Joint Optimization for Federated Learning Over the Air

Xin Fan[1] , Yue Wang[2] , Yan Huo[1] , and Zhi Tian[2]
[1]Beijing Jiaotong University, China
[2]George Mason University, USA
E-mail: {yhuo, fanxin}@bjtu.edu.cn, {ywang56,ztian1}@gmu.edu

IEEE ICC®
IEEE International Conference on Communications
16–20 May 2022 // Seoul, South Korea // Hybrid: In-Person and Virtual Conference
Intelligent Connectivity for Smart World

Outline

- Introduction
- System Model
- Convergence Analysis
- Performance Optimization
- Simulation Results
- Conclusion

Outline

- **Introduction**
- System Model
- Convergence Analysis
- Performance Optimization
- Simulation Results
- Conclusion

Introduction --Background

Conventional centralized machine learning

- **Characteristics:**
  - Collect data for training
- **Cons:**
  - **Big data**
  - **Privacy**

Federated learning (FL)

- **Characteristics:**
  - Keep data locally
  - Send weights or gradients to a center
- **Cons:**
  - **Massive devices**

FL over the air

- **Characteristics:**
  - Uncoding
  - Linear analog modulation
  - Analog aggregation
- **Pros:**
  - **Allow all local devices transmit updates using the same time-frequency resources.**

Introduction --Motivation

Challenges

- Aggregation errors, such as **channel fading**, **noise perturbation**, and so on.
- How these aggregation errors affect FL?
- Without local model parameters known in advance, how to achieve power control?
- Simple maximization of the number of participated devices is not necessarily optimal
- How to select local devices?

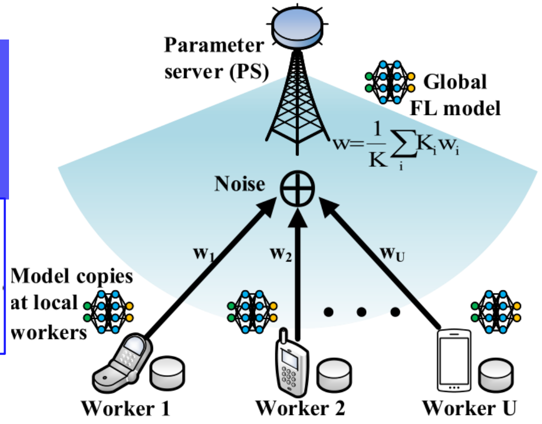
Contributions

- Convergence analysis.
- Optimization scheme.

Outline

- Introduction
- **System Model**
- Convergence Analysis
- Performance Optimization
- Simulation Results
- Conclusion

System Model



$$\textbf{P1:} \quad \min_{\mathbf{w}} \quad \frac{1}{K}\sum_{i=1}^{U}\sum_{k=1}^{K_i} f(\mathbf{w}_i; \mathbf{x}_{i,k}, y_{i,k})$$

- **Federated learning (FL)**
  - Local devices (workers)
    - Receive $\mathbf{w} = [w^1, \ldots, w^D] \in \mathcal{R}^D$ from a parameter server (PS)
    - Train to get the updates (local parameters, $\mathbf{w}_i$)
    - Send $\mathbf{w}_i$ to the PS
  - PS
    - Receive $\mathbf{w}_i$ and average them to obtain the sharing model $\boxed{\mathbf{w} = \frac{\sum_{i=1}^{U} K_i \mathbf{w}_i}{K}}$
    - Broadcast $\mathbf{w}$ to local workers
- FL over the air
  - Local worker
    - Send $\mathbf{w}_{i,t}$ with the power control policy $\mathbf{p}_{i,t} = [p_{i,t}^1, \ldots, p_{i,t}^d, \ldots, p_{i,t}^D]$ where $p_{i,t}^d = \frac{\beta_{i,t}^d K_i b_t^d}{h_{i,t}^d}$
  - PS
    - Receive $\mathbf{y}_t = \sum_{i=1}^{U} \mathbf{p}_{i,t} \odot \mathbf{w}_{i,t} \odot \mathbf{h}_{i,t} + \mathbf{z}_t$
    - Estimate $\mathbf{w}_t$ via a post-processing operation as

$$\boxed{\begin{aligned}\mathbf{w}_t = \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \mathbf{b}_t\right)^{\odot -1} \odot \mathbf{y}_t = \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \mathbf{b}_t\right)^{\odot -1} \odot \mathbf{z}_t \\ + \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t}\right)^{\odot -1} \sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \mathbf{w}_{i,t}, \qquad (8)\end{aligned}}$$

