A Lightweight Method for Tackling Unknown Participation Statistics in Federated Averaging

Shiqiang Wang [1], Mingyue Ji [2]

[1] IBM T. J. Watson Research Center, Yorktown Heights, NY, USA
[2] Department of ECE, University of Utah, Salt Lake City, UT, USA

Mathematical Formulation of Federated Learning

- A machine learning model with parameter $\mathbf{x}$
- How good is $\mathbf{x}$: Individual loss function for data sample ξ_n, $\ell_n(\mathbf{x}, \xi_n)$
- Local objective at client *n*:

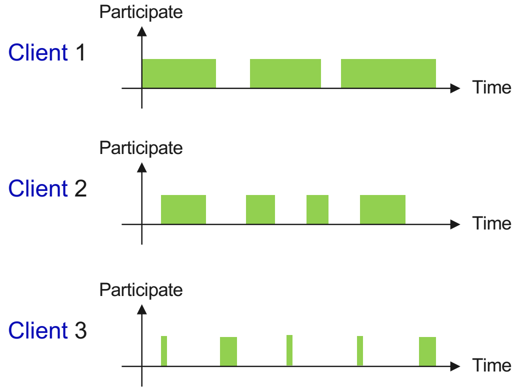
$$F_n(\mathbf{x}) := \mathbb{E}_{\xi_n \sim \mathcal{D}_n} \left[\ell_n(\mathbf{x}, \xi_n) \right]$$

- Global objective (*not directly observable*):

$$f(\mathbf{x}) := \frac{1}{N} \sum_{n=1}^{N} F_n(\mathbf{x})$$

Find $\mathbf{x}^*$ to minimize $f(\mathbf{x})$ → Optimization problem

FedAvg Algorithm



Algorithm 1: FedAvg with pluggable aggregation weights

Input: $\gamma, \eta, \mathbf{x}_0, I$; **Output:** $\{\mathbf{x}_t : \forall t\}$;

Initialize $t_0 \leftarrow 0$, $\mathbf{u} \leftarrow \mathbf{0}$;
for $t = 0, \ldots, T-1$ **do**
  for $n = 1, \ldots, N$ *in parallel* **do**
    Sample $\mathbb{1}_t^n$ from an *unknown* process;
    if $\mathbb{1}_t^n = 1$ **then**
      $\mathbf{y}_{t,0}^n \leftarrow \mathbf{x}_t$;
      for $i = 0, \ldots, I-1$ **do**
        $\mathbf{y}_{t,i+1}^n \leftarrow \mathbf{y}_{t,i}^n - \gamma \mathbf{g}_n(\mathbf{y}_{t,i}^n)$;
      $\Delta_t^n \leftarrow \mathbf{y}_{t,I}^n - \mathbf{x}_t$;
    else
      $\Delta_t^n \leftarrow \mathbf{0}$;
    $\omega_t^n \leftarrow \texttt{ComputeWeight}(\{\mathbb{1}_\tau^n : \tau < t\})$;
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \frac{\eta}{N} \sum_{n=1}^{N} \omega_t^n \Delta_t^n$;

Local updates

Aggregation →

Randomized participation with unknown statistics

Aggregation weights

Challenge: The participation statistics of clients are often *unknown*, *uncontrollable*, and *heterogeneous*

Improper Choice of Aggregation Weights Causes Bias

Only assumed for theoretical analysis

Theorem 1 (Objective minimized at convergence, informal). *When* $\mathbb{I}_t^n \sim \text{Bernoulli}(p_n)$ *and the weights are time-constant, i.e.,* $\omega_t^n = \omega_n$ *but generally* ω_n *may not be equal to* $\omega_{n'}$ $(n \neq n')$, *with properly chosen learning rates* γ *and* η *and some other assumptions, Algorithm 1 minimizes the following objective:*

$$h(\mathbf{x}) := \frac{1}{P} \sum_{n=1}^{N} \omega_n p_n F_n(\mathbf{x}),$$

where $P := \sum_{n=1}^{N} \omega_n p_n$.

