Communication-efficient Federated Learning Through 1-Bit Compressive Sensing and Analog Aggregation

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

Outline

- Introduction
- System Model
- Convergence Analysis
- Minimization of the Error Floor
- Simulation Results and Evaluation
- Conclusion

Outline

- **Introduction**
- System Model
- Convergence Analysis
- Minimization of the Error Floor
- Simulation Results and Evaluation
- Conclusion

Introduction

Background

Conventional centralized machine learning

- Collect data for training
- Big data
- Privacy

Federated learning (FL)

- Keep data locally
- Send weights or gradients
- Dimension curse

Pre-processing of updates

- Quantization
- Sparsification
- Communication censoring
- Proportional to the number of active workers

Analog Aggregation

- Non-coding
- Linear analog modulation
- Over the air computation
- Allow all local workers transmit updates using the same time-frequency resources.

Introduction

- **Motivation**
  - There may be aggregation errors, such as **channel fading**, **noise perturbation**, **sparsification** and so on.
  - How these aggregation errors affect FL?
  - Prior works assume that E(g*g^H)=I, so they can achieve power control like E(p*h*g)<=Pmax → p=Pmax/h.   (So called channel inversion power control)
  - Without local gradients known in advance, how to achieve power control?
  - Simple maximization of the number of participated workers is learning-agnostic and hence not necessarily optimal
  - How to select local workers?

Outline

- Introduction
- **System Model**
- Convergence Analysis
- Minimization of the Error Floor
- Simulation Results and Evaluation
- Conclusion

System Model

- **Federated learning (FL)**
  - Local devices (workers)
    - Receive the current sharing model $\mathbf{w} = [w^1, \ldots, w^D] \in \mathcal{R}^D$ from a parameter server (PS)
    - Use local data to train the received model and get the updates (local gradients, $\mathbf{g}_i$)
    - Send $\mathbf{g}_i$ to the PS
  - PS
    - Receive the updates $\mathbf{g}_i$ and average them $\mathbf{g} = \frac{1}{K}\sum_{i=1}^{U} K_i \mathbf{g}_i$
    - Update the sharing model $\mathbf{w} = \mathbf{w} - \alpha \mathbf{g}$
    - Broadcast the updated sharing model to local workers
- **1-bit compressive sensing and analog aggregation transmission**
  - Sparsification
    - Top-k
  - Dimension Reduction
    - Random Gaussian matrix $\mathbf{\Phi} \in \mathbb{R}^{S \times D}$ $(S \ll D)$
  - Quantization
    - The overall operation $\mathcal{C}(\mathbf{g}_{i,t}) = \mathtt{sign}(\mathbf{\Phi}\mathtt{sparse}_{\kappa}(\mathbf{g}_{i,t}))$
  - Transmission
    - Power control factor $p_{i,t} = \frac{\beta_{i,t} K_i b_t}{h_{i,t}}$

System Model

- **Federated learning (FL)**
  - Local devices (workers)
    - Receive the current sharing model $\mathbf{w} = [w^1, \ldots, w^D] \in \mathcal{R}^D$ from a parameter server (PS)
    - Use local data to train the received model and get the updates (local gradients, $\mathbf{g}_i$)
    - Send $\mathbf{g}_i$ to the PS
  - PS
    - Receive the updates $\mathbf{g}_i$ and average them $\mathbf{g} = \frac{1}{K}\sum_{i=1}^{U} K_i \mathbf{g}_i$
    - Update the sharing model $\mathbf{w} = \mathbf{w} - \alpha \mathbf{g}$
    - Broadcast the updated sharing model to local workers
- **1-bit compressive sensing and analog aggregation transmission**
  - Sparsification
    - Top-k
  - Dimension Reduction
    - Random Gaussian matrix $\mathbf{\Phi} \in \mathbb{R}^{S \times D}$ $(S \ll D)$
  - Quantization
    - The overall operation $\mathcal{C}(\mathbf{g}_{i,t}) = \mathtt{sign}(\mathbf{\Phi}\mathtt{sparse}_{\kappa}(\mathbf{g}_{i,t}))$
  - Transmission
    - Power control factor $p_{i,t} = \frac{\beta_{i,t} K_i b_t}{h_{i,t}}$

