source: federated.withgoogle.com

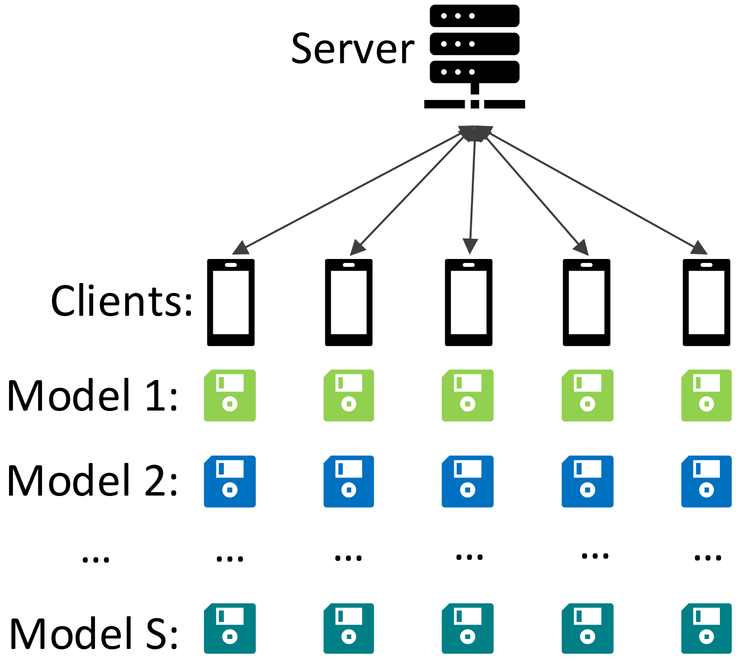
Multi-Model Federated Learning

Key assumptions from previous work [1]

In each round, the server only allows partial participation, and each active client can only train one model.

1) Partial Participation: reduce communication cost

2) Only train one model: computational constraints



Multi-model federated learning

[1] Bhuyan, Neelkamal, Sharayu Moharir, and Gauri Joshi. "Multi-model federated learning with provable guarantees." *EAI International Conference on Performance Evaluation Methodologies and Tools*. Cham: Springer Nature Switzerland, 2022.

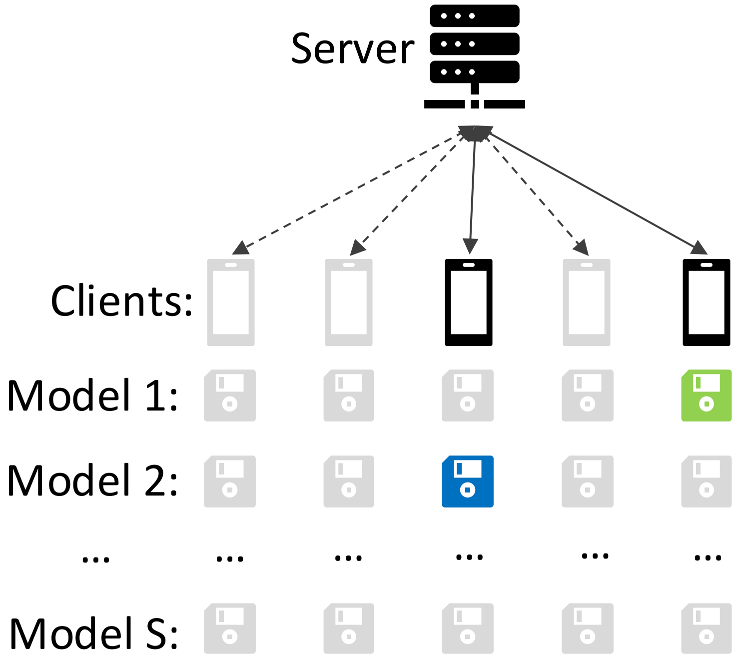
Multi-Model Federated Learning

Key assumptions from previous work [1]

In each round, the server only allows partial participation, and each active client can only train one model.

1) Partial Participation: reduce communication cost

2) Only train one model: computational constraints



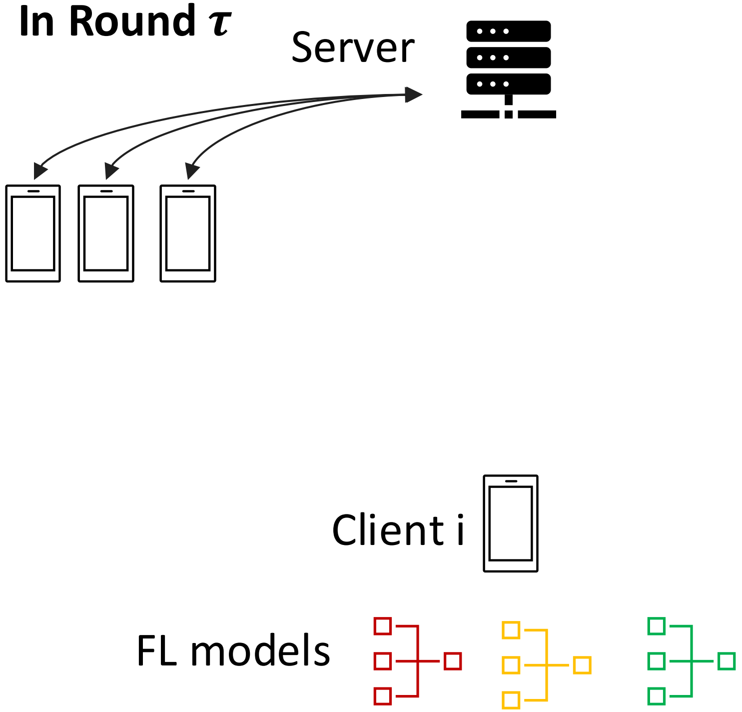
Multi-model federated learning

[1] Bhuyan, Neelkamal, Sharayu Moharir, and Gauri Joshi. “Multi-model federated learning with provable guarantees.” *EAI International Conference on Performance Evaluation Methodologies and Tools*. Cham: Springer Nature Switzerland, 2022.

MMFL Optimal Variance-Reduced Sampling

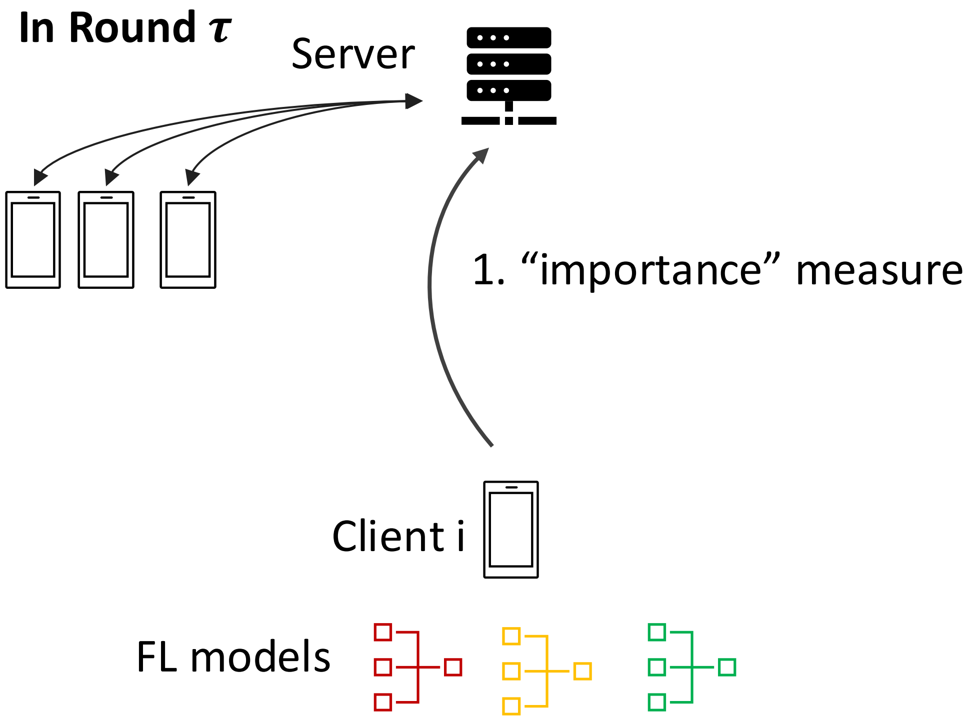
Idea: the server prefers selecting more "important" clients.

In Round τ



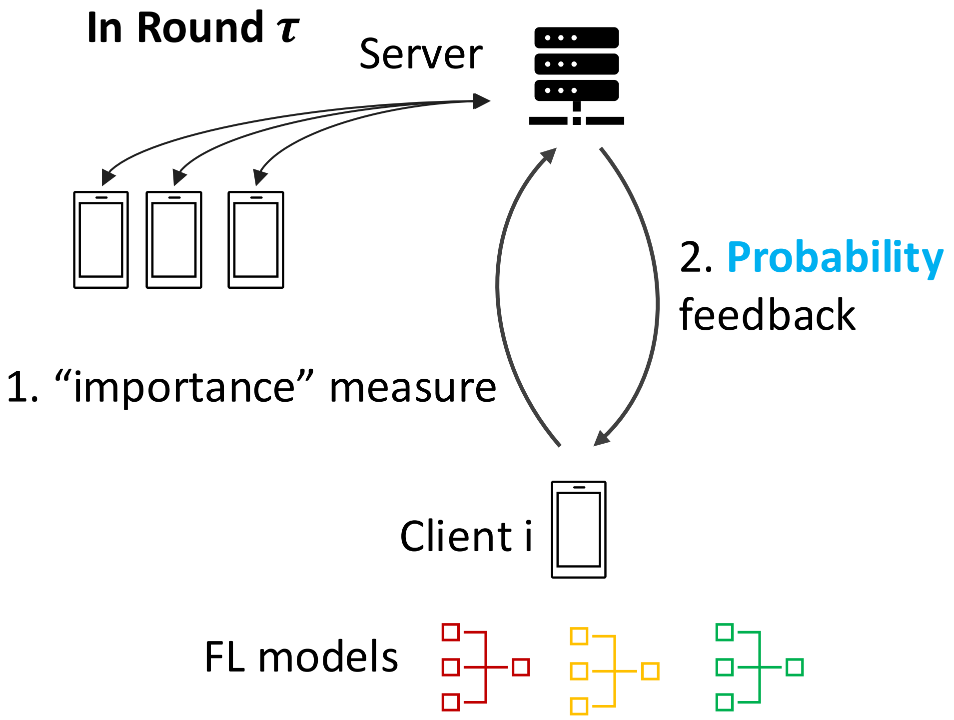
MMFL Optimal Variance-Reduced Sampling

Idea: the server prefers selecting more “important” clients.



