Carnegie Mellon University

Optimal Variance-Reduced Client Sampling in Multiple Models Federated Learning

MAY 17, 2024

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

Outline

- Introduction
  - Federated Learning (FL)
  - Multi-Model Federated Learning (MMFL)
- Optimal variance-reduced client sampling in MMFL
- Experiments
- Future directions about this work

Federated Learning

Decentralized learning with unshared local data

Local client (device):

1 Get global model parameters

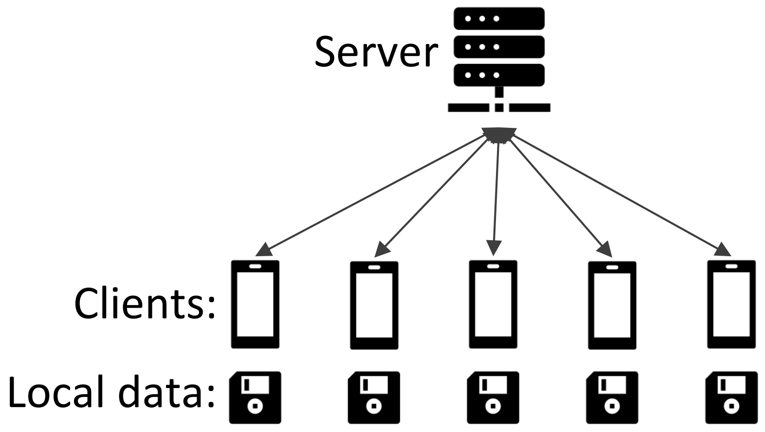
2 Train model parameters with local private data

3 Send updated parameters to the server

Server:

1 Receive updates from clients

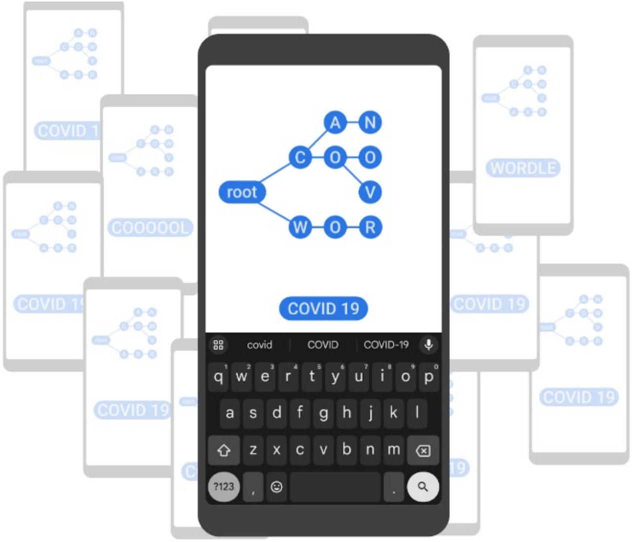
2 Aggregate local updates for a better global model



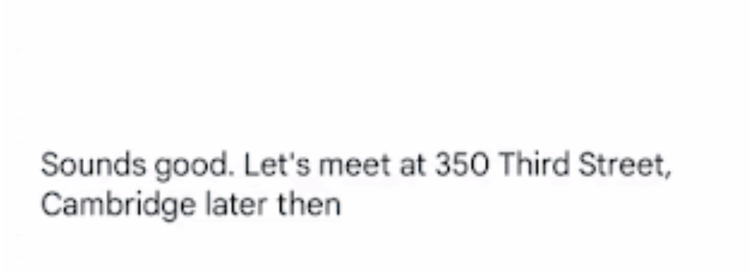
Multi-Model Federated Learning

(Google example) Multiple FL models are running on your phone.