Outline

- Introduction
- System Model
- **Convergence Analysis**
- Performance Optimization
- Simulation Results
- Conclusion

Convergence Analysis

Basic Assumptions

- Assumption 1 (Lipschitz continuity, smoothness): $\|\nabla F(\mathbf{w}_{t+1}) - \nabla F(\mathbf{w}_t)\| \leq L\|\mathbf{w}_{t+1} - \mathbf{w}_t\|$
- Assumption 2 (strongly convex): $F(\mathbf{w}_{t+1}) \geq F(\mathbf{w}_t) + (\mathbf{w}_{t+1} - \mathbf{w}_t)^T \nabla F(\mathbf{w}_t) + \frac{\mu}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2, \quad \forall \mathbf{w}_t, \mathbf{w}_{t+1}.$
- Assumption 3 (bounded sample-wise gradient): $\| \nabla f(\mathbf{w}_t) \|^2 \leq \rho_1 + \rho_2 \| \nabla F(\mathbf{w}_t) \|^2$

Convex Case with Full Gradient Descent (GD)

Theorem 1. *Adopt **Assumptions 1-3**, and the model updating rule for $\mathbf{w}_t$ of the FL-over-the-air scheme is given by (8), $\forall t$. Given the learning rate $\alpha = \frac{1}{L}$, the expected performance gap $\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)]$ of $\mathbf{w}_t$ at the t-th iteration is given by*

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \leq \underbrace{\sum_{i=1}^{t-1} \prod_{j=1}^{i} A_{t+1-j} B_{t-i} + B_t}_{\Delta_t} + \prod_{j=1}^{t} A_j \mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)], \quad (12)$$

where $A_t = 1 - \frac{\mu}{L} + \rho_2 \sum_{d=1}^{D} \left(\frac{K}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} - 1\right)$ *and* $B_t = \frac{\rho_1}{2L} \sum_{d=1}^{D} \left(\frac{K}{\sum_{i=1}^{U} K_i \beta_{i,t}^d} - 1\right) + \|(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \mathbf{b}_t)^{\odot -1}\|^2 \frac{L\sigma^2}{2}$.

Performance gap

Guideline for optimization

Imposes a convergence condition

$$0 < \rho_2 < \frac{\mu}{(\frac{K}{K_{min}} - 1)DL}$$

Convergence Analysis

- Non-convex Case

Theorem 2. *Under the* **Assumptions** *1 and 3 for the non-convex case, given the learning rate* $\alpha = \frac{1}{L}$, *the convergence at the* T*-th iteration is given by*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_{t-1})\|^2 \leq \frac{2L}{T(1-\rho_2 D(\frac{K}{K_{min}}-1))}\mathbb{E}[F(\mathbf{w}_0)] - F(\mathbf{w}^*)] + \frac{2L\sum_{t=1}^{T}B_t}{T(1-\rho_2 D(\frac{K}{K_{min}}-1))}. \quad (13)$$

$$\min_{0,1,\dots,T}\mathbb{E}[\|\nabla F(\mathbf{w}_{t-1})\|^2] \leq \frac{1}{T}\sum_{t=1}^{T}\|\nabla F(\mathbf{w}_{t-1})\|^2 \overset{T\to\infty}{\leq} \underbrace{\frac{2L\sum_{t=1}^{T}B_t}{T(1-\rho_2 D(\frac{K}{K_{min}}-1))}}_{\triangle_T^{NC}}$$

Guideline for optimization

Performance gap

Convergence Analysis

- Stochastic gradient descent

Theorem 3. *Under the* ***Assumptions 1, 2*** *and* ***3*** *for the convex case, given the learning rate $\alpha = \frac{1}{L}$ and the mini-batch size K_b, the convergence behavior of the SGD implementation of FL over the air is given by*

$$\mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \leq \underbrace{\sum_{i=1}^{t-1} \prod_{j=1}^{i} A_{t+1-j}^{SGD} B_{t-i}^{SGD} + B_t^{SGD}}_{\Delta_t^{SGD}} + \prod_{j=1}^{t} A_j^{SGD} \mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)], \quad (15)$$