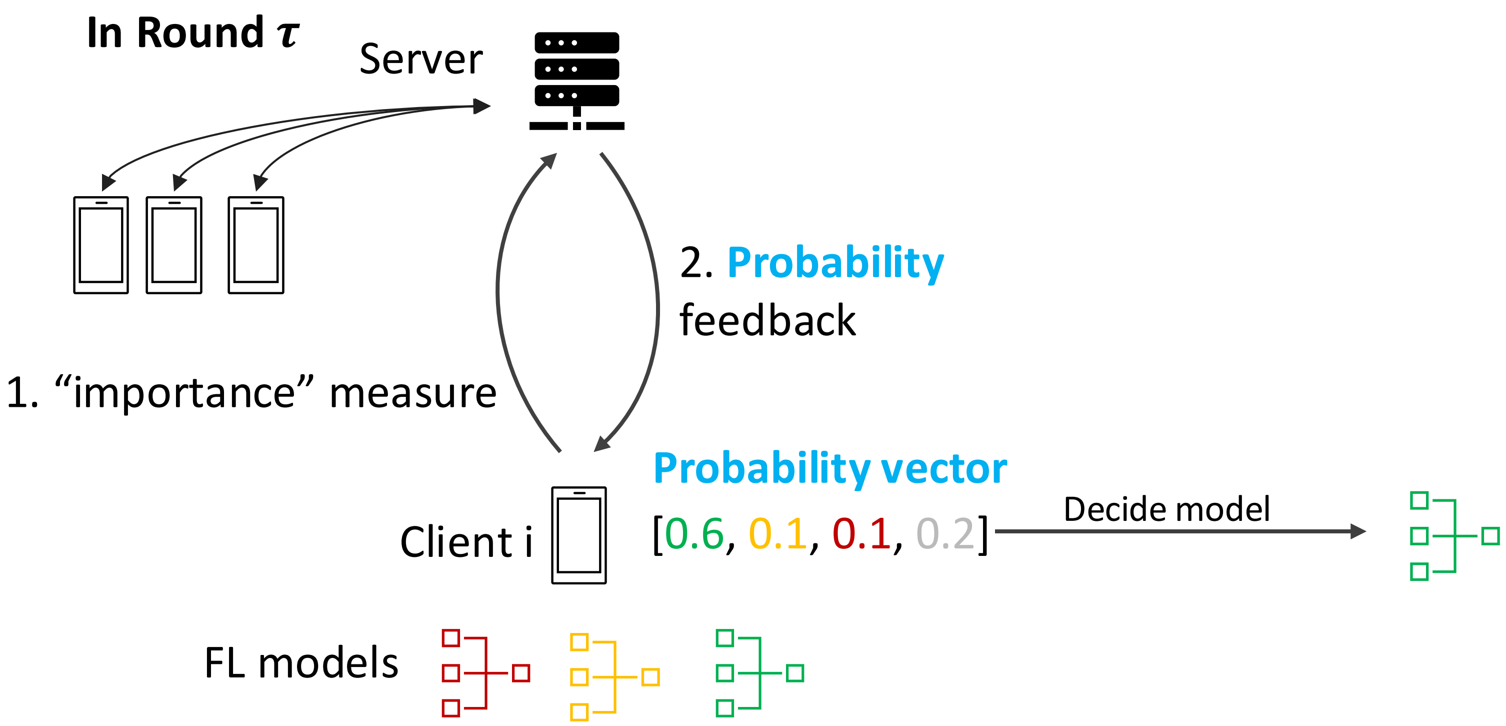
MMFL Optimal Variance-Reduced Sampling

Idea: the server prefers selecting more "important" clients.



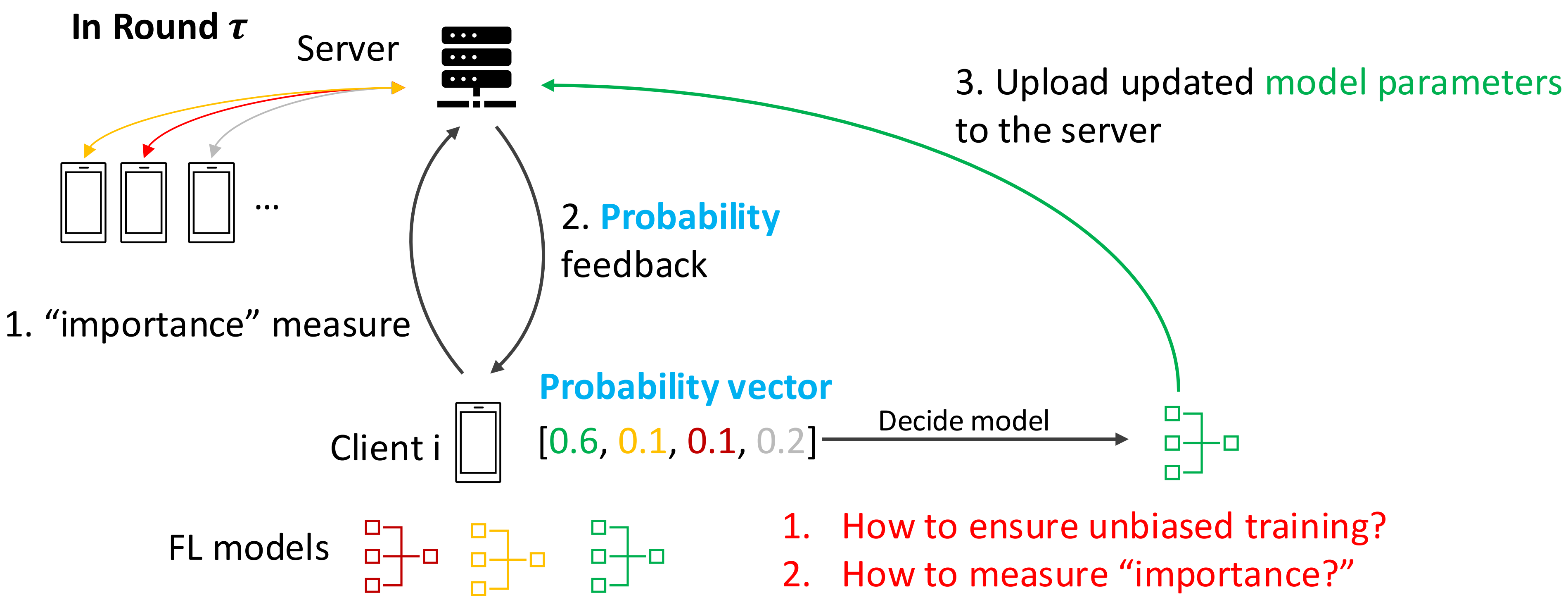
MMFL Optimal Variance-Reduced Sampling

Idea: the server prefers selecting more “important” clients.



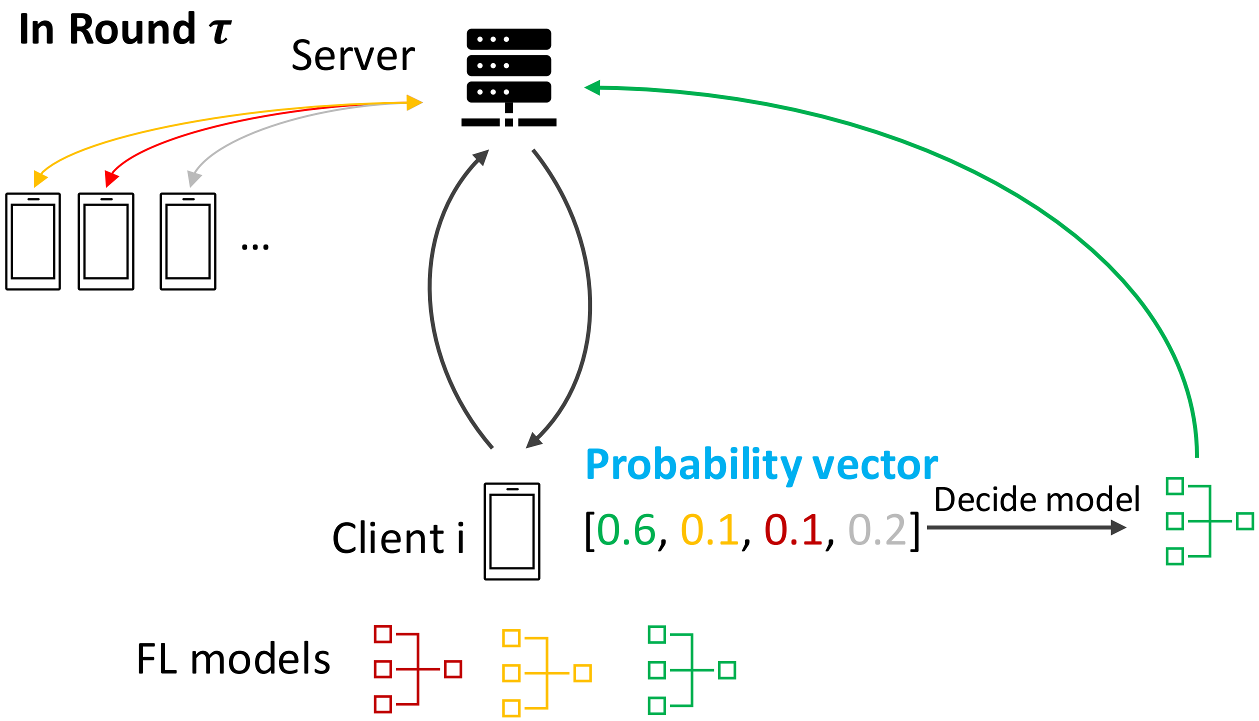
MMFL Optimal Variance-Reduced Sampling

Idea: the server prefers selecting more "important" clients.