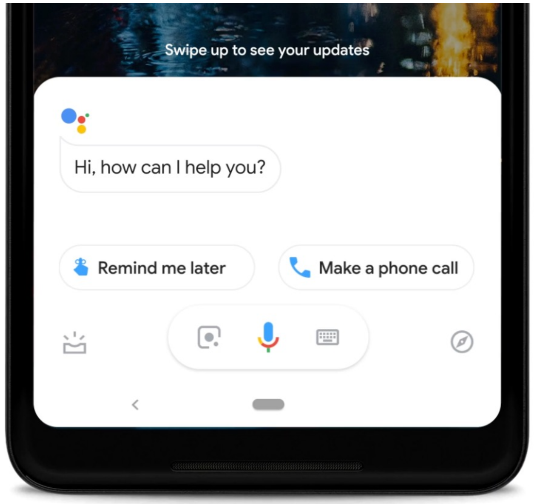
Keyboard prediction



Predicting text selection

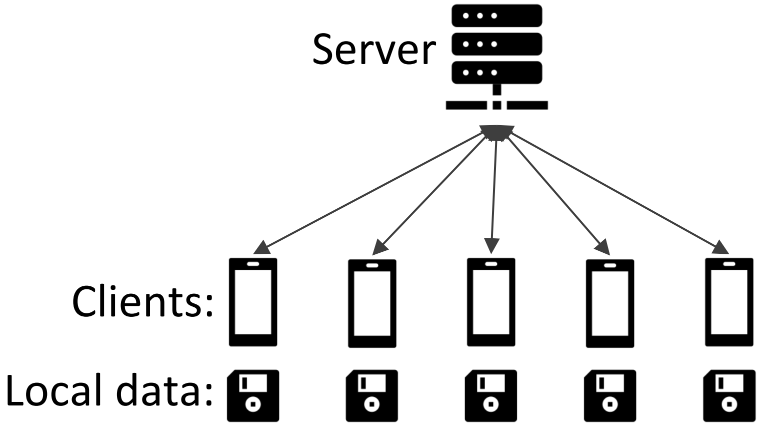


Speech model

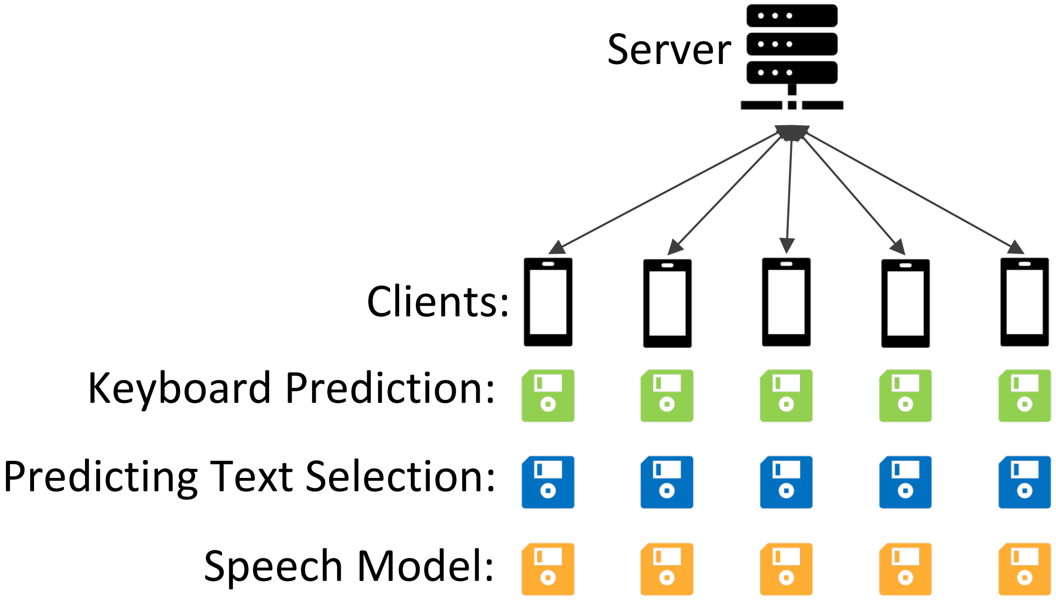


 Source: federated.withgoogle.com

Multi-Model Federated Learning



Single-model federated learning

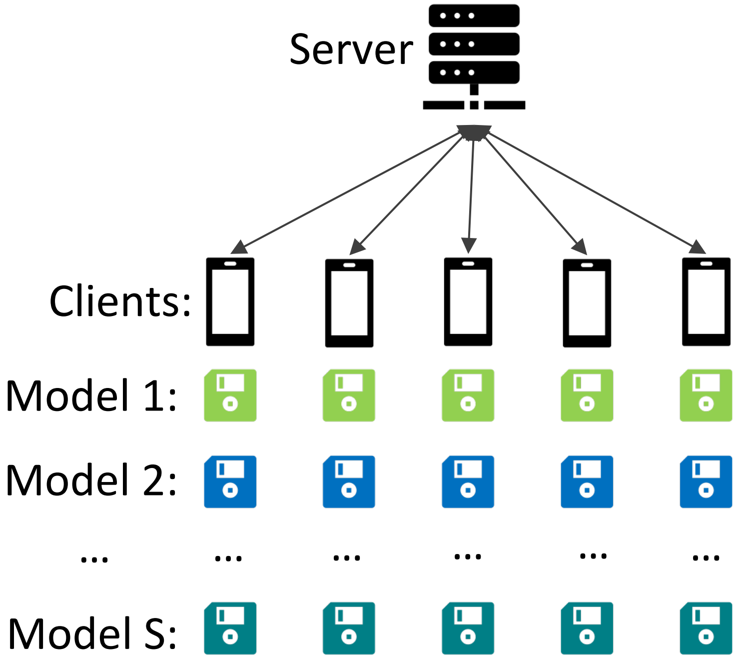


Multi-model federated learning

Multi-Model Federated Learning

Problems we may have

Server communication cost



Multi-model federated learning

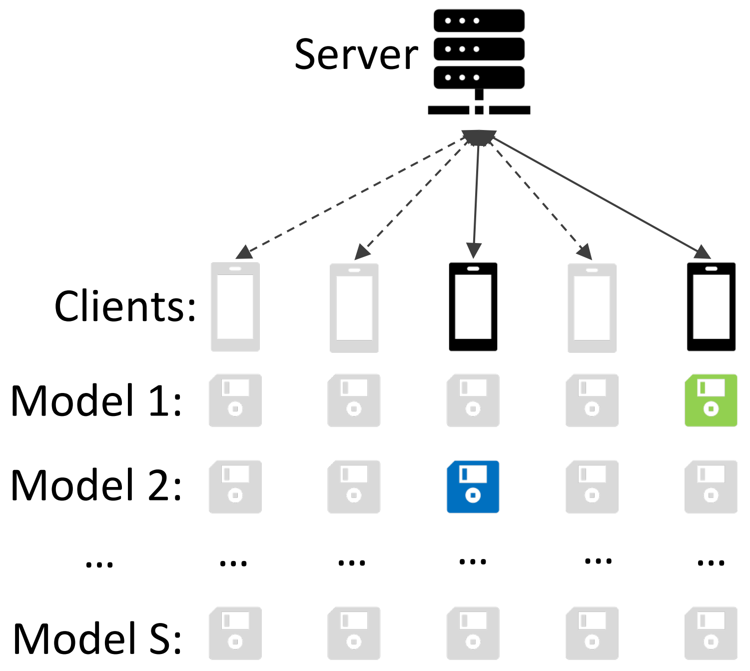
Multi-Model Federated Learning

Problems we may have

Server communication cost

Assumptions:

1. Partial client participation/communication
2. Each client can only send 1 model to the server



Multi-model federated learning

Multi-Model Federated Learning

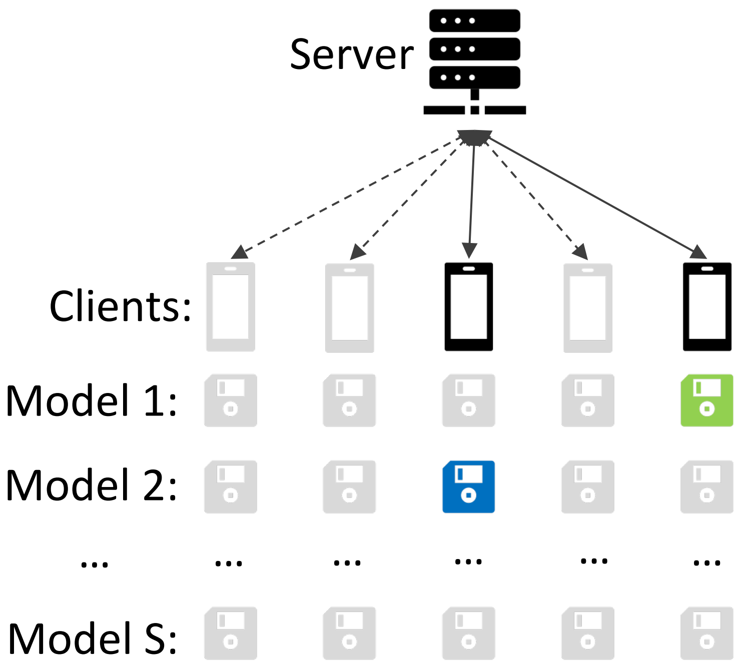
Problems we may have

Server communication cost

Assumptions:

1. Partial client participation/communication
2. Each client can only send 1 model to the server

Questions:
How to sample clients? How to sample models?

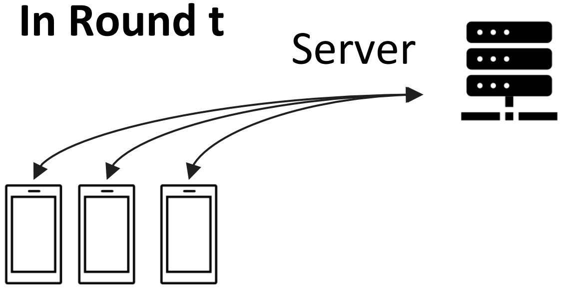


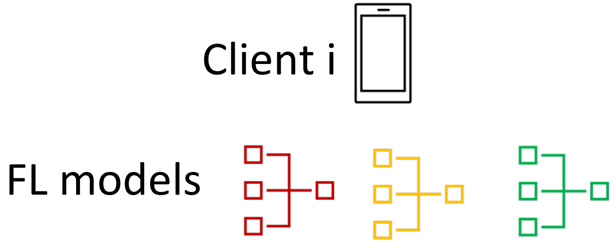
Multi-model federated learning

MMFL Optimal Variance-Reduced Sampling

Idea: the client with higher gradient norm can provide more informative updates.

In Round t

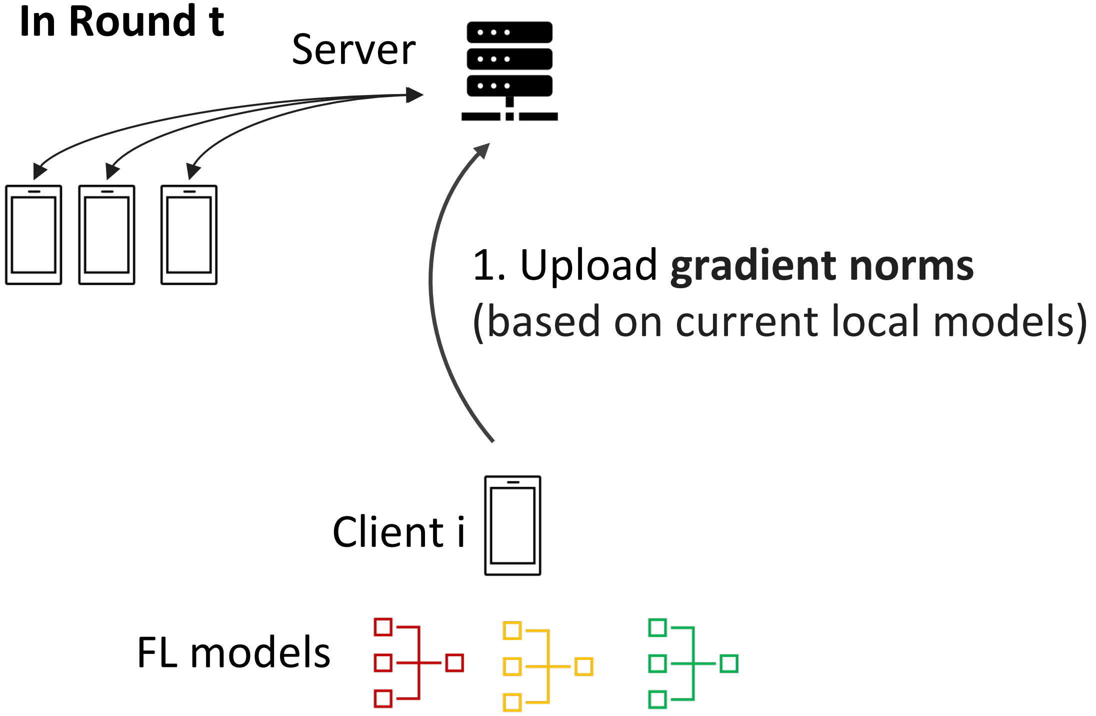




MMFL Optimal Variance-Reduced Sampling

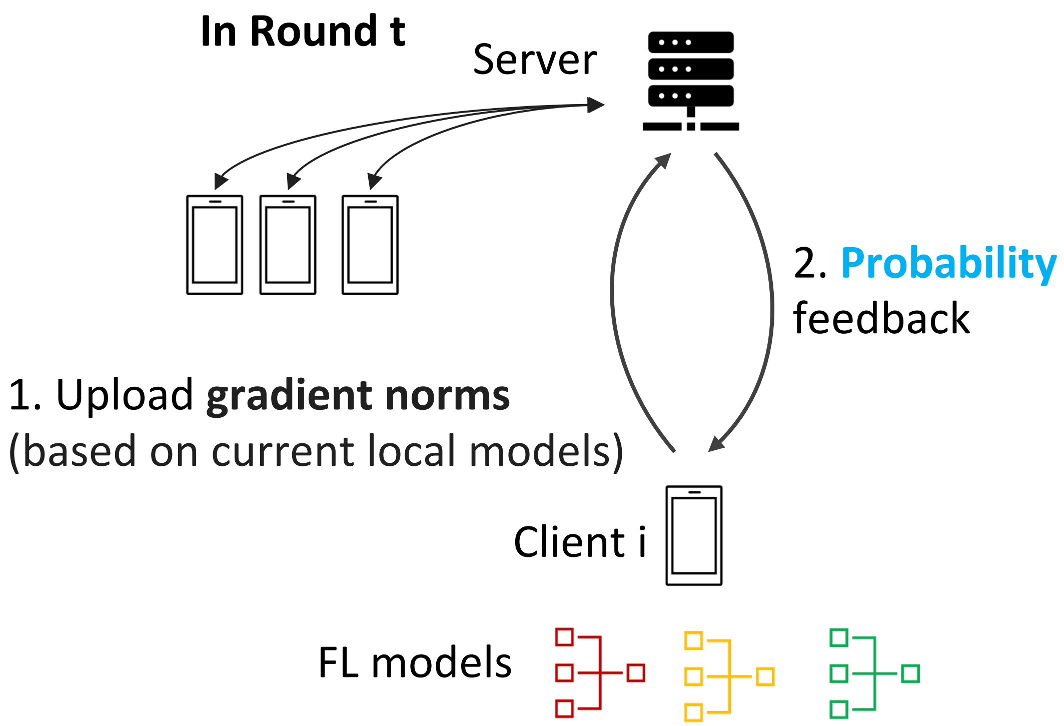
Idea: the client with higher gradient norm can provide more informative updates.