- The received signal at PS

$$\mathbf{y}_t = \sum_{i=1}^{U} h_{i,t} p_{i,t} \mathcal{C}(\mathbf{g}_{i,t}) + \mathbf{z}_t$$

- Reconstruction
  - Reconstruct the sparse averaged gradient

System Model

- **1-bit compressive sensing**

$$\mathbf{y} = Q(\Phi\mathbf{x}) = \Phi\mathbf{x} + \mathbf{n}.$$

  - Quantization can be modeled as measurement noise

$$\|\mathbf{n}\|_2 = \left(\sum_i \|n_i\|^2\right)^{1/2} \leq \epsilon.$$

  - A robust reconstruction can be achieved by solving

$$\widehat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x}\|_1 \text{ s.t. } \|\mathbf{y} - \Phi\mathbf{x}\|_2 \leq \epsilon$$

  - Binary iterative hard thresholding (BIHT) algorithm

System Model

- **Aggregation errors**
  - Sparsification errors
  - Quantization errors
  - Additive white Gaussian noise (AWGN)
  - Reconstruction errors

  - **How these aggregation errors affect FL?**
  - **Can we mitigate these errors?**

Outline

- Introduction
- System Model
- **Convergence Analysis**
- Minimization of the Error Floor
- Simulation Results and Evaluation
- 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 (twice-continuously differentiable): $\nabla^2 F(\mathbf{w}_t) \preceq L\mathbf{I}.$
  - Assumption 3 (sample-wise gradient bounded): $\| \nabla f(\mathbf{w}_t) \|^2 \leq \rho_1 + \rho_2 \| \nabla F(\mathbf{w}_t) \|^2$
  - Assumption 4 (local gradient bounded): $\|\mathbf{g}_{i,t}\|^2 \leq G^2, \forall i, t$

Theorem 1. *Given the power scaling factor b_t, worker selection vectors $\beta_{i,t}$, and the learning rate $\alpha = \frac{1}{L}$, we have the following convergence rate at the T-th iteration.*

$$\frac{1}{T}\sum_{t=1}^{T} \| \nabla F(\mathbf{w}_{t-1}) \|^2 \leq \frac{2L}{T(1-\rho_2)}\mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)] + \frac{2L}{T(1-\rho_2)}\sum_{t=1}^{T} B_t, \tag{20}$$

where

$$B_t = \frac{\sum_{i=1}^{U} K_i\rho_1(1-\beta_{i,t})}{2LK\sum_{i=1}^{U} K_i\beta_{i,t}} + \frac{C^2}{2L}\left(1 + (1+\delta)\frac{D-\kappa}{SD}G^2 + \frac{\sigma^2}{\left(\sum_{i=1}^{U} K_i\beta_{i,t}b_t\right)^2}\right)$$
$$+ \sum_{i=1}^{U} \beta_{i,t}(1+\delta)\frac{D-\kappa}{2LD}G^2, \tag{21}$$

and $\mathbf{w}_t$ converges to $\mathbf{w}^*$.

Convergence Analysis

Theorem 1. *Given the power scaling factor b_t, worker selection vectors $\beta_{i,t}$, and the learning rate $\alpha = \frac{1}{L}$, we have the following convergence rate at the T-th iteration.*

$$\frac{1}{T}\sum_{t=1}^{T} \| \nabla F(\mathbf{w}_{t-1}) \|^2 \leq \frac{2L}{T(1-\rho_2)}\mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)] + \frac{2L}{T(1-\rho_2)}\sum_{t=1}^{T} B_t, \tag{20}$$

- From **Theorem 1**, we have

$$\frac{1}{T}\sum_{t=1}^{T} \| \nabla F(\mathbf{w}_{t-1}) \|^2 \leq \frac{2L}{T(1-\rho_2)}\mathbb{E}[F(\mathbf{w}_0) - F(\mathbf{w}^*)] + \frac{2L}{T(1-\rho_2)}\sum_{t=1}^{T} B_t$$
$$\xrightarrow{T\to\infty} \frac{2L}{T(1-\rho_2)}\sum_{t=1}^{T} B_t.$$