1. How to ensure unbiased training?
2. How to measure "importance?"

MMFL Optimal Variance-Reduced Sampling



In each global round (Aggregation):

$$w_s^{\tau+1} = w_s^{\tau} - \sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p_{s|i}^{\tau}} U_{i,s}^{\tau}$$

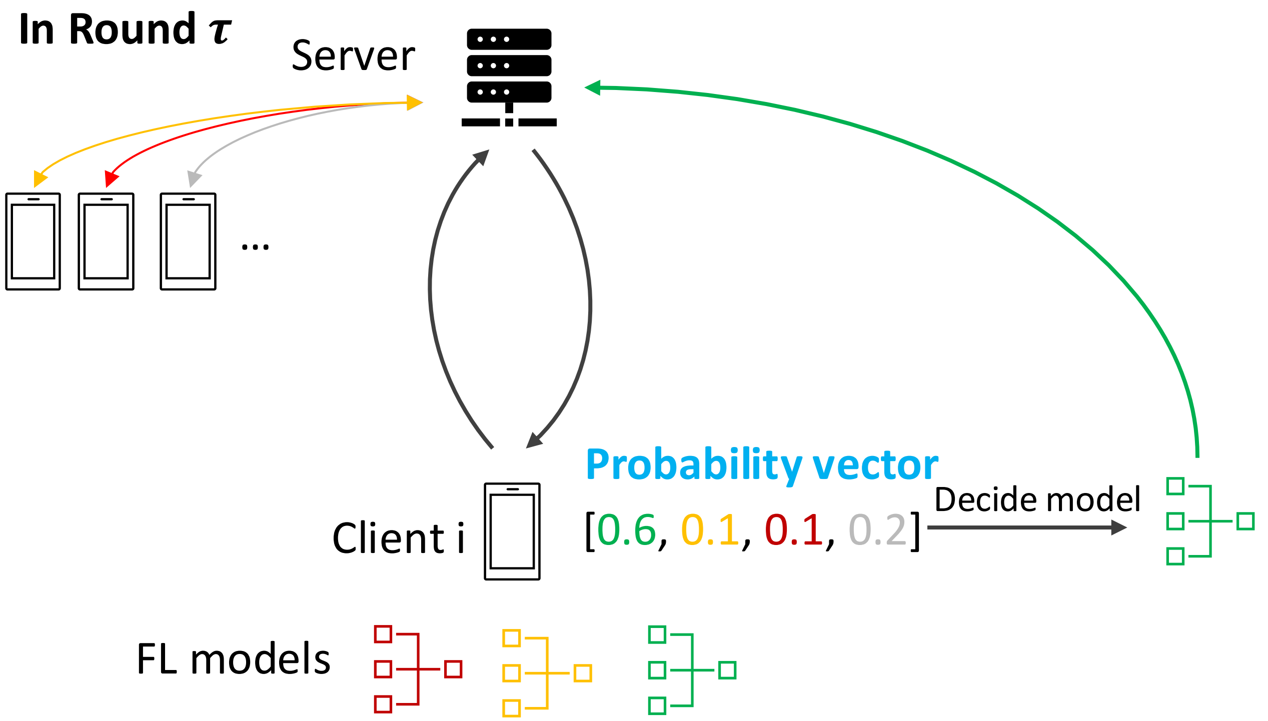
$d_{i,s} = \frac{n_{i,s}}{\sum_{j=1}^{N} n_{j,s}}$: dataset size ratio.

$U_{i,s}^{\tau} = \eta_\tau \sum_{t=1}^{K} \nabla f_{i,s}^{t,\tau}$: local update.

$p_{s|i}^{\tau}$: probability of assigning client i to model s.

$\mathcal{A}_{\tau,s}$: set of assigned clients for model s.

MMFL Optimal Variance-Reduced Sampling



In each global round (Aggregation):

$$w_s^{\tau+1} = w_s^{\tau} - \sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p_{s|i}^{\tau}} U_{i,s}^{\tau}$$

Unbiased Training:

$$\mathbb{E}\left[\sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p_{s|i}^{\tau}} U_{i,s}^{\tau}\right]$$

$$= \mathbb{E}\left[\sum_{i=1}^{N} \frac{d_{i,s}}{p_{s|i}^{\tau}} U_{i,s}^{\tau} 1_{i \in \mathcal{A}_{\tau,s}}\right]$$

$$= \sum_{i=1}^{N} d_{i,s} U_{i,s}^{\tau}$$

MMFL optimal variance-reduced sampling

Aggregation:

$$w_s^{\tau+1} = w_s^{\tau} - \boxed{\sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p_{s|i}^{\tau}} U_{i,s}^{\tau}}$$

Random Variable X

$\mathbb{E}[X]$ is given.

MMFL optimal variance-reduced sampling

Aggregation:

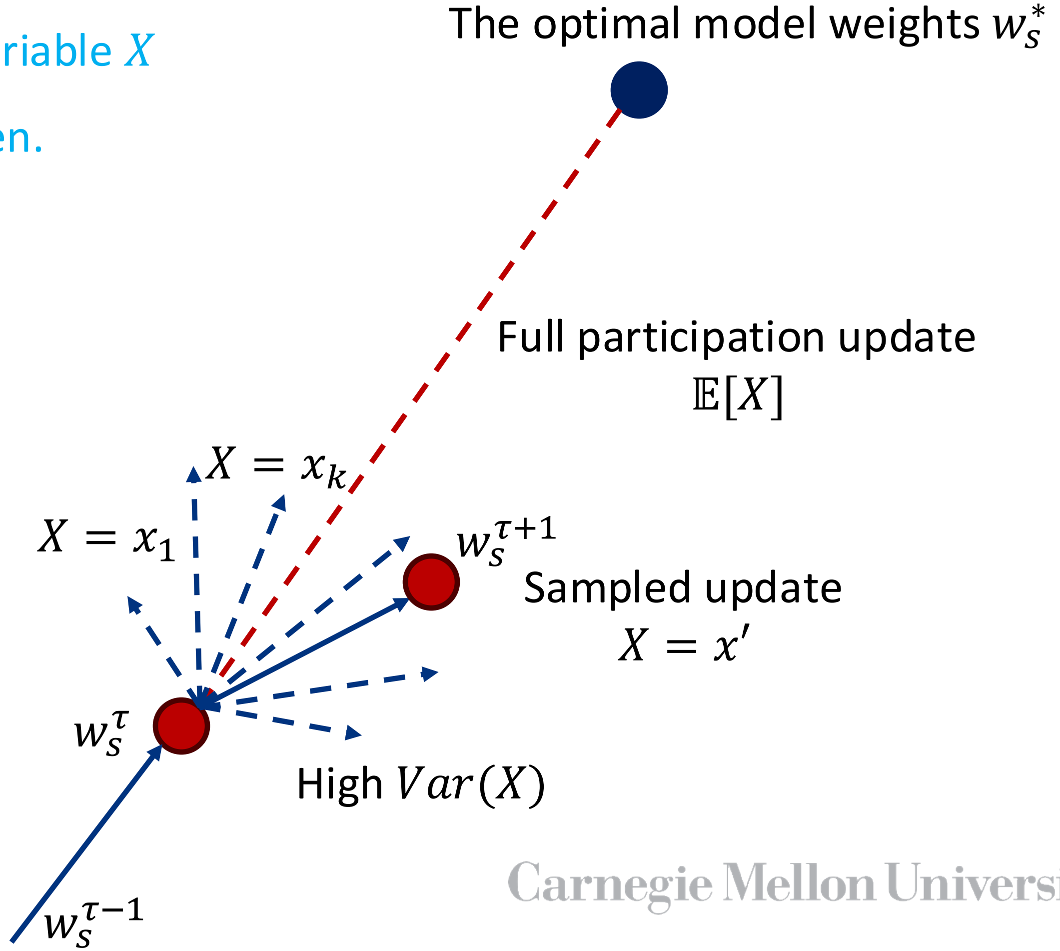
$$w_s^{\tau+1} = w_s^{\tau} - \boxed{\sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p_{s|i}^{\tau}} U_{i,s}^{\tau}}$$

Random Variable X

$\mathbb{E}[X]$ is given.

High variance of X can make the training unstable…
Therefore, define our objective:

$$\min_{\{p_{s|i}^{\tau}\}} \sum_{s=1}^{S} \mathbb{E}_{\mathcal{A}_{\tau,s}} \left[\| \sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p_{s|i}^{\tau}} U_{i,s}^{\tau} - \sum_{i=1}^{N} d_{i,s} U_{i,s}^{\tau} \|^2 \right]$$



MMFL optimal variance-reduced sampling

Aggregation:

$$w_s^{\tau+1} = w_s^{\tau} - \sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p_{s|i}^{\tau}} U_{i,s}^{\tau}$$