In Round t



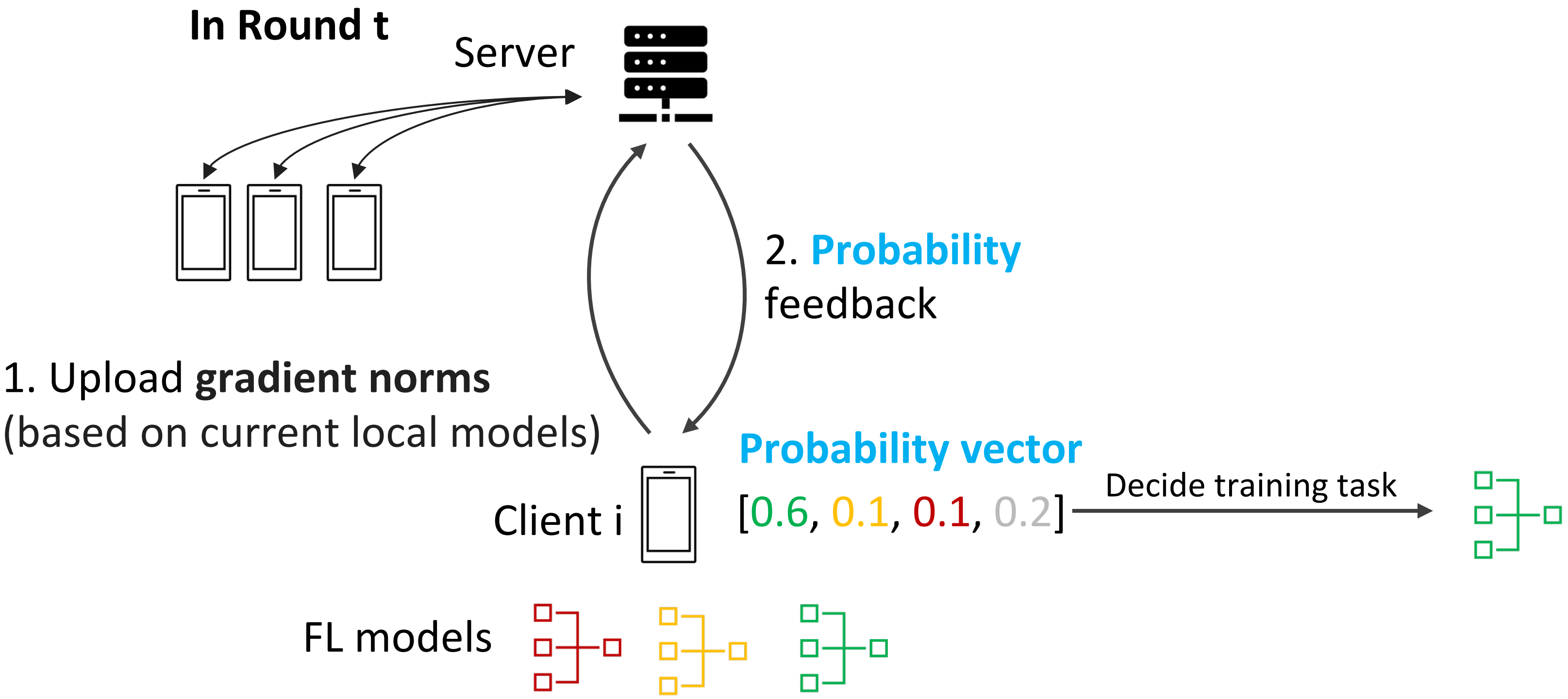
MMFL Optimal Variance-Reduced Sampling

Idea: the client with higher gradient norm can provide more informative updates.



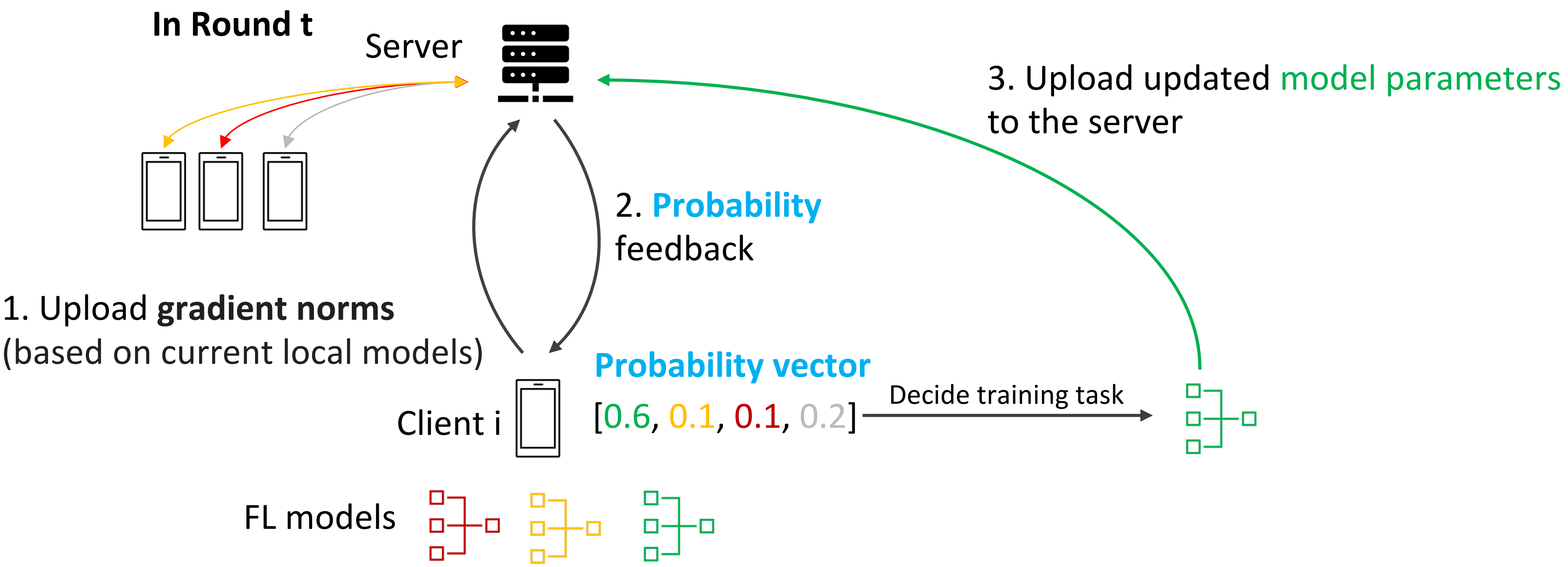
MMFL Optimal Variance-Reduced Sampling

Idea: the client with higher gradient norm can provide more informative updates.



MMFL Optimal Variance-Reduced Sampling

Idea: the client with higher gradient norm can provide more informative updates.



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

MMFL Optimal Variance-Reduced Sampling

Minimizing the variance of update

τ: global round number
i: client index
s: model index
m: expected number of active clients
$d_{i,s}$: dataset size ratio

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

Sampling Probability distribution

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

MMFL Optimal Variance-Reduced Sampling

Minimizing the variance of update

$U_{i,s}^{\tau} = \sum_{t=1}^{E} \nabla f_{i,s}(w_{i,s,\tau}^{t})$

Client i update (gradient)

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

Sampling Probability distribution

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

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

MMFL Optimal Variance-Reduced Sampling

Minimizing the variance of update

$U^\tau_{i,s} = \sum_{t=1}^{E} \nabla f_{i,s}(w^t_{i,s,\tau})$
Client i update (gradient)

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

Sampling Probability distribution

Full participation update

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

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

MMFL Optimal Variance-Reduced Sampling

Minimizing the variance of update

$U^{\tau}_{i,s} = \sum_{t=1}^{E} \nabla f_{i,s}(w^{t}_{i,s,\tau})$
Client i update (gradient)

Sampled update

$$\min_{\underbrace{\{p^{\tau}_{s|i}\}}_{\text{Sampling Probability distribution}}} \sum_{s=1}^{S} \mathbb{E}_{\mathcal{A}_{\tau,s}} \left[\| \underbrace{\sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p^{\tau}_{s|i}} U^{\tau}_{i,s}}_{\text{Sampled update}} - \underbrace{\sum_{i=1}^{N} d_{i,s} U^{\tau}_{i,s}}_{\text{Full participation update}} \|^2 \right]$$