where $A_t^{SGD} = 1 - \frac{\mu}{L} + \rho_2 (\sum_{d=1}^{D} (\frac{(\sum_{i=1}^{U} K_b)^2 - 2K(\sum_{i=1}^{U} K_b)}{K^2} + \frac{(\sum_{i=1}^{U} K_b)}{\sum_{i=1}^{U} K_b \beta_{i,t}^d}) + \frac{(\sum_{i=1}^{U} (K_i - K_b))^2}{K^2})$ *and* $B_t^{SGD} = \frac{\rho_1}{2L} (\sum_{d=1}^{D} (\frac{(\sum_{i=1}^{U} K_b)^2 - 2K(\sum_{i=1}^{U} K_b)}{K^2} + \frac{(\sum_{i=1}^{U} K_b)}{\sum_{i=1}^{U} K_b \beta_{i,t}^d}) + \frac{(\sum_{i=1}^{U} (K_i - K_b))^2}{K^2}) + \left\| \left(\sum_{i=1}^{U} K_i \boldsymbol{\beta}_{i,t} \odot \boldsymbol{b}_t \right)^{\odot -1} \right\|^2 \frac{L\sigma^2}{2}.$

Guideline for optimization

Performance gap

Imposes a convergence condition

$$0 < \rho_2 < \frac{\mu}{(\frac{2UK_b}{K} + \frac{U^2K_b^2}{K^2} + DU - \frac{2DUK_b}{K} + \frac{DU^2K_b^2}{K^2})L}$$

Outline

- Introduction
- System Model
- Convergence Analysis
- **Performance Optimization**
- Simulation Results
- Conclusion

Performance Optimization -- Problem Formulation

- Minimizing performance gap

$$\begin{aligned} \triangle_t &= B_t + A_t \triangle_{t-1}, \\ \triangle_t^{NC} &= B_t, \\ \triangle_t^{SGD} &= B_t^{SGD} + A_t^{SGD} \triangle_{t-1}^{SGD}. \end{aligned}$$

- For entry-wise optimization

$$\begin{aligned} R_t[d] &= \frac{L\sigma^2}{2\left(\sum_{i=1}^U \beta_{i,t}^d K_i b_t^d\right)^2} + \frac{K\rho_1 + 2KL\rho_2\triangle_{t-1}}{2L\sum_{i=1}^U K_i \beta_{i,t}^d}, \quad \forall d, \\ R_t^{NC}[d] &= \frac{L\sigma^2}{2\left(\sum_{i=1}^U \beta_{i,t}^d K_i b_t^d\right)^2} + \frac{K\rho_1}{2L\sum_{i=1}^U K_i \beta_{i,t}^d}, \quad \forall d, \\ R_t^{SGD}[d] &= \frac{L\sigma^2}{2\left(\sum_{i=1}^U \beta_{i,t}^d K_i b_t^d\right)^2} + \frac{U(\rho_1 + 2L\rho_2\triangle_{t-1})}{2L\sum_{i=1}^U K_i \beta_{i,t}^d}, \quad \forall d. \end{aligned}$$

- **Optimization problem P2:**

$$\begin{aligned} \min_{\{b_t, \beta_{i,t}\}_{i=1}^U} \quad & R_t \\ \text{s.t.} \quad & \left| \frac{\beta_{i,t} K_i b_t}{h_{i,t}} w_{i,t} \right|^2 \le P_i^{\max}, \\ & \beta_{i,t} \in \{0,1\}, i \in \{1,2,\dots,U\} \end{aligned}$$

Assumption 4 (bounded local gradients):

$$|w_{t-1} - w_{i,t}| \le \eta$$

- **Optimization problem P3:**

$$\begin{aligned} \min_{\{b_t, \beta_{i,t}\}_{i=1}^U} \quad & R_t \\ \text{s.t.} \quad & \left| \frac{\beta_{i,t} K_i b_t}{h_{i,t}} \right|^2 (|w_{t-1}| + \eta)^2 \le P_i^{\max} \\ & \beta_{i,t} \in \{0,1\}, i \in \{1,2,\dots,U\}, \end{aligned}$$

Performance Optimization --Solution

- Tight search space

Theorem 4. *When all the required parameters in* ***P3*** *i.e.,* $\{P_i^{\max}, w_{t-1}, h_{i,t}, K_i, \eta\}_{i=1}^{U}$, *are available at the PS, the solution space of* $(b_t, \beta_{i,t})$ *in* ***P3*** *can be reduced to the following tight search space without loss of optimality as*

$$\mathcal{S} = \left\{ \left\{ \left(b_t^{(k)}, \beta_{i,t}^{(k)} \right) \right\}_{k=1}^{U} \middle| b_t^{(k)} = \left| \frac{\sqrt{P_k^{\max}} h_{k,t}}{K_k(|w_{t-1}| + \eta)} \right|, \right.$$
$$\left. \boldsymbol{\beta}_t^{(k)}(b_t^{(k)}) = \left[\beta_{1,t}^{(k)}, \ldots, \beta_{U,t}^{(k)} \right], k = 1, \ldots, U \right\}, \quad (23)$$