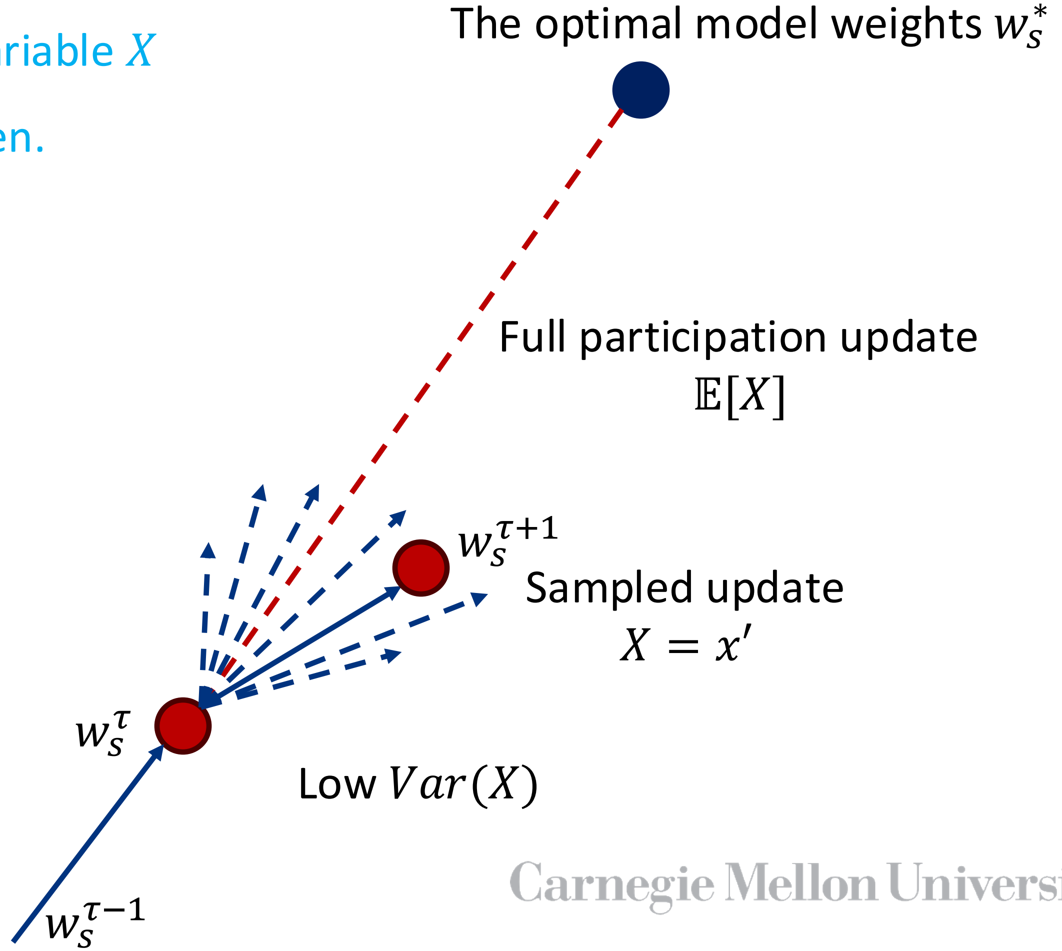
Random Variable X

$\mathbb{E}[X]$ is given.

High variance of X can make the training unstable…
Therefore, define our objective:

$$\min_{\{p_{s|i}^{\tau}\}} \sum_{s=1}^{S} \mathbb{E}_{\mathcal{A}_{\tau,s}} \left[\| \sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p_{s|i}^{\tau}} U_{i,s}^{\tau} - \sum_{i=1}^{N} d_{i,s} U_{i,s}^{\tau} \|^2 \right]$$

Notice: variance is an ideal objective to stabilize the training, but there could be other factors…
(will further discuss later)



MMFL Optimal Variance-Reduced Sampling

Minimizing the variance of update

$$\min_{\{p^{\tau}_{s|i}\}} \sum_{s=1}^{S} \mathbb{E}_{\mathcal{A}_{\tau,s}} \left[\| \sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p^{\tau}_{s|i}} U^{\tau}_{i,s} - \sum_{i=1}^{N} d_{i,s} U^{\tau}_{i,s} \|^2 \right]$$

$$\text{s.t. } p^{\tau}_{s|i} \geq 0, \ \sum_{s=1}^{S} p^{\tau}_{s|i} \leq 1, \ \sum_{s=1}^{S}\sum_{i=1}^{N} p^{\tau}_{s|i} = m \quad \forall i, s$$

τ: global round number
i: client index
s: model index
m: expected number of active clients
$d_{i,s}$: dataset size ratio
t: local epoch number
$\mathcal{A}_{\tau,s}$: set of active clients

MMFL Optimal Variance-Reduced Sampling

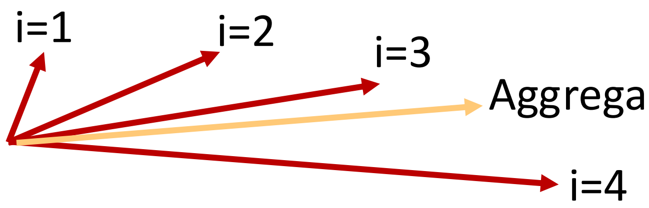
Closed-form solution of the problem

$$p_{s|i}^{\tau} = \begin{cases} (m - N + k)\frac{\|\tilde{U}_{i,s}^{\tau}\|}{\sum_{j=1}^{k} M_j^{\tau}} & \text{if } i = 1, 2, \cdots, k, \\ \frac{\|\tilde{U}_{i,s}^{\tau}\|}{M_i^{\tau}} & \text{if } i = k+1, \cdots, N. \end{cases} \quad (5)$$

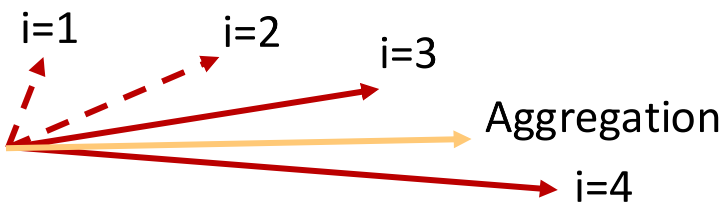
where $\|\tilde{U}_{i,s}^{\tau}\| = \|d_{i,s} U_{i,s}^{\tau}\|$ and $M_i^{\tau} = \sum_{s=1}^{S} \|\tilde{U}_{i,s}^{\tau}\|$. We reorder clients such that $M_i^{\tau} \leq M_{i+1}^{\tau}$ for all i, and k is the largest integer for which $0 < (m - N + k) \leq \frac{\sum_{j=1}^{k} M_j^{\tau}}{M_k^{\tau}}$.

τ: global round number
i: client index
s: model index
m: expected number of active clients
$d_{i,s}$: dataset size ratio
t: local epoch number
$\mathcal{A}_{\tau,s}$: set of active clients

Full participation (N=4)



Partial participation (active=2)



 Proof: https://tinyurl.com/mmflos

MMFL Optimal Variance-Reduced Sampling

Closed-form solution of the problem

$$p_{s|i}^{\tau} = \begin{cases} (m - N + k)\dfrac{\|\tilde{U}_{i,s}^{\tau}\|}{\sum_{j=1}^{k} M_j^{\tau}} & \text{if } i = 1, 2, \cdots, k, \\ \dfrac{\|\tilde{U}_{i,s}^{\tau}\|}{M_i^{\tau}} & \text{if } i = k+1, \cdots, N. \end{cases} \quad (5)$$

where $\|\tilde{U}_{i,s}^{\tau}\| = \|d_{i,s}U_{i,s}^{\tau}\|$ and $M_i^{\tau} = \sum_{s=1}^{S}\|\tilde{U}_{i,s}^{\tau}\|$. We reorder clients such that $M_i^{\tau} \leq M_{i+1}^{\tau}$ for all i, and k is the largest integer for which $0 < (m - N + k) \leq \frac{\sum_{j=1}^{k} M_j^{\tau}}{M_k^{\tau}}$.

τ: global round number
i: client index
s: model index
m: expected number of active clients
$d_{i,s}$: dataset size ratio
t: local epoch number
$\mathcal{A}_{\tau,s}$: set of active clients

Gradient-based Variance-Reduce Sampling (GVR)

Computing the gradient norm is too expensive on the client side!

Proof: https://tinyurl.com/mmflos

Reduce computational cost

Computing the gradient norm is too expensive on the client side.

Client i loss value

$$\min_{\{p^{\tau}_{s|i}\}} \sum_{s=1}^{S} \mathbb{E}_{\mathcal{A}_{\tau,s}} \left[\| \sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p^{\tau}_{s|i}} \boxed{U^{\tau}_{i,s}} - \sum_{i=1}^{N} d_{i,s} \boxed{U^{\tau}_{i,s}} \|^2 \right]$$

$$\text{s.t. } p^{\tau}_{s|i} \geq 0, \sum_{s=1}^{S} p^{\tau}_{s|i} \leq 1, \sum_{s=1}^{S} \sum_{i=1}^{N} p^{\tau}_{s|i} = m \quad \forall i, s$$

Loss-based **V**ariance-**R**educed Sampling **(LVR)**

τ: global round number
i: client index
s: model index
m: expected number of active clients
$d_{i,s}$: dataset size ratio
t: local epoch number
$\mathcal{A}_{\tau,s}$: set of active clients

Reduce computational cost

Computing the gradient norm is too expensive on the client side.

τ: global round number
i: client index
s: model index
m: expected number of active clients
$d_{i,s}$: dataset size ratio
t: local epoch number
$\mathcal{A}_{\tau,s}$: set of active clients

Client i loss value

$$\min_{\{p^{\tau}_{s|i}\}} \sum_{s=1}^{S} \mathbb{E}_{\mathcal{A}_{\tau,s}} \left[\| \sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p^{\tau}_{s|i}} \boxed{U^{\tau}_{i,s}} - \sum_{i=1}^{N} d_{i,s} \boxed{U^{\tau}_{i,s}} \|^2 \right]$$

$$\text{s.t. } p^{\tau}_{s|i} \geq 0, \sum_{s=1}^{S} p^{\tau}_{s|i} \leq 1, \sum_{s=1}^{S} \sum_{i=1}^{N} p^{\tau}_{s|i} = m \quad \forall i, s$$

Now we have two methods to optimize the sampling distribution.
Can we analyze their influence on convergence speed?

Convergence proof

Based on some common assumptions (L-smoothness, mu-strongly convex, etc.)
We modified and adapted the proof from [2].

Theorem 4 (Convergence). *Let w_s^* denote the optimal weights of model s. If the learning rate $\eta_\tau = \frac{16}{\mu}\frac{1}{(\tau+1)K+\gamma}$, then*

$$\mathbb{E}\left(\|w_s^\tau - w_s^*\|^2\right) \leq \frac{V_\tau}{(\tau K + \gamma_\tau)^2} \tag{413}$$

Here we define $\gamma_\tau = \max\{\frac{32L}{\mu}, 4K\sum_{i\in\mathcal{N}_s} \mathbb{1}_i^{s,\tau} P_{i,s}^\tau\}$
$V_\tau = \max\{\gamma_\tau^2 \mathbb{E}(\|w_s^0 - w_s^*\|^2), (\frac{16}{\mu})^2 \sum_{\tau'=0}^{\tau-1} z_{\tau'}\}$,
$z_{\tau'} = \mathbb{E}[Z_g^{\tau'} + Z_l^{\tau'} + Z_p^{\tau'}]$,
$\mathbb{E}[Z_g^\tau] = K\sum_{i\in\mathcal{N}_s} \frac{(d_{i,s}\sigma_{i,s})^2}{p_{s|i}^\tau} + 4LK\sum_{i\in\mathcal{N}_s} d_{i,s}\Gamma_{i,s} + \max(\frac{1}{d_{i,s}})\mathbb{E}[\sum_{i\in\mathcal{N}_s} \frac{(d_{i,s})^2 \sum_{t=1}^K \|\nabla f_{i,s}(w_{i,s}^{t,\tau})\|^2}{p_{s|i}^\tau}]$,
$\mathbb{E}[Z_l^\tau] = R\mathbb{E}[|\mathcal{N}_s| \sum_{i\in\mathcal{N}_s} (\mathbb{1}_i^{s,\tau} P_{i,s}^\tau f_{i,s}(w_s^\tau) - d_{i,s} f_{i,s}(w_s^\tau))^2]$, *where* $R = \frac{2K^3\bar{\sigma}^2}{e_w^2 e_f^2 \theta}$,
$\mathbb{E}[Z_p^\tau] = (\frac{2}{\theta} + K(2 + \frac{\mu}{2L}))K^2\bar{\sigma}^2 + \frac{2K^3\bar{\sigma}^2}{\theta}\mathbb{E}[(\sum_{i\in\mathcal{N}_s} \mathbb{1}_i^{s,\tau} P_{i,s}^\tau - 1)^2]$.