Sampling Probability distribution

Full participation update

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

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

Carnegie Mellon University

MMFL Optimal Variance-Reduced Sampling

Minimizing the variance of update

$U_{i,s}^{\tau} = \sum_{t=1}^{E} \nabla f_{i,s}(w_{i,s,\tau}^{t})$
Client i update (gradient)

Sampled update

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

Sampling Probability distribution

Full participation update

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

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

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

MMFL Optimal Variance-Reduced Sampling

Closed-form solution of the problem

$U_{i,s}^{\tau} = \sum_{t=1}^{E} \nabla f_{i,s}(w_{i,s,\tau}^{t})$

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

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

Proof: https://tinyurl.com/mmflos

MMFL Optimal Variance-Reduced Sampling

τ: global round number
i: client index
s: model index
m: expected number of active clients
$d_{i,s}$: dataset ratio
t: local epoch number

Algorithm 1 MMFL optimal variance-reduced sampling

```
Input: expected active client number m, clients: 1, 2, 3, ··· , N, models: 1, ··· , S,
learning rate η_τ
Initialization: The global model weights w_s^1 for each model
for global round τ = 1, ··· , T do
    for each client i = 1, ··· , N, in parallel do
        for each model s = 1, ··· , S do
            w_{i,s}^1 = w_s^τ
            for local epochs t = 1, ··· , E do
                w_{i,s}^{t+1} = w_{i,s}^t − η_τ ∇f_i(w_{i,s}^t)
            end for
            record weights difference U_{i,s} = Σ_{t=1}^E ∇f_i(w_{i,s}^t)
            Send ||U_{i,s}|| to the server
        end for
    end for
    At server:
    Receive ||U_{i,s}|| for all clients and all models
    p_{s|i}=ClientSampling({||U_{i,s}||}) using the closed-form solution
    Select clients by p_{s|i}, obtain the set of active clients for each model: A_{τ,s}
    for each model s = 1, ··· , S, in parallel do
        Request U_{i,s} from clients i ∈ A_{τ,s}
        Server aggregation: w_s^{τ+1} = w_s^τ − η_τ Σ_{i∈A_{τ,s}} (d_{i,s}/p_{s|i}) U_{i,s}
        Broadcast w_s^{τ+1} to clients
    end for
end for
```

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

Minimize its variance

MMFL optimal variance-reduced sampling

$$\{p_{s|i}^{\tau}\} \rightarrow \mathcal{A}_{\tau,1}, \cdots, \mathcal{A}_{\tau,S} \rightarrow G_1^{\tau}, \cdots, G_s^{\tau}, \cdots, G_S^{\tau}$$

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

MMFL optimal variance-reduced sampling

Sampling probability distribution

$\{p^\tau_{s|i}\} \to \mathcal{A}_{\tau,1}, \cdots, \mathcal{A}_{\tau,S} \to G^\tau_1, \cdots, G^\tau_s, \cdots, G^\tau_S$

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

MMFL optimal variance-reduced sampling

Sampling probability distribution

$$\{p_{s|i}^{\tau}\} \rightarrow \mathcal{A}_{\tau,1}, \cdots, \mathcal{A}_{\tau,S} \rightarrow G_1^{\tau}, \cdots, G_s^{\tau}, \cdots, G_S^{\tau}$$

Set of active clients for each model

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

MMFL optimal variance-reduced sampling

Sampling probability distribution

Sampled update for each model

$$\{p_{s|i}^{\tau}\} \rightarrow \mathcal{A}_{\tau,1}, \cdots, \mathcal{A}_{\tau,S} \rightarrow G_1^{\tau}, \cdots, G_s^{\tau}, \cdots, G_S^{\tau}$$

Set of active clients for each model

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

MMFL optimal variance-reduced sampling

Sampling probability distribution

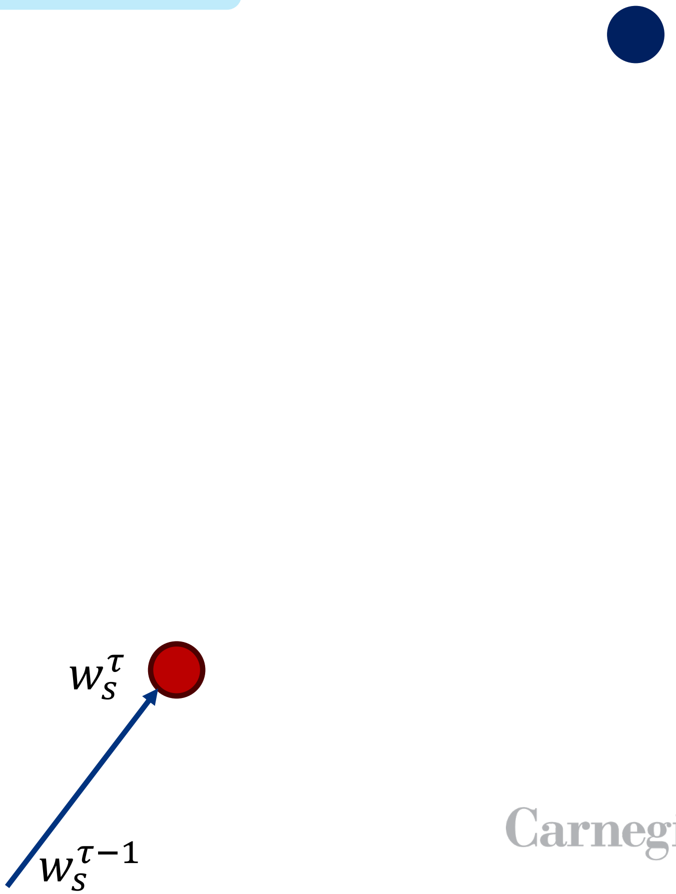
Sampled update for each model

$\{p_{s|i}^{\tau}\} \rightarrow \mathcal{A}_{\tau,1}, \cdots, \mathcal{A}_{\tau,S} \rightarrow G_1^{\tau}, \cdots, G_s^{\tau}, \cdots, G_S^{\tau}$

Set of active clients for each model

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

The optimal model weights w_s^*



MMFL optimal variance-reduced sampling

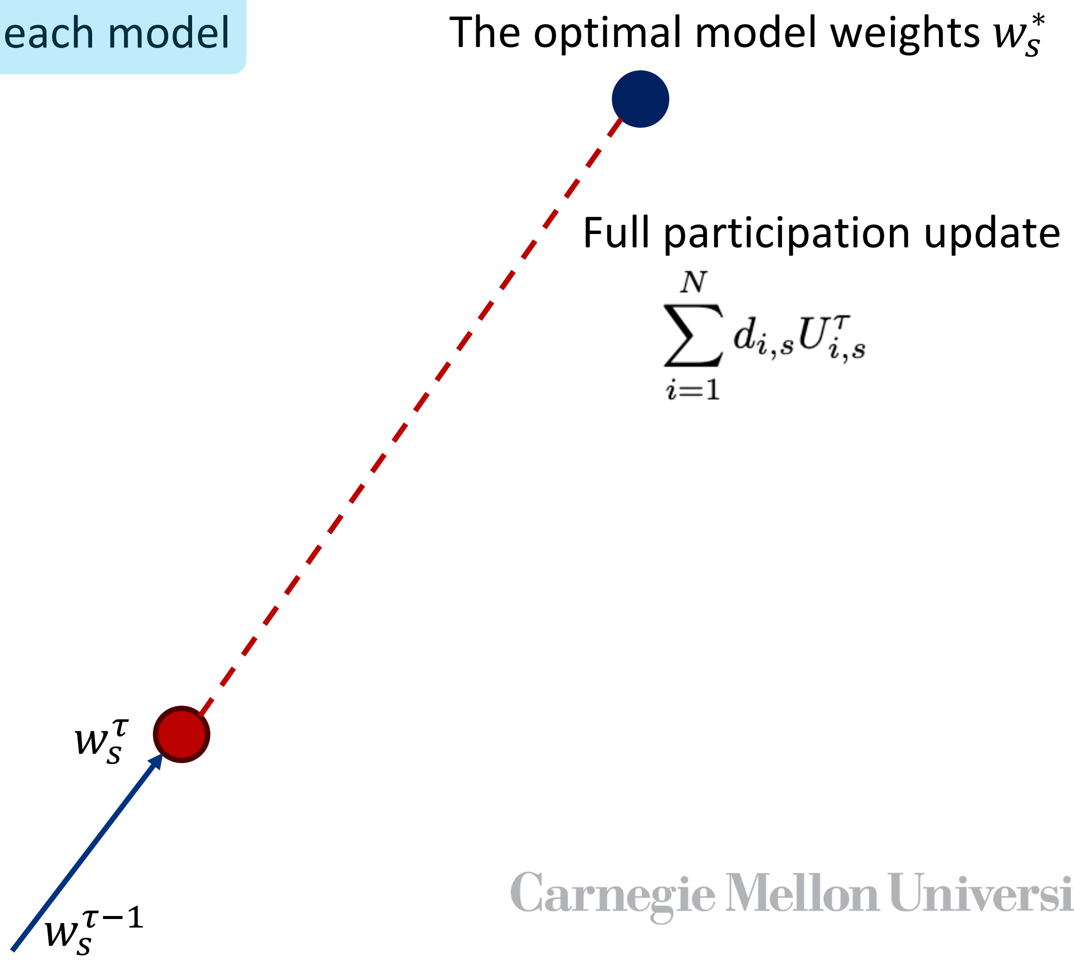
Sampling probability distribution

Sampled update for each model

$\{p_{s|i}^{\tau}\} \rightarrow \mathcal{A}_{\tau,1}, \cdots, \mathcal{A}_{\tau,S} \rightarrow G_1^{\tau}, \cdots, G_s^{\tau}, \cdots, G_S^{\tau}$