- **To mitigate errors, minimize B_t**

Convergence Analysis

Theorem 1. *Given the power scaling factor b_t, worker selection vectors $\beta_{i,t}$, and the learning rate $\alpha = \frac{1}{L}$, we have the following convergence rate at the T-th iteration.*

$$\frac{1}{T} \quad (20)$$

- From **Th**

$$\frac{1}{T}\sum_{t}$$

Introduction

- **Motivation**
  - There may be aggregation errors, such as **channel fading**, **noise perturbation**, **local worker selection**, **sparsification** and so on.
  - How these aggregation errors affect FL?
  - Prior works assume that E(g*g^H)=I, so they can achieve power control like E(p*h*g)<=Pmax → p=Pmax/h.   (So called channel inversion power control)
  - Without local gradients known in advance, how to achieve power control?
  - Simple maximization of the number of participated workers is learning-agnostic and hence not necessarily optimal
  - How to select local workers?

- **To mitigate errors, minimize B_t**

Outline

- Introduction
- System Model
- Convergence Analysis
- **Minimization of the Error Floor**
- Simulation Results and Evaluation
- Conclusion

Minimization of the Error Floor

- To mitigate errors, minimize B_t

$$B_t = \frac{\sum_{i=1}^{U} K_i \rho_1 (1-\beta_{i,t})}{2LK \sum_{i=1}^{U} K_i \beta_{i,t}} + \frac{C^2}{2L}\left(1 + (1+\delta)\frac{D-\kappa}{SD}G^2 + \frac{\sigma^2}{\left(\sum_{i=1}^{U} K_i \beta_{i,t} b_t\right)^2}\right)$$

$$+ \sum_{i=1}^{U} \beta_{i,t}(1+\delta)\frac{D-\kappa}{2LD}G^2,$$

- The PS aims to determine the power scaling factor b_t, and the scheduling indicator $\boldsymbol{\beta}_t = [\beta_{1,t}, \beta_{2,t}, ..., \beta_{U,t}]$, Such a joint optimization problem is formulated as

$$\min_{b_t, \boldsymbol{\beta}_t} \quad B_t$$

$$\text{s.t.} \quad \frac{\beta_{i,t}^2 K_i^2 b_t^2}{h_{i,t}^2} \le P_i^{\mathrm{Max}},$$

$$\beta_{i,t} \in \{0, 1\}, i \in \{1, 2, ..., U\}$$

Minimization of the Error Floor

- **The power control**

$$p_{i,t} = \frac{\beta_{i,t} K_i b_t}{h_{i,t}}$$

$$\mathbf{g} = \frac{1}{K} \sum_{i=1}^{U} K_i \mathbf{g}_i$$

$$\mathbf{y}_t = \sum_{i=1}^{U} h_{i,t} p_{i,t} \mathcal{C}(\mathbf{g}_{i,t}) + \mathbf{z}_t$$

$$\mathbf{y}_t^{desired} = \frac{\sum_{i=1}^{U} K_i \beta_{i,t} \mathcal{C}(\mathbf{g}_{i,t})}{\sum_{i=1}^{U} K_i \beta_{i,t}}$$

$$\mathbf{y}_t = \sum_{i=1}^{U} K_i b_t \beta_{i,t} \mathcal{C}(\mathbf{g}_{i,t}) + \mathbf{z}_t$$

$$\hat{\mathbf{y}}_t^{desired} = \frac{\mathbf{y}_t}{\sum_{i=1}^{U} K_i \beta_{i,t} b_t} = \mathbf{y}_t^{desired} + \frac{\mathbf{z}_t}{\sum_{i=1}^{U} K_i \beta_{i,t} b_t}$$

- **Maximum power limitation**

$$|p_{i,t} c_{i,t}^s|^2 = \left(\frac{\beta_{i,t} K_i b_t}{h_{i,t}} c_{i,t}^s \right)^2 = \frac{\beta_{i,t}^2 K_i^2 b_t^2}{h_{i,t}^2} \leq P_i^{\text{Max}}$$

$$\begin{aligned} \min_{b_t, \boldsymbol{\beta}_t} \quad & B_t \\ \text{s.t.} \quad & \frac{\beta_{i,t}^2 K_i^2 b_t^2}{h_{i,t}^2} \leq P_i^{\text{Max}}, \\ & \beta_{i,t} \in \{0,1\}, i \in \{1,2,...,U\} \end{aligned}$$

where $\mathcal{C}(\mathbf{g}_{i,t}) = [c_{i,t}^1, ..., c_{i,t}^s, ..., c_{i,t}^S]^T$ $\quad c_{i,t}^s = \pm 1$