[2] Ruan, Yichen, et al. "Towards flexible device participation in federated learning." *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021.

Convergence proof

Based on some common assumptions (L-smoothness, mu-strongly convex, etc.)
We modified and adapted the proof from [2].

$$\mathbb{E}[Z_g^\tau] = K \sum_{i\in\mathcal{N}_s} \frac{(d_{i,s}\sigma_{i,s})^2}{p^\tau_{s|i}} + 4LK \sum_{i\in\mathcal{N}_s} d_{i,s}\Gamma_{i,s} + \max\left(\frac{1}{d_{i,s}}\right)\boxed{\mathbb{E}\left[\sum_{i\in\mathcal{N}_s} \frac{(d_{i,s})^2 \sum_{t=1}^{K} \|\nabla f_{i,s}(w^{t,\tau}_{i,s})\|^2}{p^\tau_{s|i}}\right]},$$

$$\mathbb{E}[Z_l^\tau] = R\mathbb{E}[|\mathcal{N}_s| \sum_{i\in\mathcal{N}_s} (\mathbb{1}_i^{s,\tau} P^\tau_{i,s} f_{i,s}(w^\tau_s) - d_{i,s} f_{i,s}(w^\tau_s))^2],\ where\ R = \frac{2K^3\bar{\sigma}^2}{e_w^2 e_f^2 \theta},$$

$$\mathbb{E}[Z_p^\tau] = \left(\frac{2}{\theta} + K\left(2 + \frac{\mu}{2L}\right)\right)K^2\bar{\sigma}^2 + \frac{2K^3\bar{\sigma}^2}{\theta}\mathbb{E}[(\sum_{i\in\mathcal{N}_s} \mathbb{1}_i^{s,\tau} P^\tau_{i,s} - 1)^2].$$

$\mathbb{E}[Z_g^\tau]$ -> Sampled update variance (GVR)

In the proof: https://tinyurl.com/mmflos

From the upper bound to variance term:
$\left\|\sum_{t=1}^{K} \nabla f_{i,s}\right\|^2 \le K \sum_{t=1}^{K} \left\|\nabla f_{i,s}\right\|^2$ (GM-HM inequality)

$$= \boxed{\sum_{s=1}^{S}\left[\mathbb{E}\left[\left\|\sum_{i\in\mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p^\tau_{s|i}} U^\tau_{i,s}\right\|^2\right] - \left\|\sum_{i=1}^{N} d_{i,s} U^\tau_{i,s}\right\|^2\right]} \quad (9)$$

$$= \sum_{s=1}^{S}\left[\mathbb{E}\left[\sum_{i,j} \frac{d_{i,s}(U^\tau_{i,s})^\top}{p^\tau_{s|i}} \frac{d_{j,s}U^\tau_{j,s}}{p^\tau_{s|j}} \mathbb{1}_{i,j\in\mathcal{A}_{\tau,s}}\right] - \sum_{i,j} d_{i,s}d_{j,s}(U^\tau_{i,s})^\top U^\tau_{j,s}\right] \quad (10)$$

$$= \sum_{s=1}^{S}\left[\sum_{i\neq j} d_{i,s}(U^\tau_{i,s})^\top d_{j,s}U^\tau_{j,s} + \sum_{i=1}^{N} \frac{d^2_{i,s}(U^\tau_{i,s})^\top U^\tau_{i,s}}{p^\tau_{s|i}} - \sum_{i,j} d_{i,s}d_{j,s}(U^\tau_{i,s})^\top U^\tau_{j,s}\right] \quad (11)$$

$$= \sum_{s=1}^{S}\left(\sum_{i=1}^{N}\left(\frac{\|d_{i,s}U^\tau_{i,s}\|^2}{p^\tau_{s|i}} - \|d_{i,s}U^\tau_{i,s}\|^2\right)\right) \quad (12)$$

$$= \sum_{s=1}^{S}\boxed{\sum_{i=1}^{N} \frac{\|d_{i,s}U^\tau_{i,s}\|^2}{p^\tau_{s|i}}} - \sum_{s=1}^{S}\sum_{i=1}^{N} \|d_{i,s}U^\tau_{i,s}\|^2 \quad (13)$$

Convergence proof

Based on some common assumptions (L-smoothness, mu-strongly convex, etc.)
We modified and adapted the proof from [2].

$$\mathbb{E}[Z_g^\tau] = K\sum_{i\in\mathcal{N}_s}\frac{(d_{i,s}\sigma_{i,s})^2}{p_{s|i}^\tau} + 4LK\sum_{i\in\mathcal{N}_s} d_{i,s}\Gamma_{i,s} + \max(\frac{1}{d_{i,s}})\mathbb{E}[\sum_{i\in\mathcal{N}_s}\frac{(d_{i,s})^2\sum_{t=1}^{K}\|\nabla f_{i,s}(w_{i,s}^{t,\tau})\|^2}{p_{s|i}^\tau}],$$

$$\mathbb{E}[Z_l^\tau] = R\mathbb{E}[|\mathcal{N}_s|\boxed{\sum_{i\in\mathcal{N}_s}(\mathbb{1}_i^{s,\tau}P_{i,s}^\tau f_{i,s}(w_s^\tau) - d_{i,s}f_{i,s}(w_s^\tau))^2}], \textit{ where } R = \frac{2K^3\bar{\sigma}^2}{e_w^2 e_f^2\theta},$$

$$\mathbb{E}[Z_p^\tau] = (\frac{2}{\theta} + K(2+\frac{\mu}{2L}))K^2\bar{\sigma}^2 + \frac{2K^3\bar{\sigma}^2}{\theta}\mathbb{E}[(\sum_{i\in\mathcal{N}_s}\mathbb{1}_i^{s,\tau}P_{i,s}^\tau - 1)^2].$$

$\mathbb{E}[Z_l^\tau]$ -> Sampled loss variance (LVR), with similar GM-HM inequality.

Client i loss value

$$\min_{\{p_{s|i}^\tau\}}\sum_{s=1}^{S}\mathbb{E}_{\mathcal{A}_{\tau,s}}\left[\|\sum_{i\in\mathcal{A}_{\tau,s}}\frac{d_{i,s}}{p_{s|i}^\tau}U_{i,s}^\tau - \sum_{i=1}^{N}d_{i,s}U_{i,s}^\tau\|^2\right]$$

$$\text{s.t. } p_{s|i}^\tau \geq 0, \sum_{s=1}^{S}p_{s|i}^\tau \leq 1, \sum_{s=1}^{S}\sum_{i=1}^{N}p_{s|i}^\tau = m \quad \forall i, s$$

Convergence proof

Based on some common assumptions (L-smoothness, mu-strongly convex, etc.)
We modified and adapted the proof from [2].

$$\mathbb{E}[Z_g^\tau] = K\sum_{i\in\mathcal{N}_s}\frac{(d_{i,s}\sigma_{i,s})^2}{p_{s|i}^\tau}+4LK\sum_{i\in\mathcal{N}_s}d_{i,s}\Gamma_{i,s}+\max(\frac{1}{d_{i,s}})\mathbb{E}[\sum_{i\in\mathcal{N}_s}\frac{(d_{i,s})^2\sum_{t=1}^K\|\nabla f_{i,s}(w_{i,s}^{t,\tau})\|^2}{p_{s|i}^\tau}],$$

$$\mathbb{E}[Z_l^\tau] = R\mathbb{E}[|\mathcal{N}_s|\sum_{i\in\mathcal{N}_s}(\mathbb{1}_i^{s,\tau}P_{i,s}^\tau f_{i,s}(w_s^\tau) - d_{i,s}f_{i,s}(w_s^\tau))^2],\ where\ R = \frac{2K^3\bar{\sigma}^2}{e_w^2 e_f^2\theta},$$

$$\mathbb{E}[Z_p^\tau] = (\frac{2}{\theta}+K(2+\frac{\mu}{2L}))K^2\bar{\sigma}^2+\frac{2K^3\bar{\sigma}^2}{\theta}\boxed{\mathbb{E}[(\sum_{i\in\mathcal{N}_s}\mathbb{1}_i^{s,\tau}P_{i,s}^\tau-1)^2]}.$$

$$P_{i,s}^\tau = \frac{d_{i,s}}{p_{s|i}^\tau}$$

$\mathbb{E}[Z_p^\tau]$ -> Participation heterogeneity (or variance).

$\boxed{\text{The red term}}$ is only related to dataset distribution and sampling distribution.

What is the meaning of this term?