Set of active clients for each model

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



MMFL optimal variance-reduced sampling

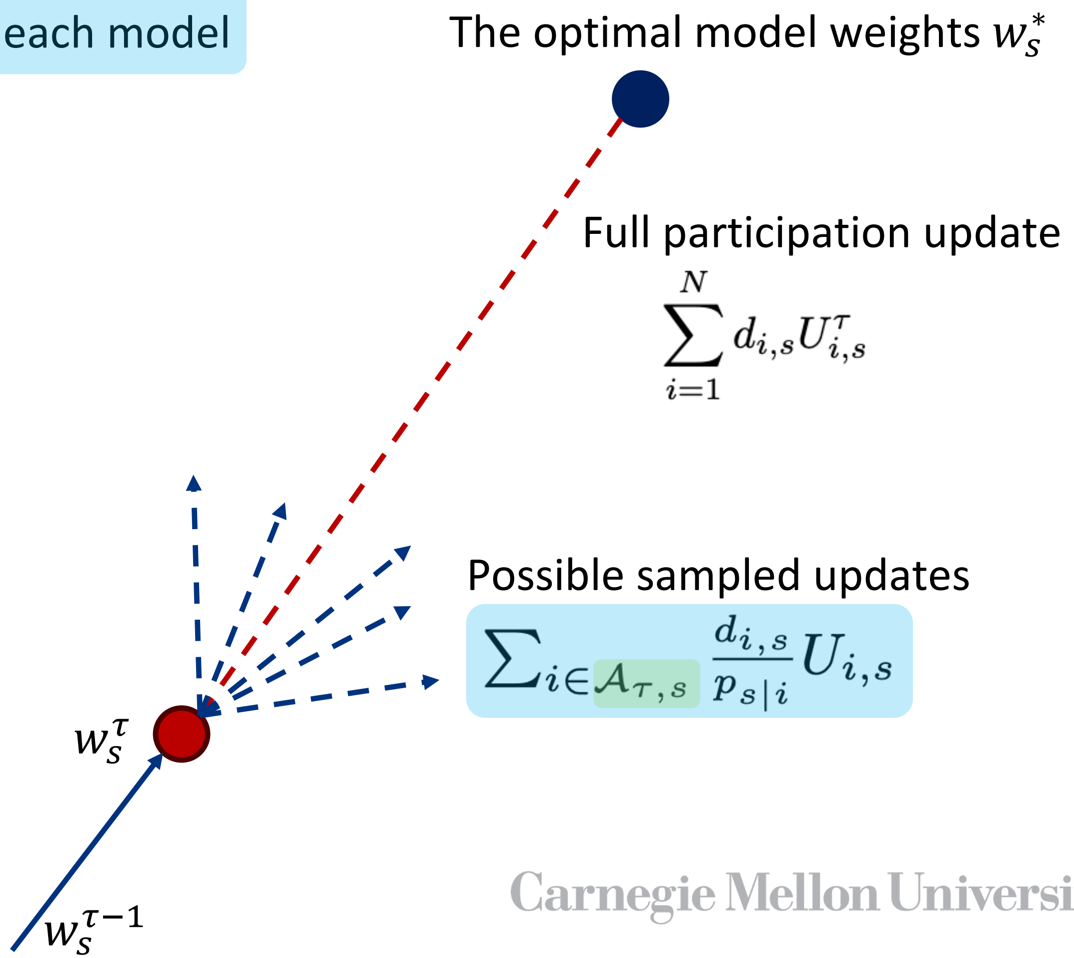
Sampling probability distribution

Sampled update for each model

$$\{p^{\tau}_{s|i}\} \rightarrow \mathcal{A}_{\tau,1}, \cdots, \mathcal{A}_{\tau,S} \rightarrow G^{\tau}_1, \cdots, G^{\tau}_s, \cdots, G^{\tau}_S$$

Set of active clients for each model

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



MMFL optimal variance-reduced sampling

Sampling probability distribution

Sampled update for each model

$\{p^{\tau}_{s|i}\} \rightarrow \mathcal{A}_{\tau,1}, \cdots, \mathcal{A}_{\tau,S} \rightarrow G^{\tau}_{1}, \cdots, G^{\tau}_{s}, \cdots, G^{\tau}_{S}$

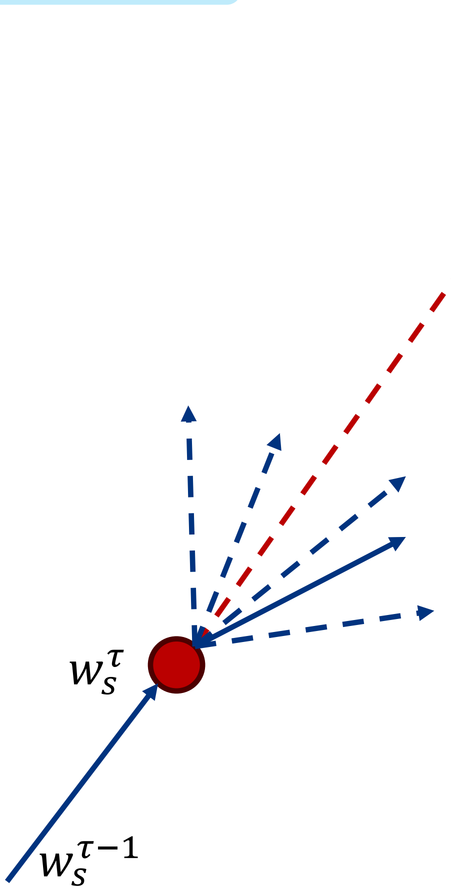
Set of active clients for each model

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

The optimal model weights w_s^*

Full participation update

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



Sampled update

$$\sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p_{s|i}} U_{i,s}$$

MMFL optimal variance-reduced sampling

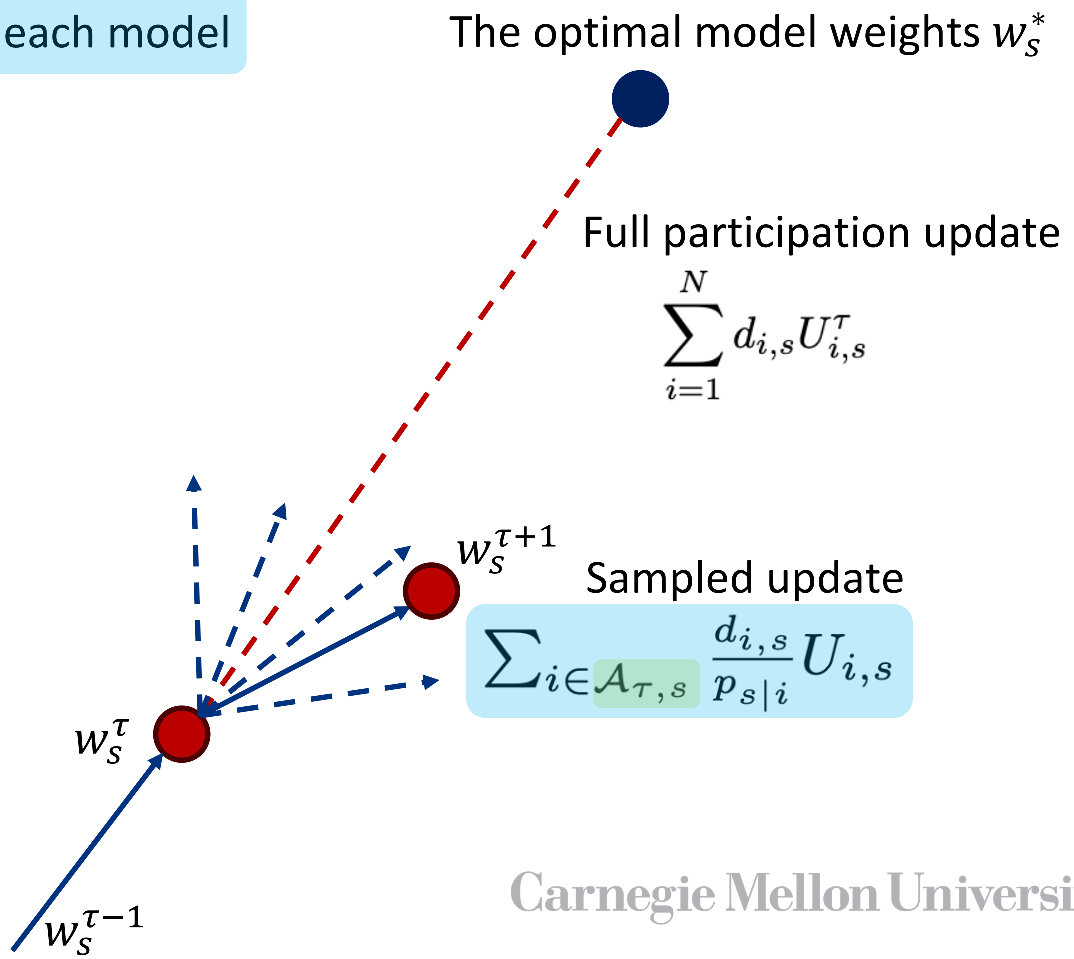
Sampling probability distribution

Sampled update for each model

$$\{p^{\tau}_{s|i}\} \rightarrow \mathcal{A}_{\tau,1}, \cdots, \mathcal{A}_{\tau,S} \rightarrow G_1^{\tau}, \cdots, G_s^{\tau}, \cdots, G_S^{\tau}$$