where $\boldsymbol{\beta}_t^{(k)}$ *is a function of* $b_t^{(k)}$, *in the form* $\beta_{i,t}^{(k)} = H(P_i^{\max} - |\frac{K_i b_t^{(k)}(|w_{t-1}|+\eta)}{h_{i,t}}|)$ *and* $H(x)$ *is the Heaviside step function, i.e.,* $H(x) = 1$ *for* $x > 0$, *and* $H(x) = 0$ *otherwise.*

Discrete Programming:

$$\boxed{\min_{(b_t, \boldsymbol{\beta}_t) \in \mathcal{S}} R_t = R_t(b_t, \boldsymbol{\beta}_t)}$$

Complexity: $\mathcal{O}(U)$

Outline

- Introduction
- System Model
- Convergence Analysis
- Performance Optimization
- **Simulation Results**
- Conclusion

Simulation Results

- Wireless network setting
  - $U = 20$, $P_i^{\max} = P^{\max} = 10$ mW
  - Rayleigh fading model.
- Two baseline methods for comparison
  - *Perfect aggregation*
  - *Random policy*
- Two tasks
  - Linear regression with a synthetic dataset
  - Image classification with the MNIST dataset

Simulation Results

- Linear regression experiments
  - The optimal result of a linear regression is: $y = -2x + 1$
  - The input x and the output y follow the function: $y = -2x + 1 + n \times 0.4$

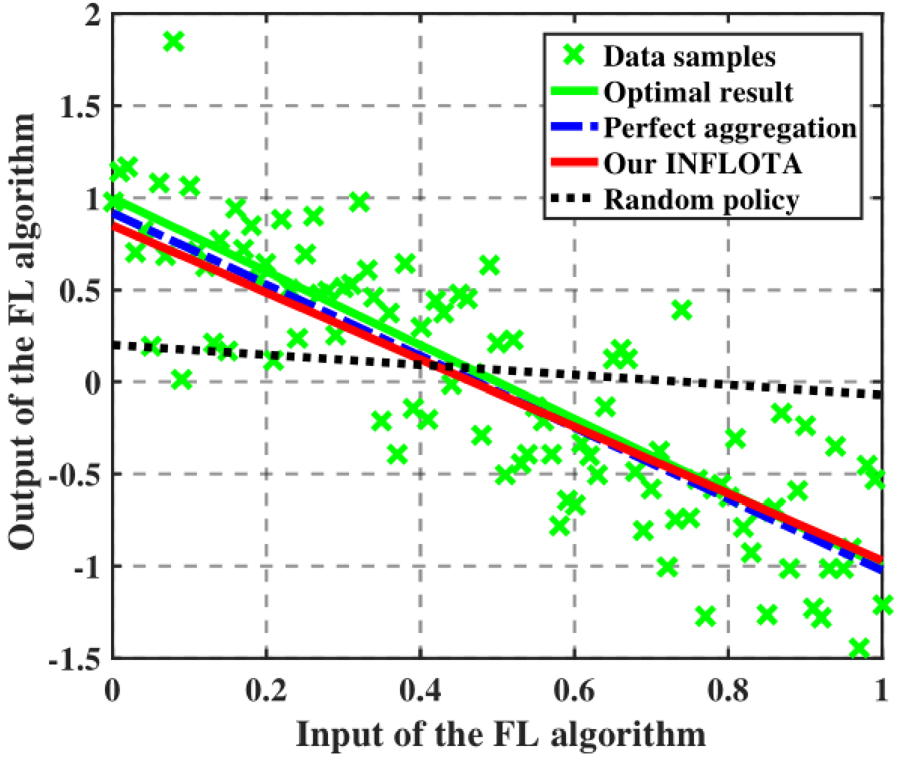


Fig. 2: An example of implementing FL for linear regression.