Minimization of the Error Floor

- **The power control**

$$p_{i,t} = \frac{\beta_{i,t} K_i b_t}{h \ldots}$$

$$\mathbf{g} = \frac{1}{K} \sum_{i=1}^{U} K_i \mathbf{g}_i$$

$$\mathbf{y}_t = \sum_{i=1}^{U} h \ldots$$

$$\ldots \frac{K_i \beta_{i,t} \mathcal{C}(\mathbf{g}_{i,t})}{\sum_{i=1}^{U} K_i \beta_{i,t}}$$

$$\mathbf{y}_t = \sum_{i=1}^{U} \ldots$$

$$\hat{\mathbf{y}}_t^{desired} \ldots$$

Introduction

- **Motivation**
  - There may be aggregation errors, such as **channel fading**, **noise perturbation**, **local worker selection**, **sparsification** and so on.
  - How these aggregation errors affect FL?
  - Prior works assume that E(g*g^H)=I, so they can achieve power control like E(p*h*g)<=Pmax → p=Pmax/h. (So called channel inversion power control)
  - Without local gradients known in advance, how to achieve power control?
  - Simple maximization of the number of participated workers is learning-agnostic and hence not necessarily optimal
  - How to select local workers?

- **Maximu…**

$$|p_{i,t} c_{i,t}^s| = \left(\frac{\ldots}{h_{i,t}} c_{i,t} \right) = \frac{\ldots}{h_{i,t}^2} \leq P_i$$

$$\text{s.t.} \quad \frac{\beta_{i,t}^2 K_i^2 b_t^2}{h_{i,t}^2} \leq P_i^{\text{Max}},$$

$$\beta_{i,t} \in \{0,1\}, i \in \{1,2,\ldots,U\}$$

where $\mathcal{C}(\mathbf{g}_{i,t}) = [c_{i,t}^1, \ldots, c_{i,t}^s, \ldots, c_{i,t}^S]^T$ $\quad c_{i,t}^s = \pm 1$

Minimization of the Error Floor

- Mixed integer programming (MIP)
  - The coupling of b_t and $\boldsymbol{\beta}_t$
  - Non-convex
- Optimal Solution via Discrete Programming
  - Given $\boldsymbol{\beta}_t = [\beta_{1,t}, \beta_{2,t}, ..., \beta_{U,t}]$, the problem is convex.
  - The enumeration-based method may be applicable for a small number of workers, e.g., U ≤ 10
  - The complexity is $\mathcal{O}(2^U)$
- ADMM-based Suboptimal Solution
  - Decomposition
    - Decompose the hard combinatorial problem into U parallel smaller convex problems.
    - Iteratively solve them.
  - The complexity is $\mathcal{O}(U)$

$$\begin{aligned} \min_{b_t, \boldsymbol{\beta}_t} \quad & B_t \\ \text{s.t.} \quad & \frac{\beta_{i,t}^2 K_i^2 b_t^2}{h_{i,t}^2} \leq P_i^{\text{Max}}, \\ & \beta_{i,t} \in \{0,1\}, i \in \{1,2,...,U\} \end{aligned}$$