Implicit weighting due to partial participation

- Choosing $\omega_n = {}^1/_{p_n}$
  - Objective is consistent with $f(\mathbf{x}) := \frac{1}{N} \sum_{n=1}^{N} F_n(\mathbf{x})$
  - However, impractical when p_n is unknown

The **ideal** case:
Choice of aggregation weight ω_n should *cancel out* the implicit weighting by p_n

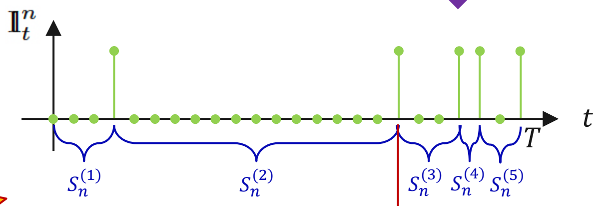
- Choosing other values of ω_n (e.g., $\omega_n = 1, \forall n$)
  - Objective inconsistency, leading to bias (preference of more frequently participating clients)

How to Estimate Aggregation Weights?

Inspired by Bernoulli-distributed participation ➔ Generalize to other participation patterns empirically

$$\omega_n = 1/p_n \quad \xrightarrow{\text{Estimate}} \quad p_n \approx \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}_t^n$$

Problem 1 (Goal of Weight Estimation, informal). *Choose $\{\omega_t^n\}$ so that its long-term average (i.e., for large T) $\frac{1}{T}\sum_{t=0}^{T-1}\omega_t^n$ is close to $\frac{1}{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{I}_t^n}$, for each n.*

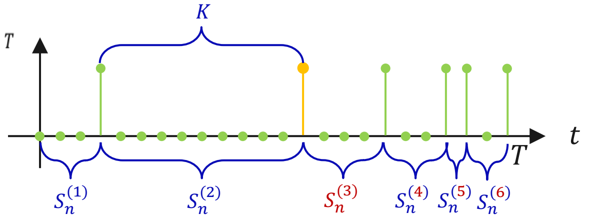


Equivalent to the average of **intervals** between every pair of adjacent participating rounds → Geometric distribution for Bernoulli participating clients (same parameter p_n)

Cannot predict the future
➔ Estimate ω_t^n based on intervals seen so far

Problem: Large overestimate of ω_t^n when large intervals exist (although with low probability) ➔ instability in training

Solution: "Cutoff" interval



Create a dummy interval when the actual interval exceeds K

- Smaller K → lower variance (more samples), but higher bias
- Larger K → higher variance (less samples), but lower bias

FedAU

- FedAvg with adaptive weighting to support unknown participation statistics

Algorithm 1: FedAvg with pluggable aggregation weights

Input: $\gamma, \eta, \mathbf{x}_0, I$; **Output:** $\{\mathbf{x}_t : \forall t\}$;

Initialize $t_0 \leftarrow 0$, $\mathbf{u} \leftarrow \mathbf{0}$;
for $t = 0, \ldots, T-1$ **do**
  for $n = 1, \ldots, N$ *in parallel* **do**
    Sample $\mathbb{I}_t^n$ from an *unknown* process;
    if $\mathbb{I}_t^n = 1$ **then**
      $\mathbf{y}_{t,0}^n \leftarrow \mathbf{x}_t$;
      for $i = 0, \ldots, I-1$ **do**
        $\mathbf{y}_{t,i+1}^n \leftarrow \mathbf{y}_{t,i}^n - \gamma \mathbf{g}_n(\mathbf{y}_{t,i}^n)$;
      $\Delta_t^n \leftarrow \mathbf{y}_{t,I}^n - \mathbf{x}_t$;
    else
      $\Delta_t^n \leftarrow \mathbf{0}$;
    $\omega_t^n \leftarrow \texttt{ComputeWeight}(\{\mathbb{I}_\tau^n : \tau < t\})$;
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \frac{\eta}{N} \sum_{n=1}^{N} \omega_t^n \Delta_t^n$;

Algorithm 2: Weight computation in FedAU

Input: $K, \{\mathbb{I}_t^n : \forall t, n\}$; **Output:** $\{\omega_t^n : \forall t, n\}$;

for $n = 1, \ldots, N$ *in parallel* **do**
  Initialize $M_n \leftarrow 0$, $S_n^\diamond \leftarrow 0$, $\omega_0^n \leftarrow 1$;
  for $t = 1, \ldots, T-1$ **do**
    $S_n^\diamond \leftarrow S_n^\diamond + 1$;
    if $\mathbb{I}_{t-1}^n = 1$ *or* $S_n^\diamond = K$ **then**
      $S_n \leftarrow S_n^\diamond$; // final interval computed

$$\omega_t^n \leftarrow \begin{cases} S_n, & \text{if } M_n = 0 \\ \frac{M_n \cdot \omega_{t-1}^n + S_n}{M_n + 1}, & \text{if } M_n \geq 1 \end{cases};$$