Set of active clients for each model

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



MMFL optimal variance-reduced sampling

Sampling probability distribution

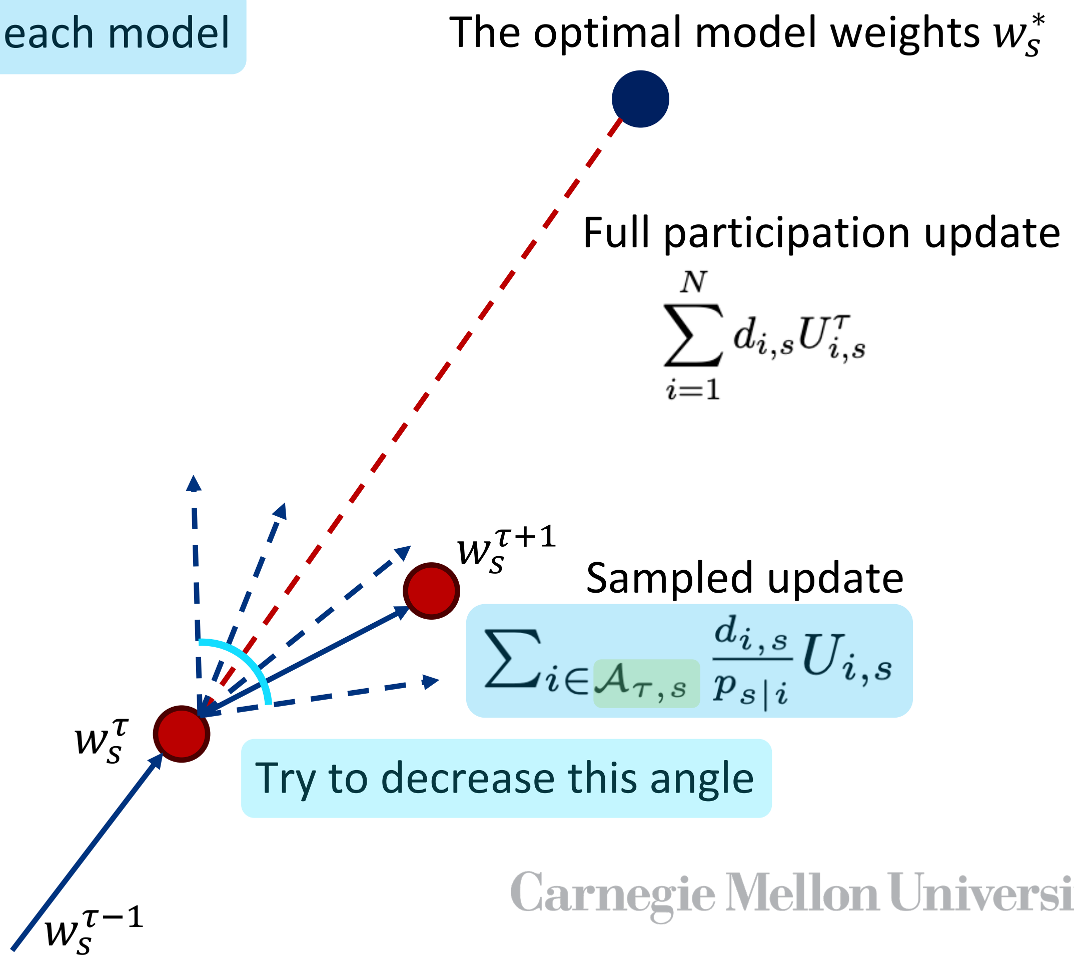
Sampled update for each model

$\{p^{\tau}_{s|i}\} \rightarrow \mathcal{A}_{\tau,1}, \cdots, \mathcal{A}_{\tau,S} \rightarrow G^{\tau}_{1}, \cdots, G^{\tau}_{s}, \cdots, G^{\tau}_{S}$

Set of active clients for each model

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

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



Preliminary experiments

Experiment settings:

N=120 (total clients)

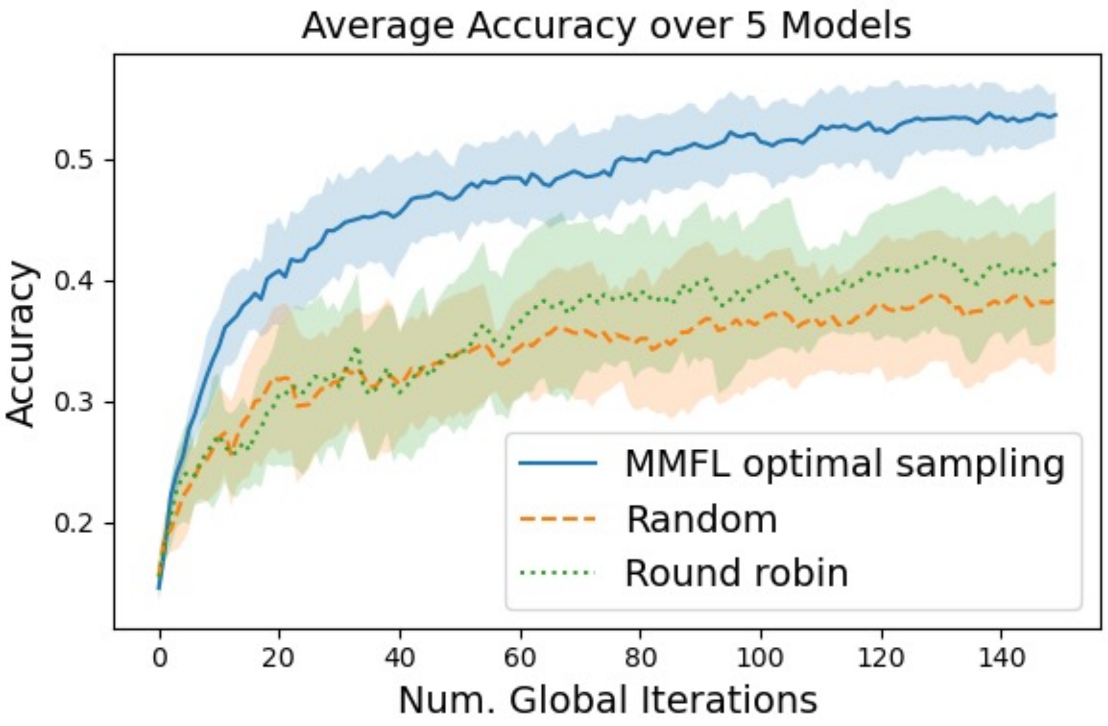
m=12 (expected active clients)

5 models (all Fashion-MNIST classification, with different non-iid level)

52.6% data belongs to 10% clients

E=5 (local epoch number)

5 random seeds



One more step

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 ratio
t: local epoch number

$U_{i,s}^{\tau} = \sum_{t=1}^{E} \nabla f_{i,s}(w_{i,s,\tau}^{t})$
Client i update (gradient)

Sampled update

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

Sampling Probability distribution

Full participation update

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

One more step

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 ratio
t: local epoch number

Client i loss value

Sampled update

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

Sampling Probability distribution

Full participation update

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

One more step

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 ratio
t: local epoch number