Outline

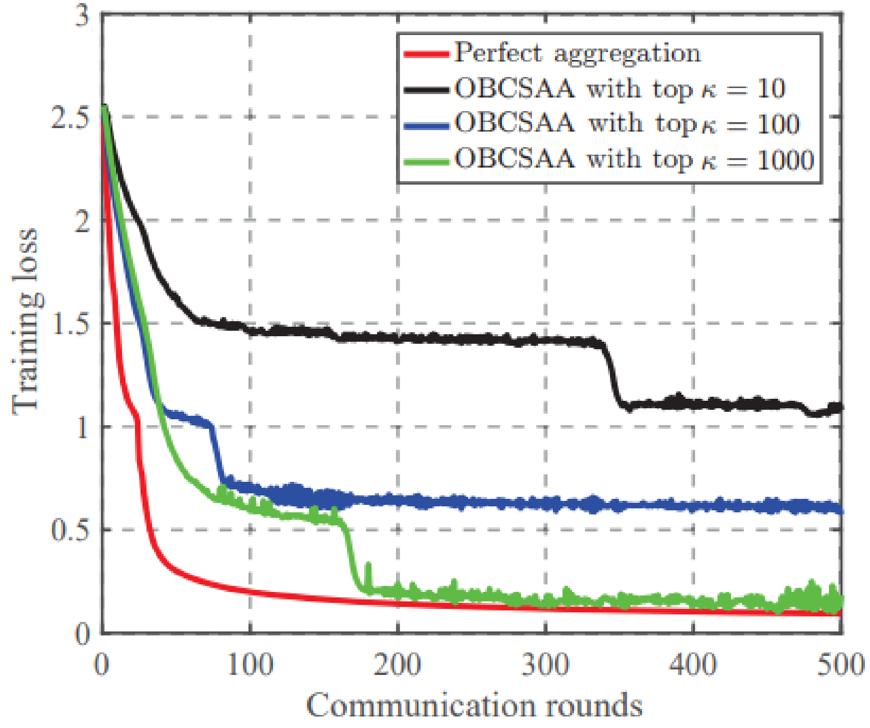
- Introduction
- System Model
- Convergence Analysis
- Minimization of the Error Floor
- **Simulation Results and Evaluation**
- Conclusion

Simulation Results and Evaluation

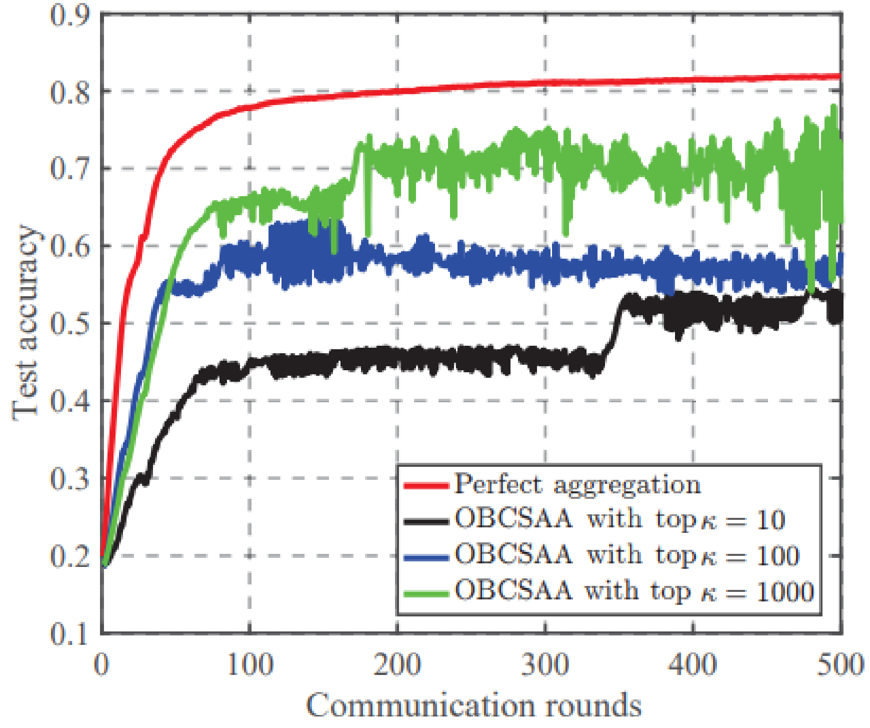
- MNIST dataset
  - 784-neuron input layer, a 64-neuron hidden layer, and a 10-neuron softmax output layer
  - The dimension of the gradient is D=50890
- Our proposed scheme:
  - **O**ne **B**it **C**ompressive **S**ensing and **A**nalog **A**ggregation (OBCSAA)
- Baseline:
  - Perfect aggregation (ideal case): without transmission and compression, PS can obtain exact gradients for aggregation

Simulation Results and Evaluation

- The performance with **Top-k**
  - The sparsity ratio are k/D: 10/50890, 100/50890, 1000/50890.