      $M_n \leftarrow M_n + 1$;
      $S_n^\diamond \leftarrow 0$;
    else
      $\omega_t^n \leftarrow \omega_{t-1}^n$;

Cutoff condition of interval length

Online interval computation and averaging

Main Result

Theorem 2 (Convergence error w.r.t. (1)). *Let* $\gamma \leq \frac{1}{4\sqrt{15LI}}$ *and* $\gamma\eta \leq \min\left\{\frac{1}{4LI}; \frac{N}{54LIQ}\right\}$, *where* $Q := \max_{t\in\{0,\dots,T-1\}} \frac{1}{N}\sum_{n=1}^{N} p_n(\omega_t^n)^2$. *When Assumptions 1–5 hold, the result* $\{\mathbf{x}_t\}$ *obtained from Algorithm 1 satisfies:*

Defined in the paper

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

Weight error term

$$\leq \mathcal{O}\left(\frac{\mathcal{F}}{\gamma\eta IT} + \frac{\Psi_G+\delta^2+\gamma^2L^2I\sigma^2}{NT}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\mathbb{E}\left[\left(p_n\omega_t^n-1\right)^2\right] + \frac{\gamma\eta LQ\left(I\delta^2+\sigma^2\right)}{N} + \gamma^2L^2I\left(I\delta^2+\sigma^2\right)\right),$$

where $\mathcal{F} := f(\mathbf{x}_0) - f^*$, *and* $f^* := \min_{\mathbf{x}} f(\mathbf{x})$ *is the truly minimum value of the objective in* (1).

Theorem 3 (Bounding the weight error term). *For* $\{\omega_t^n\}$ *obtained from Algorithm 2, when* $T \geq 2$,

$$\frac{1}{NT}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\mathbb{E}\left[\left(p_n\omega_t^n-1\right)^2\right] \leq \mathcal{O}\left(\frac{K\log T}{T} + \frac{1}{N}\sum_{n=1}^{N}(1-p_n)^{2K}\right).$$

Related to variance

Related to bias

Confirms the bias-variance tradeoff:

- Small K → low variance, high bias
- Large K → high variance, low bias

Final Convergence Rate

Corollary 4 (Convergence of FedAU). *Let* $K = \lceil \log_c T \rceil$ *with* $c := {}^1/(1-\min_n p_n)^2$, $\gamma = \min\left\{\frac{1}{LI\sqrt{T}}; \frac{1}{4\sqrt{15}LI}\right\}$, *and choose* η *such that* $\gamma\eta = \min\left\{\sqrt{\frac{\mathcal{F}N}{Q(I\delta^2+\sigma^2)LIT}}; \frac{1}{4LI}; \frac{N}{54LIQ}\right\}$. *When* $T \geq 2$, *the result* $\{\mathbf{x}_t\}$ *obtained from Algorithm 1 that uses* $\{\omega_t^n\}$ *obtained from Algorithm 2 satisfies*

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

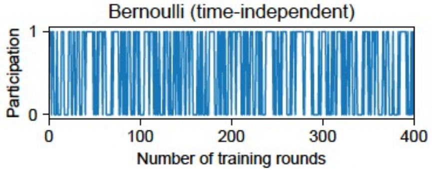
Upper bound of weight error term

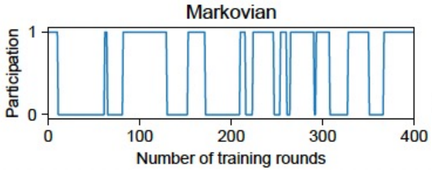
Standard in FedAvg

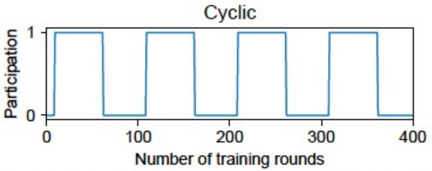
$$\leq \mathcal{O}\left(\frac{\sigma\sqrt{L\mathcal{F}Q}}{\sqrt{NIT}} + \frac{\delta\sqrt{L\mathcal{F}Q}}{\sqrt{NT}} + \frac{\left(\Psi_G + \delta^2 + \frac{\sigma^2}{IT}\right)R\log^2 T}{T} + \frac{L\mathcal{F}\left(1+\frac{Q}{N}\right)+\delta^2+\frac{\sigma^2}{I}}{T}\right),$$

where Q *and* Ψ_G *are defined in Theorem 2 and* $R := {}^1/\log c$.