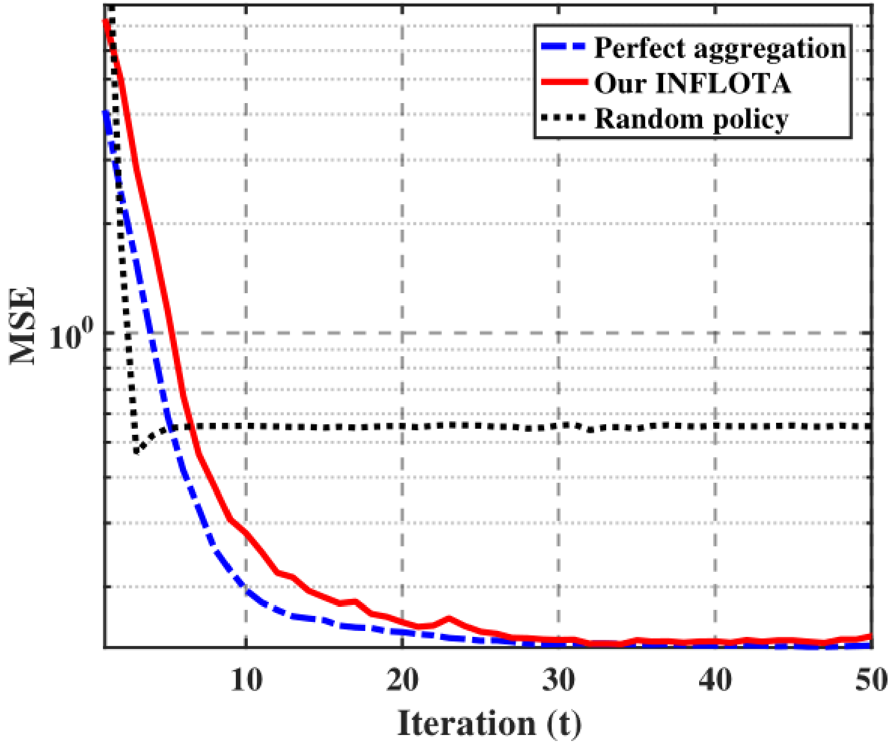


Fig. 3: MSE as the number of iteration varies.

Simulation Results

- Evaluation on the MNIST dataset
  - A multi-layer perceptron (MLP) with a 784-neuron input layer, a 64-neuron hidden layer, and a 10-neuron softmax output layer.
  - The total number of parameters in the MLP is 50890.

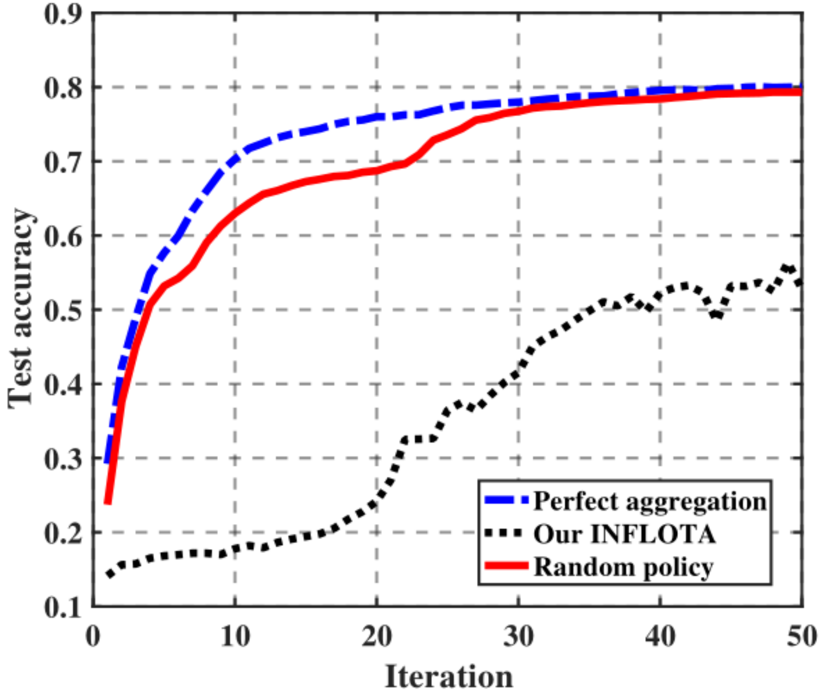


Fig. 4: The test accuracy as the iteration varies.

Outline

- Introduction
- System Model
- Convergence Analysis
- Performance Optimization
- Simulation Results and Evaluation
- **Conclusion**

Conclusion

- Under the convex and non-convex cases with either the GD or SGD implementations, we respectively derive the expected convergence rate of FL.
- We propose a joint optimization scheme of worker selection and power control.
- Our joint optimization scheme is applicable for both the convex and non-convex cases, using either GD or SGD implementations.

THANK YOU

Questions?

Xin FAN

Email: fanxin@bjtu.edu.cn