(a) Training loss

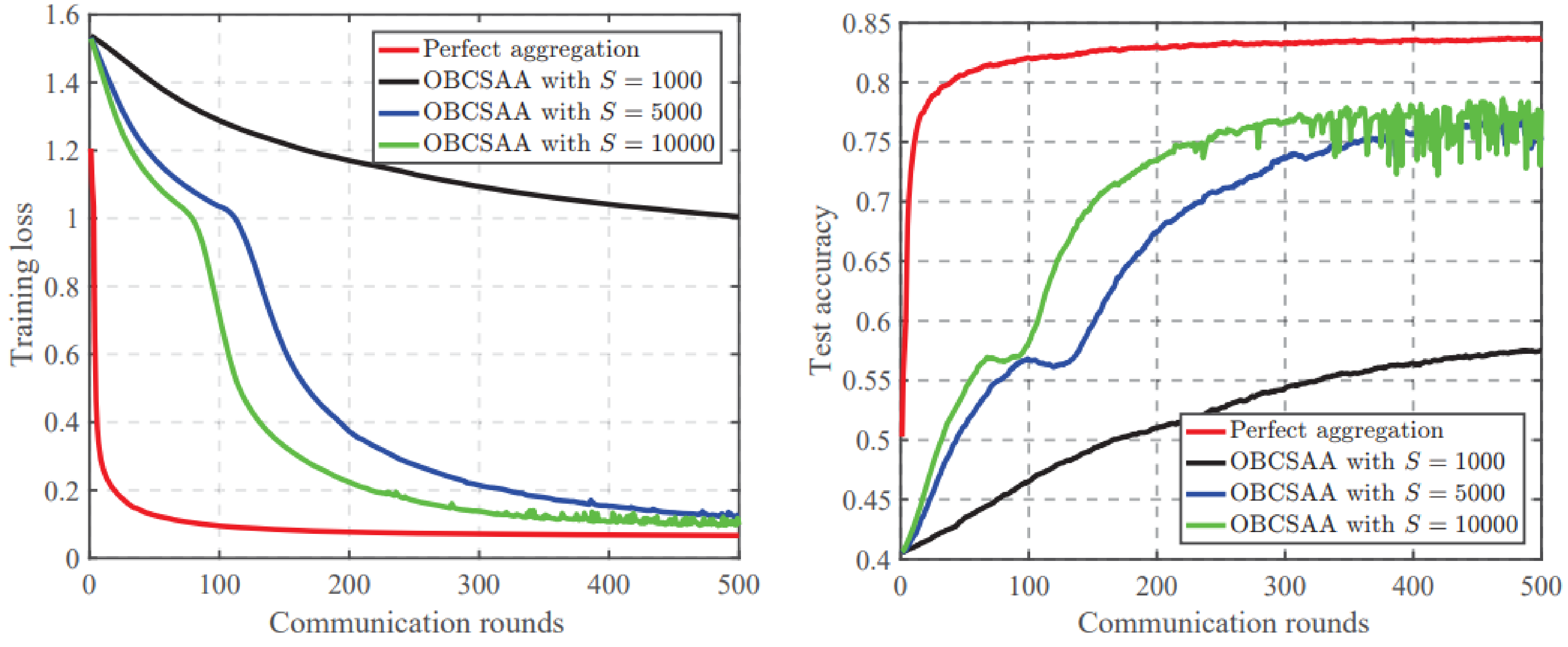


(b) Test accuracy

Fig. 1: The performance of our proposed OBCSAA under different sparsification operators.

Simulation Results and Evaluation

- The performance with **Dimension Reduction**
  - Under S = 5000 and κ = 1000, our OBCSAA only use one wireless channel and 5000/50890 transmission time



(a) Training loss (b) Test accuracy

Fig. 2: The performance of our proposed OBCSAA under different S.

Simulation Results and Evaluation

- The performance with enumeration-based and ADMM-based methods
  - Enumeration can achieve better performance with higher complexity.