Convergence proof

Based on some common assumptions (L-smoothness, mu-strongly convex, etc.)
We modified and adapted the proof from [2].

$$\mathbb{E}[Z_p^\tau] = (\tfrac{2}{\theta} + K(2 + \tfrac{\mu}{2L}))K^2\bar{\sigma}^2 + \tfrac{2K^3\bar{\sigma}^2}{\theta}\boxed{\mathbb{E}[(\textstyle\sum_{i\in\mathcal{N}_s} \mathbb{1}_i^{s,\tau} P_{i,s}^\tau - 1)^2]}.$$

$$P_{i,s}^\tau = \frac{d_{i,s}}{p_{s|i}^\tau}$$

$\mathbb{E}[Z_p^\tau]$ -> Participation heterogeneity (or variance)

Recall our global aggregation rule:

$$w_s^{\tau+1} = w_s^\tau - \sum_{i\in\mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p_{s|i}^\tau} U_{i,s}^\tau$$

Can be rewritten as:

$$w_s^{\tau+1} = w_s^\tau - (H_s^\tau)^\top U_s^\tau$$

$$H_s^\tau = [\cdots, 1_i^{s,\tau} P_{i,s}^\tau, \cdots]^\top, U_s^\tau = [\cdots, U_{i,s}^\tau, \cdots]$$

Convergence proof

Based on some common assumptions (L-smoothness, mu-strongly convex, etc.)
We modified and adapted the proof from [2].

$$\mathbb{E}[Z_p^\tau] = (\tfrac{2}{\theta} + K(2 + \tfrac{\mu}{2L}))K^2\bar{\sigma}^2 + \tfrac{2K^3\bar{\sigma}^2}{\theta}\boxed{\mathbb{E}[(\textstyle\sum_{i\in\mathcal{N}_s} \mathbb{1}_i^{s,\tau} P_{i,s}^\tau - 1)^2]}.$$

$$P_{i,s}^\tau = \frac{d_{i,s}}{p_{s|i}^\tau}$$

$\mathbb{E}[Z_p^\tau]$ -> Participation heterogeneity (or variance)

Recall our global aggregation rule:

$$w_s^{\tau+1} = w_s^\tau - \sum_{i\in\mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p_{s|i}^\tau} U_{i,s}^\tau$$

Can be rewritten as:

$$w_s^{\tau+1} = w_s^\tau - (H_s^\tau)^\top U_s^\tau$$

$$H_s^\tau = [\cdots, 1_i^{s,\tau} P_{i,s}^\tau, \cdots]^\top, U_s^\tau = [\cdots, U_{i,s}^\tau, \cdots]$$

$$|H_s^\tau|_1 = \sum_{i=1}^{N} 1_i^{s,\tau} P_{i,s}^\tau = \sum_{i=1}^{N} 1_i^{s,\tau} \frac{d_{i,s}}{p_{s|i}^\tau}$$

Notice $\mathbb{E}[|H_s^\tau|_1] = 1$, therefore

$$\boxed{\text{red term}} = \mathbb{E}[(|H_s^\tau|_1 - 1)^2]$$

This is also a variance!

How does this variance influence the training?

The influence of participation heterogeneity

$$|H_s^\tau|_1 = \sum_{i=1}^{N} 1_i^{s,\tau} P_{i,s}^\tau = \sum_{i=1}^{N} 1_i^{s,\tau} \frac{d_{i,s}}{p_{s|i}^\tau}$$

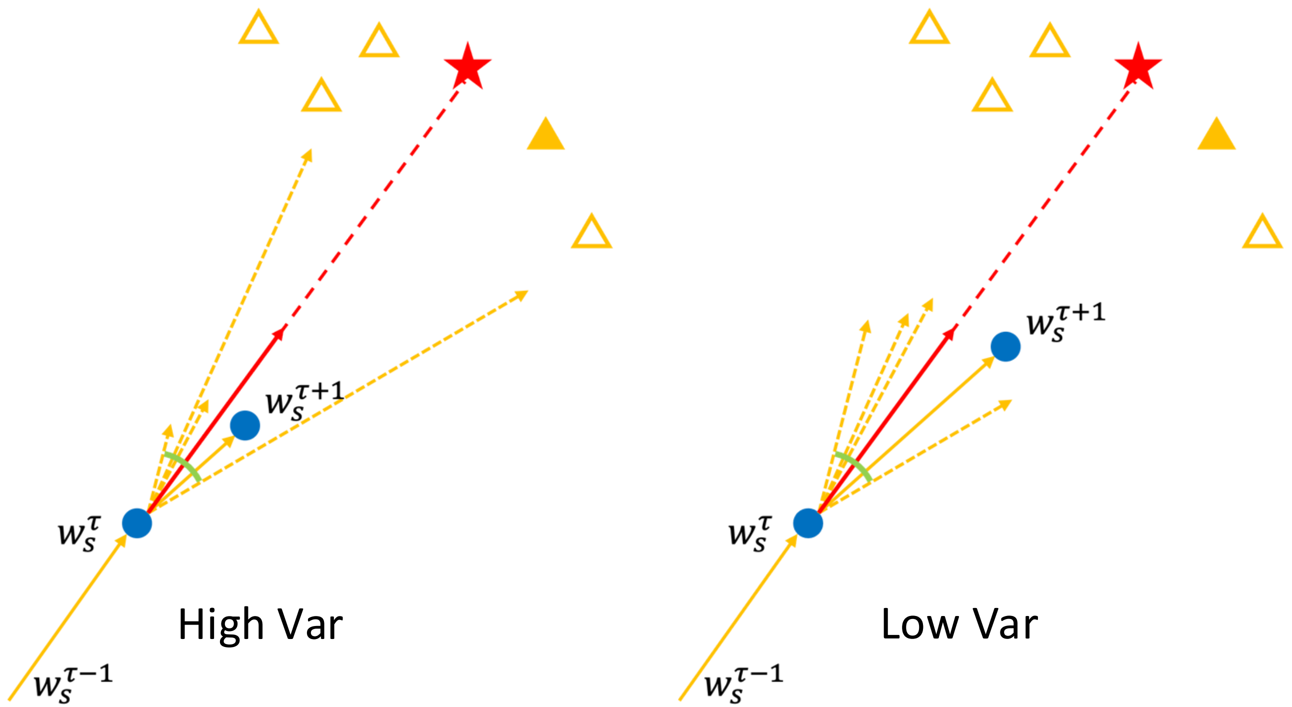
$$Var_H = \mathbb{E}[(|H_s^\tau|_1 - 1)^2]$$

High Var_H: $|H_s^\tau|_1$ may change a lot across rounds.

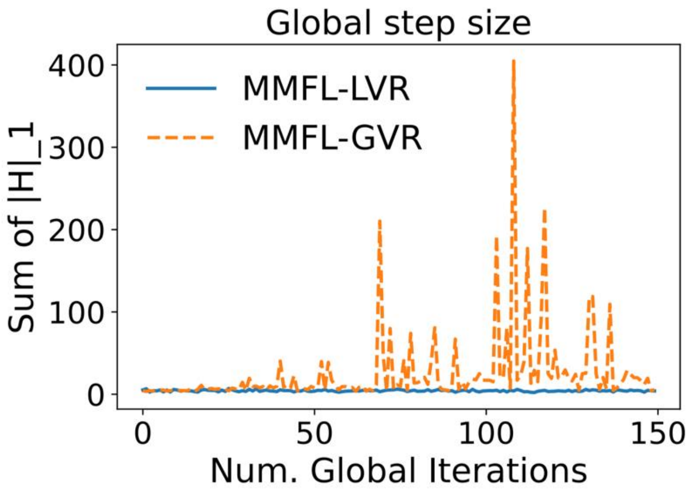
Lead to unstable "global step."

$$w_s^{\tau+1} = w_s^\tau - (H_s^\tau)^\top U_s^\tau$$

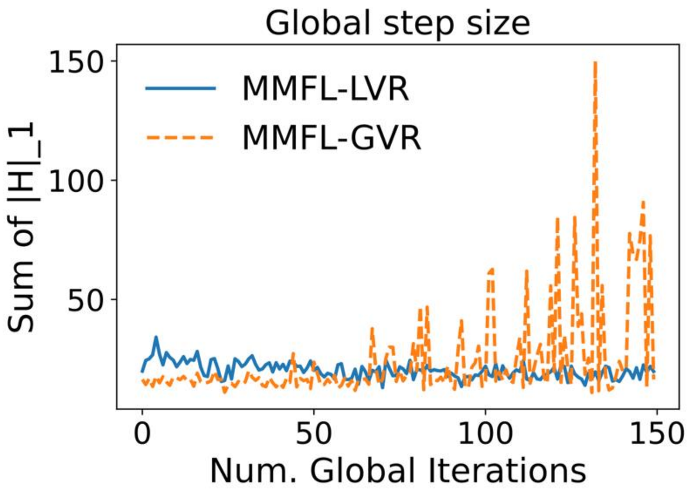
Impact the training especially at the end stage of the training.



Compare GVR and LVR



3 models



5 models

$$w_s^{\tau+1} = w_s^{\tau} - (H_s^{\tau})^{\top} U_s^{\tau}$$

How to mitigate the impact of unstable “global step?”

Mitigate the impact of participation heterogeneity

Previous Aggregation Rule:

$$|H_s^\tau|_1 = \sum_{i=1}^{N} 1_i^{s,\tau} P_{i,s}^\tau = \sum_{i=1}^{N} 1_i^{s,\tau} \frac{d_{i,s}}{p_{s|i}^\tau}$$

$$w_s^{\tau+1} = w_s^\tau - (H_s^\tau)^\top \, U_s^\tau$$

New Aggregation Rule [3]:

$$w_s^{\tau+1} = w_s^\tau - (\sum_{i=1}^{N} d_{i,s} h_{i,s}^\tau + \sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}\left(U_{i,s}^\tau - h_{i,s}^\tau\right)}{p_{s|i}^\tau})$$

$$h_{i,s}^\tau = \begin{cases} U_{i,s}^{\tau-1}, if \ i \in \mathcal{A}_{\tau-1,s} \\ h_{i,s}^{\tau-1}, if \ i \in \mathcal{A}_{\tau-1,s} \end{cases}$$

$U_{i,s}^\tau - h_{i,s}^\tau$ should be small. Even though $|H_s^\tau|_1$ has a high variance, the impact is small.

Server stores stale updates from clients, and use stale updates to stabilize the training. **GVR***

[3] Jhunjhunwala, Divyansh, et al. "Fedvarp: Tackling the variance due to partial client participation in federated learning." *Uncertainty in Artificial Intelligence*. PMLR, 2022.