Sampled update

Client i loss value

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

Full participation update

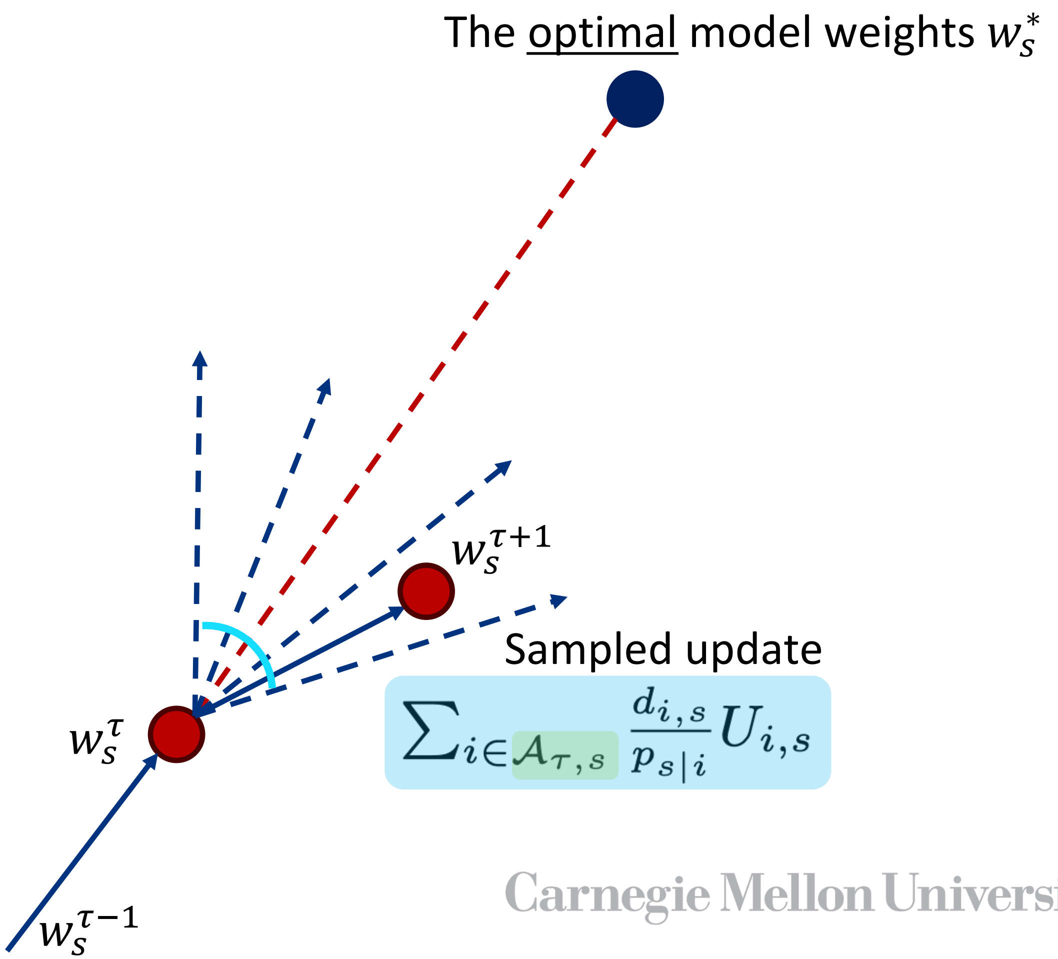
$$\text{s.t. } p^{\tau}_{s|i} \geq 0, \sum_{s=1}^{S} p^{\tau}_{s|i} \leq 1, \sum_{s=1}^{S} \sum_{i=1}^{N} p^{\tau}_{s|i} = m \quad \forall i, s$$

In single-model FL,
they believe this is just an approximation of the gradient norm…

One more step

$$\min_{w_s} F = \sum_{i=1}^{N} d_{i,s} f_{i,s}(w_s)$$

Using loss is in fact optimizing another thing...



One more step

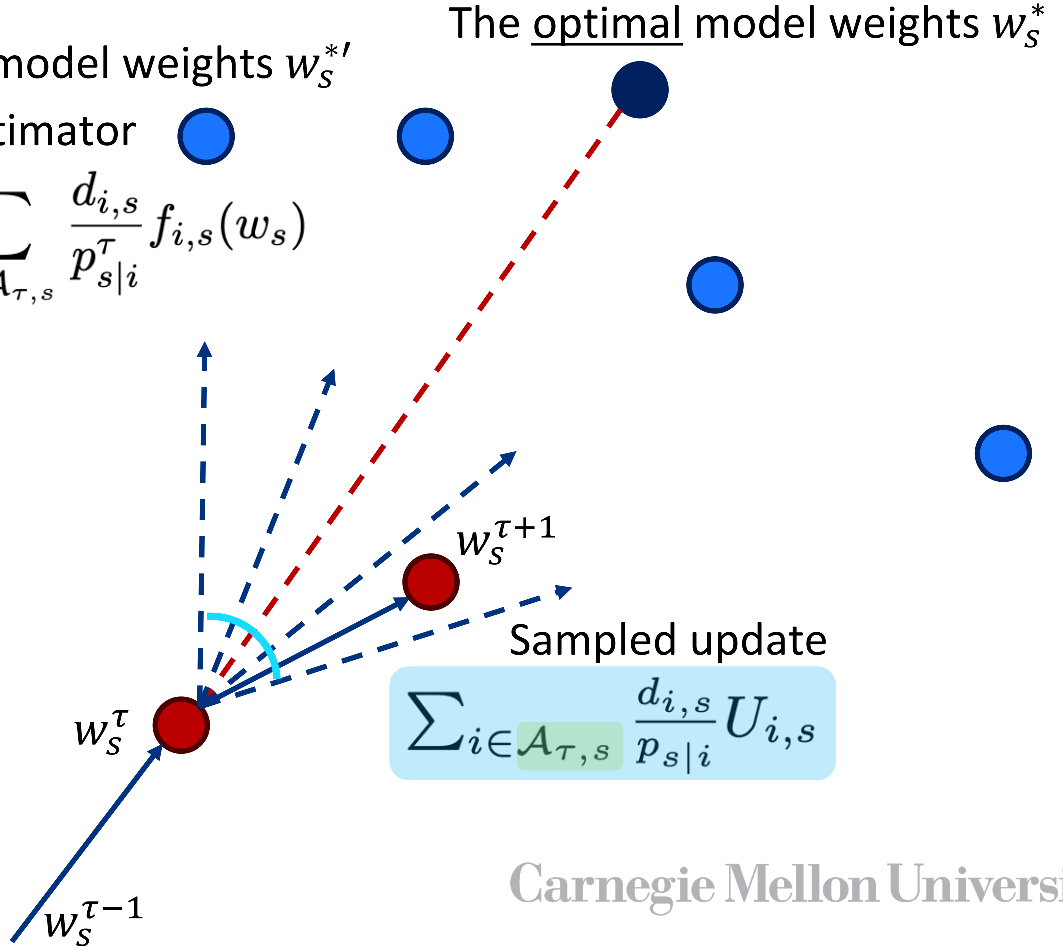
$$\min_{w_s} F = \sum_{i=1}^{N} d_{i,s} f_{i,s}(w_s)$$

Using loss is in fact optimizing another thing…

The optimal model weights $w_s^{*\prime}$

Objective estimator

$$\min_{w_s} F'_\tau = \sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p^\tau_{s|i}} f_{i,s}(w_s)$$



One more step

$$\min_{w_s} F = \sum_{i=1}^{N} d_{i,s} f_{i,s}(w_s)$$

Using loss is in fact optimizing another thing...

The optimal model weights $w_s^{*\prime}$

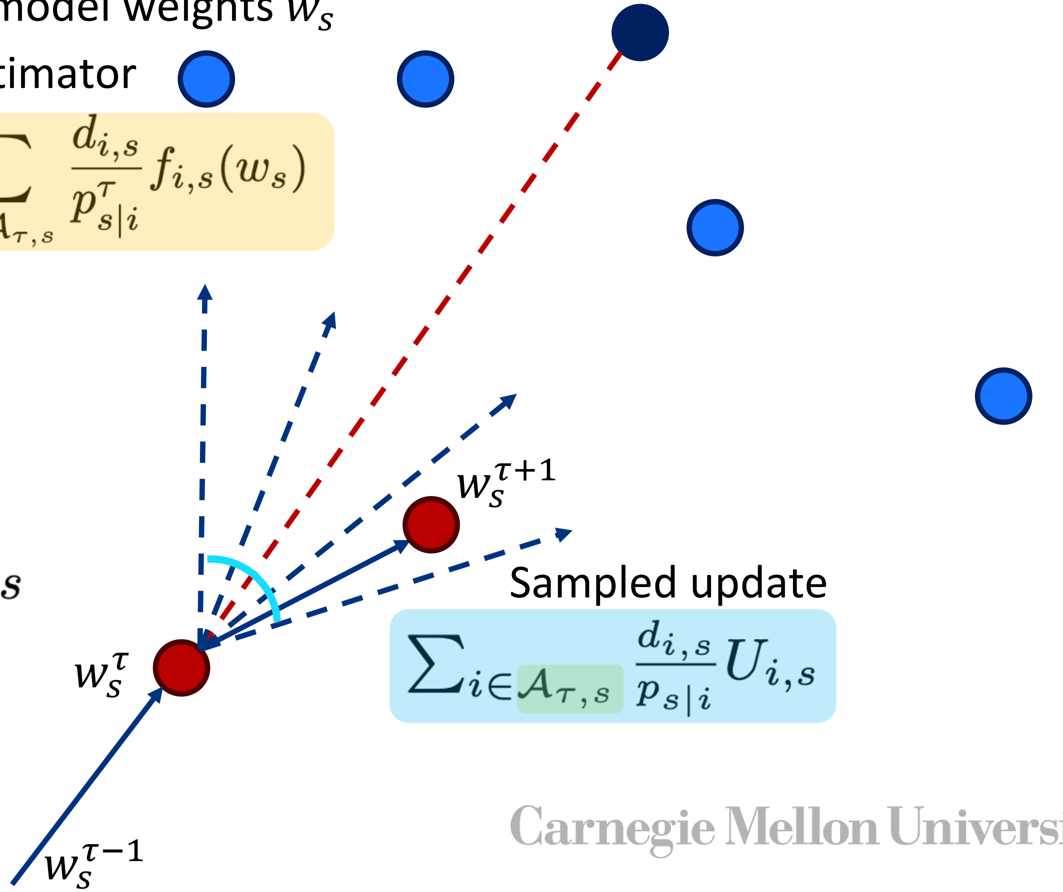
Objective estimator

$$\min_{w_s} F'_\tau = \sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p^\tau_{s|i}} f_{i,s}(w_s)$$