Experiments







Participation pattern	Dataset	SVHN		CIFAR-10		CIFAR-100		CINIC-10	
	Method / Metric	Train	Test	Train	Test	Train	Test	Train	Test
	FedAU (ours, $K \to \infty$)	90.4±0.5	89.3±0.5	85.4±0.4	77.1±0.4	63.4±0.6	**52.3**±0.4	65.2±0.5	61.5±0.4
	FedAU (ours, $K = 50$)	**90.6**±0.4	**89.6**±0.4	**86.0**±0.5	**77.3**±0.3	**63.8**±0.3	52.1±0.6	**66.7**±0.3	**62.7**±0.2
	Average participating	89.1±0.3	87.2±0.3	83.5±0.9	74.1±0.8	59.3±0.4	48.8±0.7	61.1±2.3	56.6±2.0
	Average all	88.5±0.5	87.0±0.3	81.0±0.9	72.7±0.9	58.2±0.4	47.9±0.5	60.5±2.3	56.2±2.0
	FedVarp (250× memory)	89.6±0.5	88.9±0.5	84.2±0.3	77.9±0.2	57.2±0.9	49.2±0.8	64.4±0.6	62.0±0.5
	MIFA (250× memory)	89.4±0.3	88.7±0.2	83.5±0.6	77.5±0.3	55.8±1.1	48.4±0.7	63.8±0.7	61.5±0.5
	Known participation statistics	89.2±0.5	88.4±0.5	84.3±0.5	77.0±0.5	59.4±0.7	50.6±0.4	63.2±0.6	60.5±0.5
	FedAU (ours, $K \to \infty$)	90.5±0.4	89.3±0.4	85.3±0.3	77.1±0.3	63.2±0.5	**51.8**±0.3	64.9±0.3	61.2±0.2
	FedAU (ours, $K = 50$)	**90.6**±0.3	**89.5**±0.3	**85.9**±0.5	**77.2**±0.3	**63.5**±0.4	51.7±0.3	**66.3**±0.4	**62.3**±0.2
	Average participating	89.0±0.3	87.1±0.2	83.4±0.9	74.2±0.7	59.2±0.4	48.6±0.4	61.5±2.3	56.9±1.9
	Average all	88.4±0.6	86.8±0.7	80.8±1.0	72.5±0.5	57.8±0.9	47.7±0.5	59.9±2.8	55.7±2.2
	FedVarp (250× memory)	89.6±0.3	88.6±0.2	84.0±0.3	77.8±0.2	56.4±1.1	48.8±0.5	64.6±0.4	62.1±0.4
	MIFA (250× memory)	89.1±0.3	88.4±0.2	83.0±0.4	77.2±0.4	55.1±1.2	48.1±0.6	63.5±0.7	61.2±0.6
	Known participation statistics	89.5±0.2	88.6±0.2	84.5±0.4	76.9±0.3	59.7±0.5	50.3±0.5	63.5±0.9	60.7±0.6
	FedAU (ours, $K \to \infty$)	89.8±0.6	88.7±0.6	84.2±0.8	76.3±0.7	60.9±0.6	50.6±0.3	63.5±1.0	60.0±0.8
	FedAU (ours, $K = 50$)	**89.9**±0.6	**88.8**±0.6	**84.8**±0.6	**76.6**±0.4	**61.3**±0.8	**51.0**±0.5	**64.5**±0.9	**60.9**±0.7
	Average participating	87.4±0.5	85.5±0.7	81.6±1.2	73.3±0.8	58.1±1.0	48.3±0.8	58.9±2.1	55.0±1.6
	Average all	89.1±0.8	87.4±0.8	83.1±1.0	73.8±0.8	59.7±0.3	48.8±0.4	62.9±1.7	57.6±1.5
	FedVarp (250× memory)	84.8±0.5	83.9±0.6	79.7±0.9	75.3±0.7	50.9±0.5	45.9±0.4	60.4±0.7	58.5±0.6
	MIFA (250× memory)	78.6±1.2	77.4±1.1	73.0±1.3	70.6±1.1	44.8±0.6	41.1±0.6	51.2±1.0	50.2±0.9
	Known participation statistics	89.9±0.7	88.7±0.6	83.6±0.7	76.1±0.5	60.2±0.4	50.8±0.4	62.6±0.8	59.8±0.7

- Same stationary probability for all participation patterns (but different across clients), initial state/offset is randomized
- Participation rate is correlated with heterogeneous data distribution

Thank You!

Email: wangshiq@us.ibm.com
Homepage: https://shiqiang.wang/