Outline

- Introduction
  - Federated learning (FL) ✓
  - Multi-model federated learning (MMFL) ✓
- Variance-reduced client sampling in a simple MMFL system ✓
- Modeling computational heterogeneity in MMFL
- Experiments

Recall

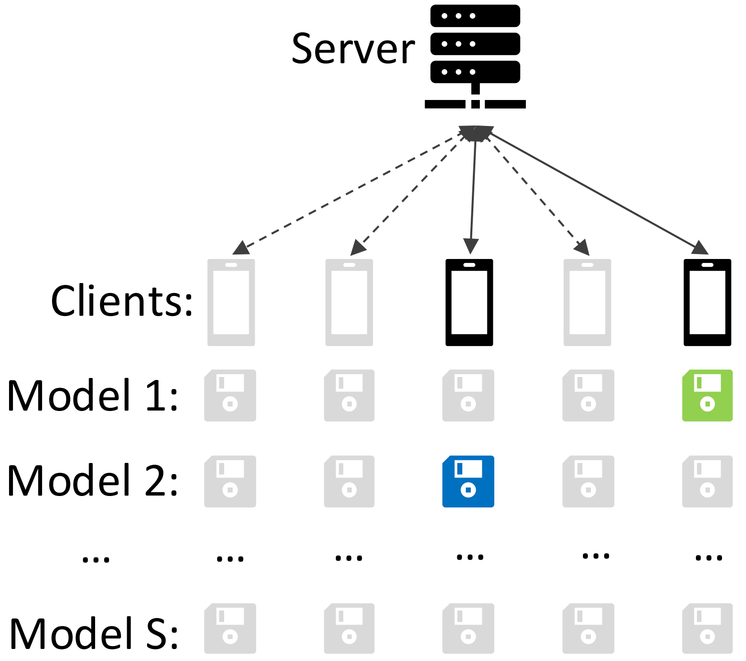
Key assumptions from previous work [1]

In each round, the server only allows partial participation, and each active client can only train one model.

1) Partial Participation: reduce communication cost

2) Only train one model: computational constraints

"Only train one model" is too ideal, without considering heterogeneity of computational abilities.



Multi-model federated learning

[1] Bhuyan, Neelkamal, Sharayu Moharir, and Gauri Joshi. "Multi-model federated learning with provable guarantees." *EAI International Conference on Performance Evaluation Methodologies and Tools*. Cham: Springer Nature Switzerland, 2022.

Multi-Model Federated Learning

Make more realistic assumptions

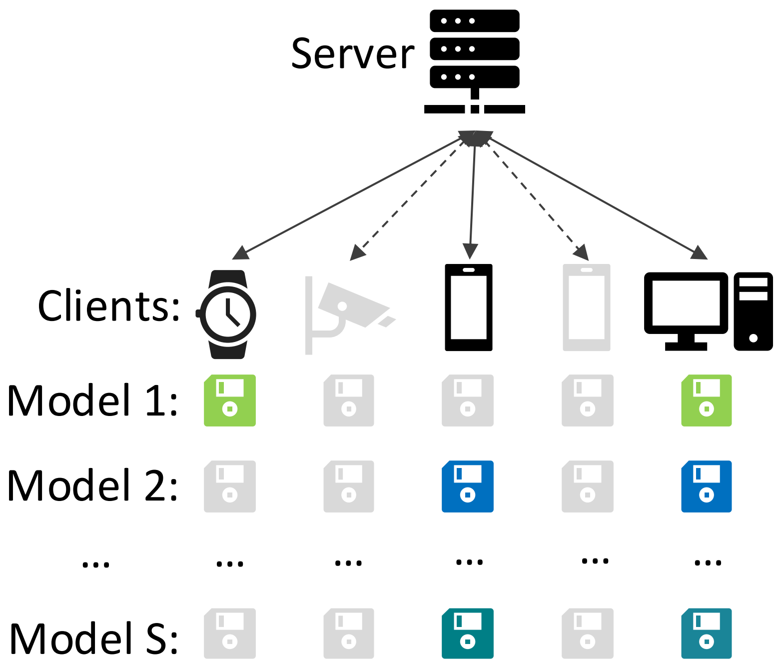
In each round, the server only allows partial participation, and each active client i can train B_i models in parallel.

1) Partial Participation: reduce communication cost

2) Client i can train B_i models ($B_i \leq S$):

Computational constraint & heterogeneity

"Powerful" clients train more models, leading to biased convergence. How to achieve unbiased training?



Multi-model federated learning

System model for heterogeneous MMFL

For ease of description, <u>assume client i has B_i processors</u>, each processor (i, b) can train one model independently.

1) Adjust the aggregation rule to ensure unbiased training

$$w_s^{\tau+1} = w_s^\tau - \sum_{(i,b)\in\mathcal{A}_{\tau,s}} P_{(i,b),s}^\tau G_{(i,b),s}^\tau$$

$$P_{(i,b),s}^\tau = \frac{d_{i,s}}{B_i p_{s|(i,b)}^\tau}, \qquad G_{(i,b),s}^\tau = \eta_\tau \sum_{t=1}^{K} \nabla f_{i,s}^{t,\tau}$$

Notations:
w_s^τ: global model parameters
$\mathcal{A}_{\tau,s}$: set of active “processors”
$d_{i,s}$: dataset size ratio
$p_{s|(i,b)}^\tau$: the probability of having processor (i, b) to train model s
τ: global round index
t: local epoch index

System model for heterogeneous MMFL

For ease of description, <u>assume client i has B_i processors</u>, each processor (i,b) can train one model independently.

1) Adjust the aggregation rule to ensure unbiased training

$$w_s^{\tau+1} = w_s^{\tau} - \sum_{(i,b)\in\mathcal{A}_{\tau,s}} P_{(i,b),s}^{\tau} G_{(i,b),s}^{\tau}$$

$$\mathbb{E}\left[\sum_{i=1}^{N}\sum_{b=1}^{B_i} 1_{(i,b),s}^{\tau} \frac{d_{i,s}}{B_i p_{s|(i,b)}} G_{(i,b),s}^{\tau}\right] = \sum_{i=1}^{N} d_{i,s}\mathbb{E}\left[G_{(i,b),s}^{\tau}\right]$$

Sampling at the "processor-level"

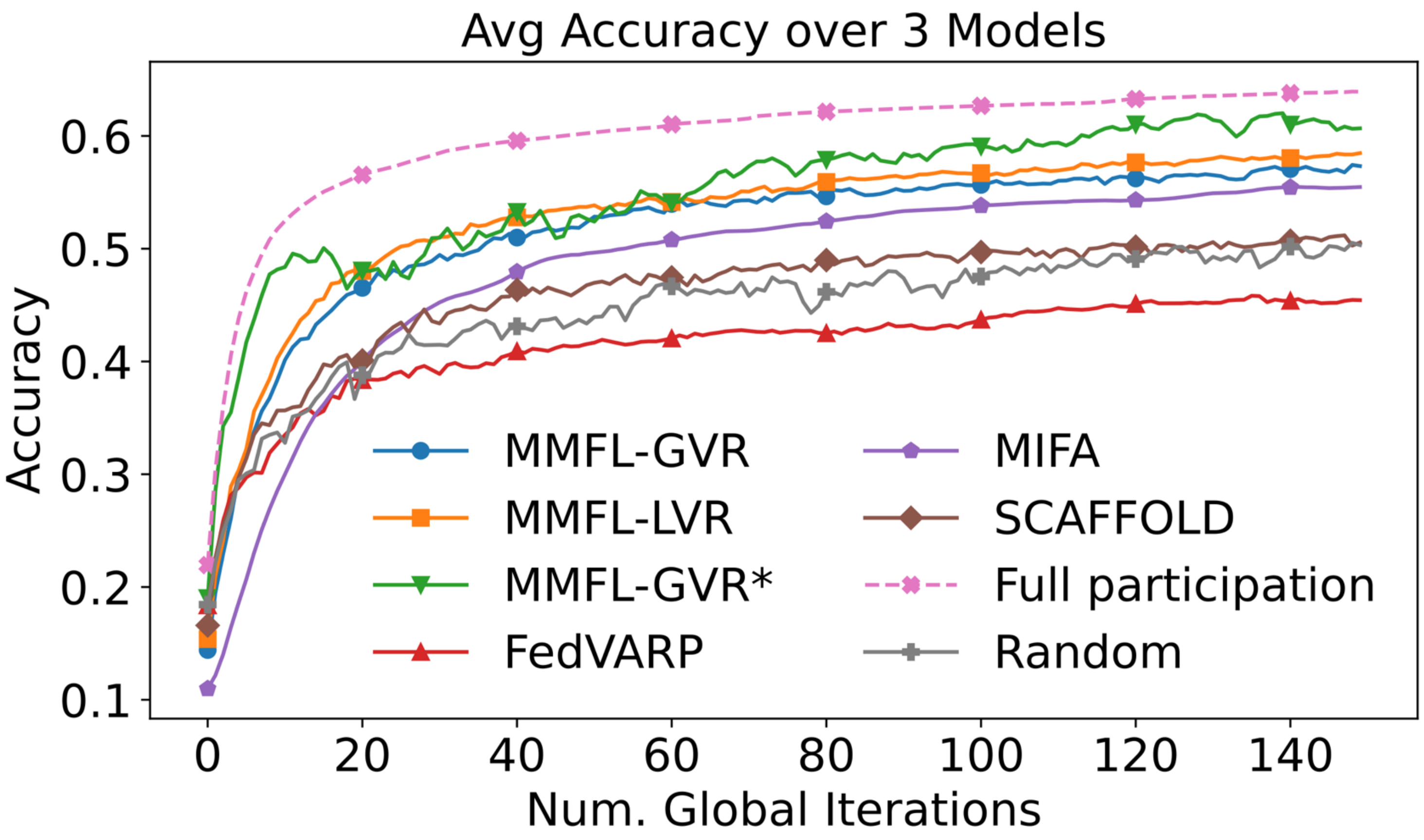
Notations:
w_s^{τ}: global model parameters
$\mathcal{A}_{\tau,s}$: set of active "processors"
$d_{i,s}$: dataset size ratio
$p_{s|(i,b)}^{\tau}$: the probability of having processor (i,b) to train model s
τ: global round index
t: local epoch index