The optimal model weights w_s^{*}

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

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



One more step

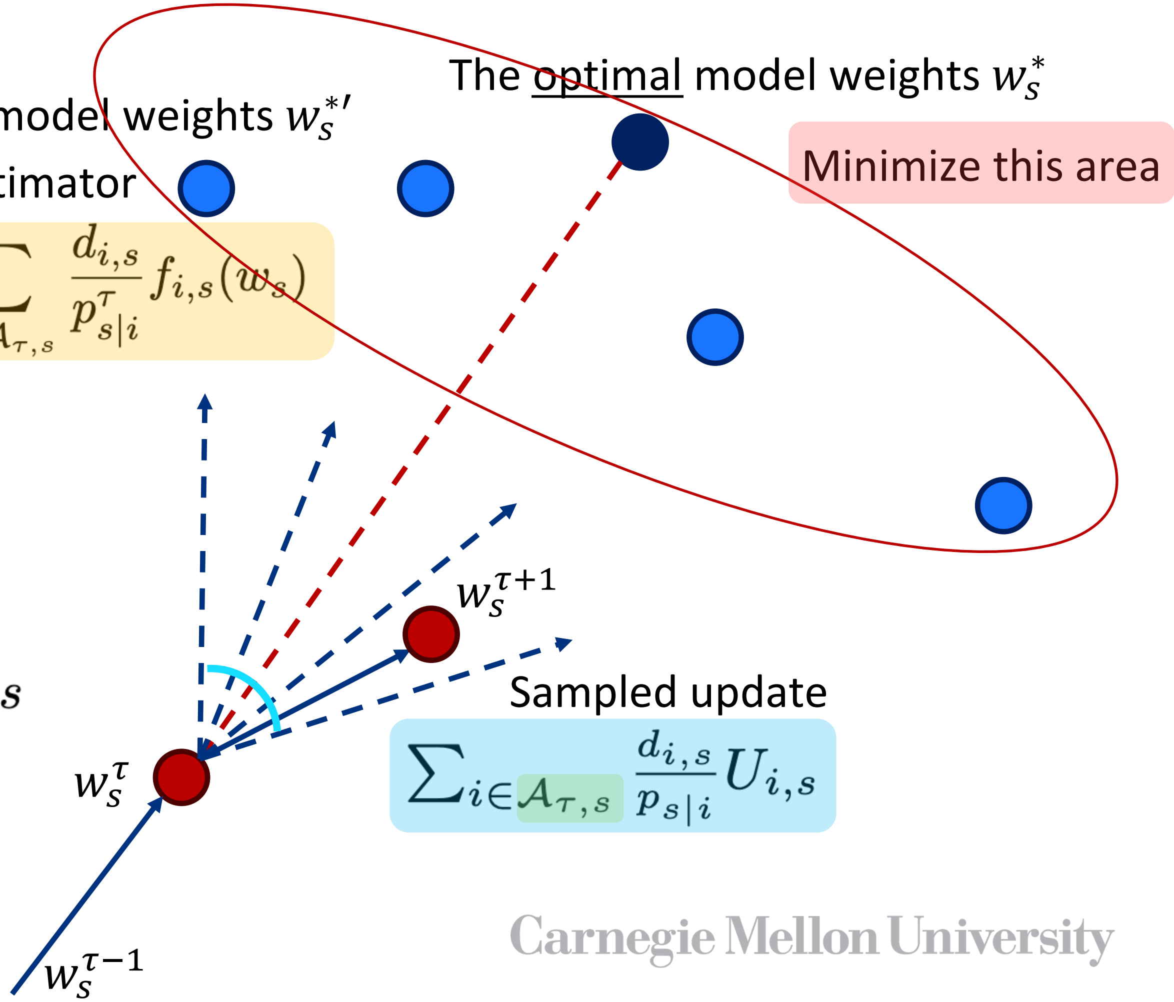
$$\min_{w_s} F = \sum_{i=1}^{N} d_{i,s} f_{i,s}(w_s)$$

Using loss is in fact optimizing another thing...

The optimal model weights $w_s^{*\prime}$

Objective estimator

$$\min_{w_s} F'_\tau = \sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p^\tau_{s|i}} f_{i,s}(w_s)$$



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

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

Experiment results

Experiment settings:

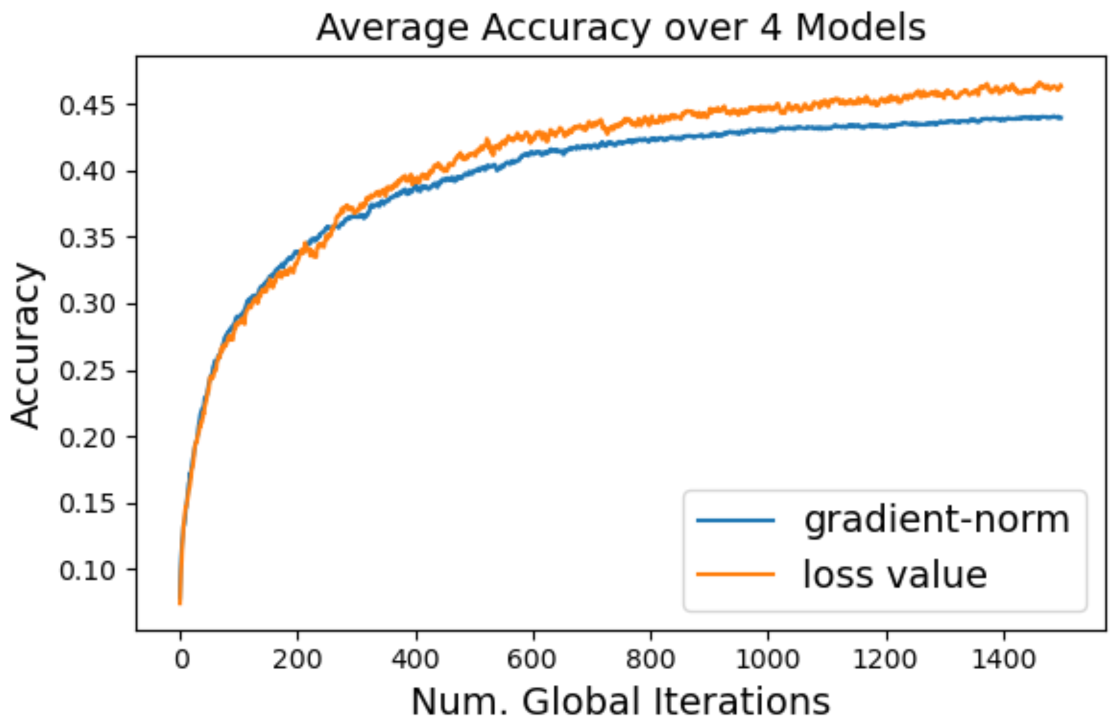
N=80 (total clients)

m=16 (expected active clients)

4 models (Fashion-MNIST, MNIST, EMNIST, Fashion-MNIST)

E=1 (local epoch number)

5 random seeds



Related Work

Optimal Sampling for Federated Learning (single-model)

N: The number of clients.   m: expected number of active clients per round

$$\min_{\{p_i^t\}} \mathbb{E}_{\mathcal{A}_t}\left[\left(\sum_{i\in\mathcal{A}_t}\frac{d_i}{p_i^t}\nabla f_i - \sum_{i=1}^{N} d_i \nabla f_i\right)^2\right]$$

$$\text{s.t. } 0 \leq p_i^t \leq 1 \quad \forall i, \ \sum_{i=1}^{N} p_i^t = m$$

The closed-form solution can be deduced by optimizing a Lagrangian function.

[2] Chen, Wenlin, Samuel Horváth, and Peter Richtárik. "Optimal Client Sampling for Federated Learning." *Transactions on Machine Learning Research* (2022).

Related Work

Optimal Sampling for Federated Learning (single-model)

N: The number of clients.   m: expected number of active clients per round

$$p_i^t = \begin{cases} (m - N + k)\frac{\|\tilde{U}_i\|}{\sum_{j=1}^k \|\tilde{U}_j\|} & \text{if } i = 1, 2, \cdots, k, \\ 1 & \text{if } i = k+1, \cdots, N. \end{cases} \quad (1)$$

where $\|\tilde{U}_i\| = \|d_i U_i\|$, $d_i = n_i / \sum_{j=1}^N n_j$ (n_i is the number of samples for client i), and U_i is an unbiased estimator of ∇f_i. Reorder clients to guarantee $\|\tilde{U}_i\| \leq \|\tilde{U}_{i+1}\|$ for all i, and k is the largest integer for which $0 < (m - N + k) \leq \frac{\sum_{j=1}^k \|\tilde{U}_j\|}{\|\tilde{U}_k\|}$. p_i is the probability of sampling client i for the current round.

[2] Chen, Wenlin, Samuel Horváth, and Peter Richtárik. "Optimal Client Sampling for Federated Learning." *Transactions on Machine Learning Research* (2022).