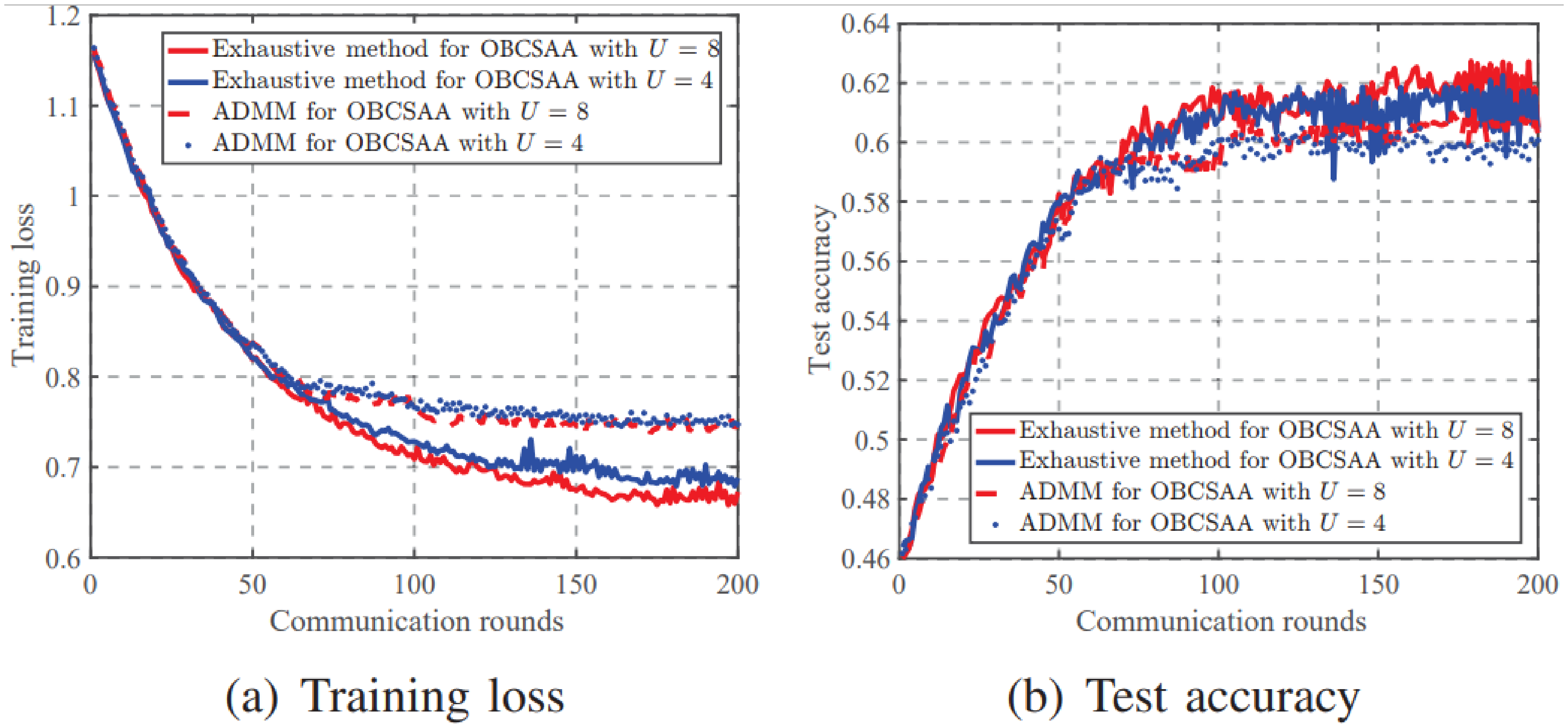


Fig. 3: The performance of joint optimization solving methods for our proposed OBCSAA under different U.

Outline

- Introduction
- System Model
- Convergence Analysis
- Minimization of the Error Floor
- Simulation Results and Evaluation
- **Conclusion**

Conclusion

- We propose a communication-efficient FL based on 1-bit CS and analog aggregation transmissions.
- We derive a closed-form expression for the expected convergence rate of the FL algorithm.
- We formulate a joint optimization problem of communication and learning.
- An enumeration-based method and an ADMM-based method are proposed to solve the non-convex problem, which are suitable for small-scale networks and approximate solution for large-scale networks.

THANK YOU

Xin FAN

Email: fanxin@bjtu.edu.cn