Experiments

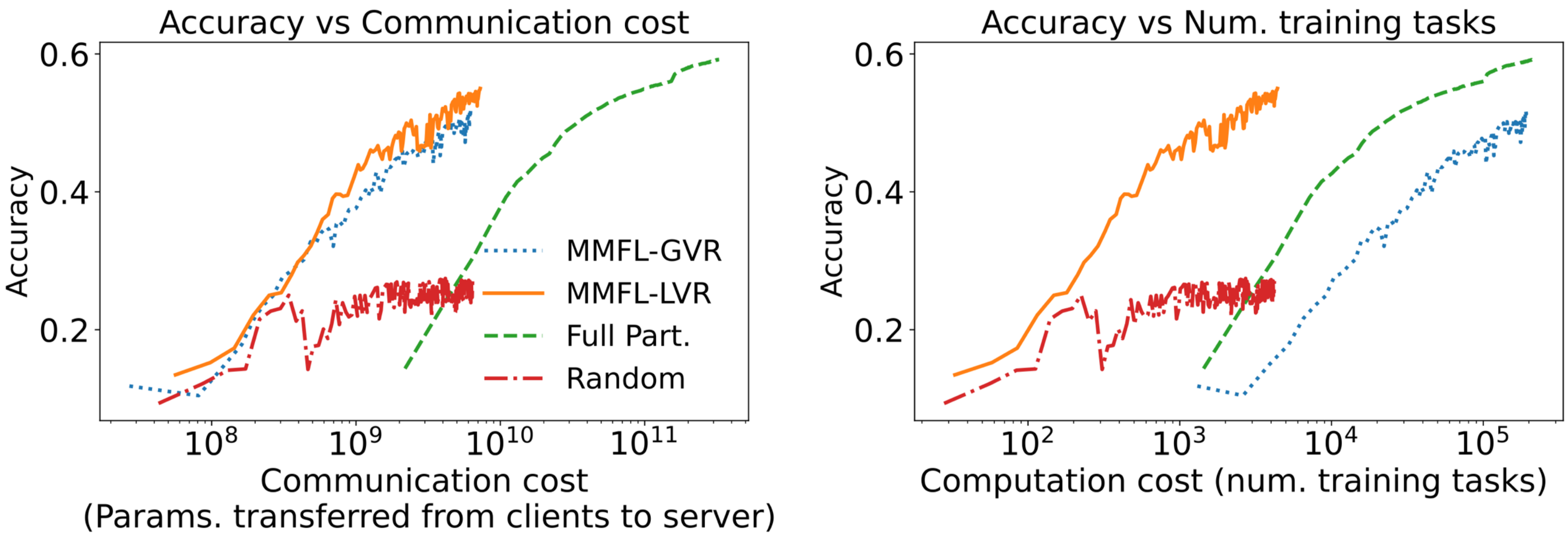
3 Models: all Fashion-MNIST.
N=120 clients
m=12 (active rate=0.1)
Each client: 30% labels.

For each model: 10% high-data clients, 90% low-data clients. 10% clients hold 52.6% data of each task.

25% clients: $B_i = 3$
50% clients: $B_i = 2$
25% clients: $B_i = 1$



Experiments



Experiments

3 Models: all Fashion-MNIST.

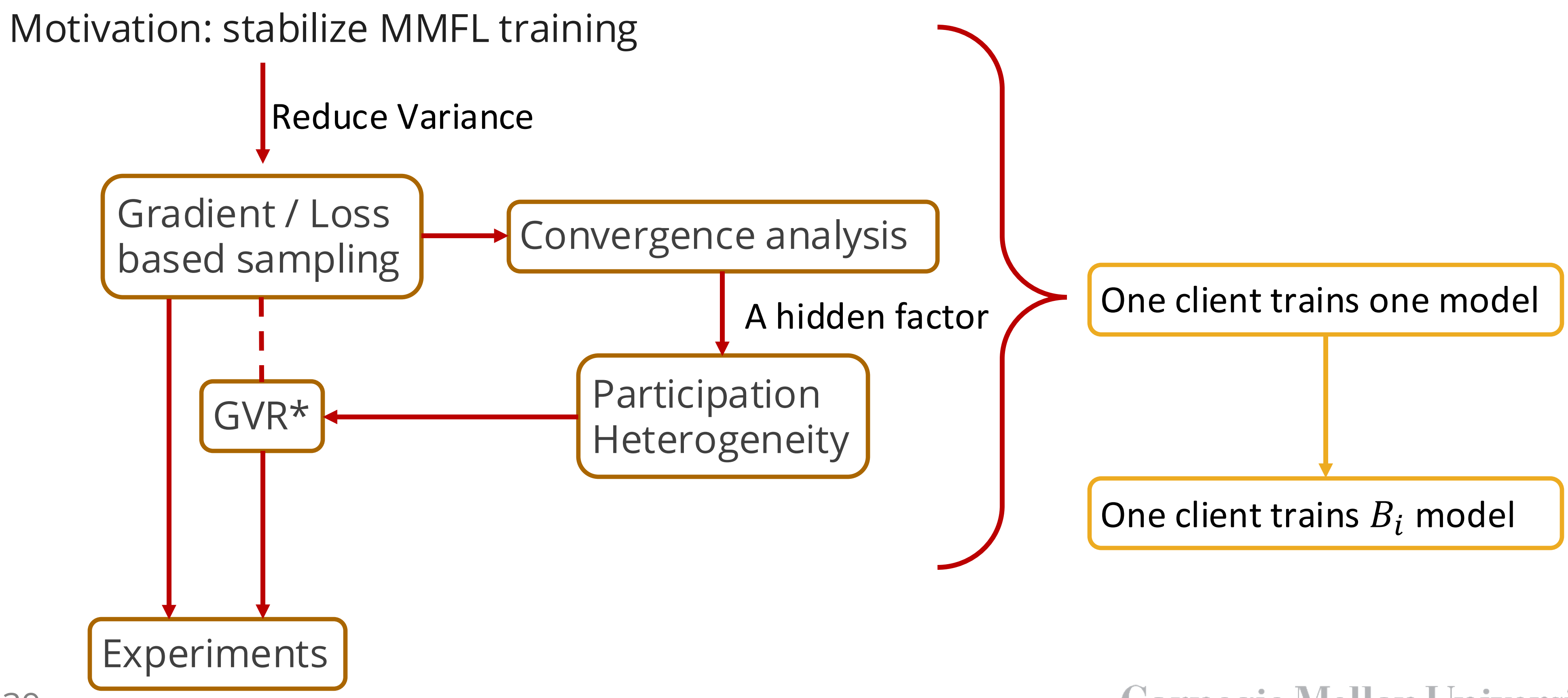
5 Models: two Fashion-MNIST, one CIFAR-10, one EMNIST, one Shakespeare.

10% clients only have data for S-1 models.

TABLE I
FINAL AVERAGE MODEL ACCURACY RELATIVE TO THAT FROM FULL PARTICIPATION (THEORETICALLY THE BEST UNDER THE SAME LOCAL TRAINING SETTINGS).

Methods	3 tasks	5 tasks	Comm. Cost	Comp. Cost	Mem. Cost
FedVARP [30]	$0.712 \pm .14$	$0.690 \pm .19$	Low	Low	High
MIFA [31]	$0.868 \pm .18$	$0.835 \pm .18$	Low	Low	High
SCAFFOLD [32]	$0.794 \pm .14$	$0.650 \pm .24$	Low	Low	Low
Random	$0.778 \pm .19$	$0.749 \pm .23$	Low	Low	Low
Full Participation	$1.000 \pm .13$	$1.000 \pm .14$	High	High	Low
MMFL-GVR	$0.893 \pm .14$	$0.842 \pm .20$	Low	High	Low
MMFL-LVR	$\underline{0.912} \pm .15$	$\underline{0.849} \pm .16$	<u>Low</u>	<u>Low</u>	<u>Low</u>
MMFL-GVR*	$\mathbf{0.960} \pm .15$	$\mathbf{0.869} \pm .18$	Low	High	High

Summary


Motivation: stabilize MMFL training
Reduce Variance
Gradient / Loss based sampling
Convergence analysis
A hidden factor
Participation Heterogeneity
GVR*
Experiments
One client trains one model
One client trains B_i model

Multi-Model Federated Learning

Make more realistic assumptions

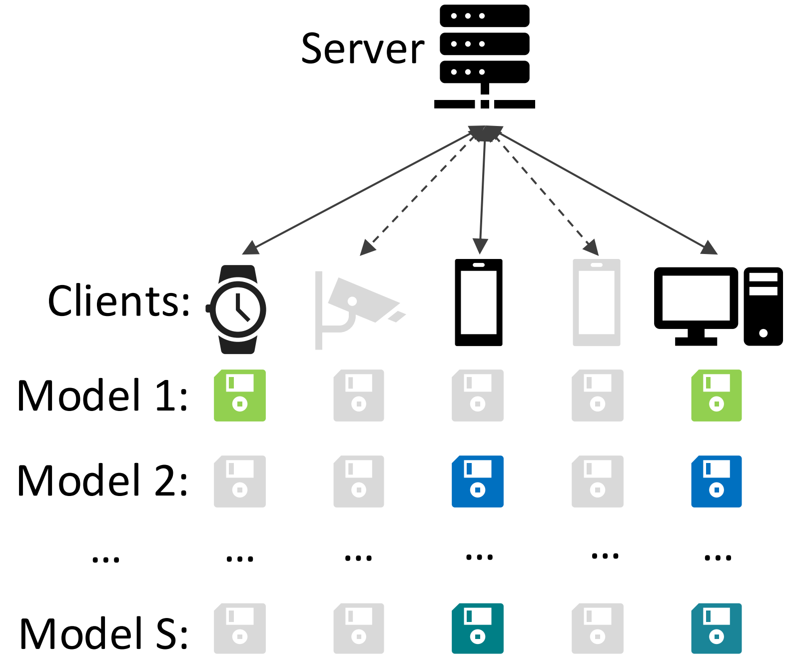
In each round, the server only allows partial participation, and each active client i can train B_i models in parallel.

Other ways to model computational heterogeneity:

1) Asynchronous training [4]

2) Flexible local epochs number [5]

3) Flexible model architectures [6]



Multi-model federated learning

[4] Askin, Baris, et al. "FedAST: Federated Asynchronous Simultaneous Training."
[5] Ruan, Yichen, et al. "Towards flexible device participation in federated learning."
[6] Park, Jong-Ik, and Carlee Joe-Wong. "Federated Learning with Flexible Architectures."