FAFedZO: Faster Zero-order Adaptive Federated Learning Algorithm

Yanbo Lu , Huimin Gao , Yi Zhang , Yong Xu *

Posted Date: 22 January 2025

doi: 10.20944/preprints202501.1650.v1

Keywords: Federated Learning; Zero-Order optimization; Adaptive; convergence

 doi:10.20944/preprints202501.1650.v1

Article

FAFedZO: Faster Zero-Order Adaptive Federated Learning Algorithm

Yanbo Lu, Huimin Gao, Yi Zhang and Yong Xu *

School of Mathematics and Statistics, Henan University of Science and Technology, Luoyang 471023, China
* Correspondence: xuyong@haust.edu.cn

Abstract: Federated Learning represents a newly emerging methodology in the field of machine learning, it enables distributed agents to collaboratively learn a centralized model without sharing their raw data. Some scholars have already proposed many first-order algorithms and second-order algorithms for federated learning to reduce communication costs and speed up convergence. However, these algorithms generally rely on gradient or Hessian information, and we find it difficult to solve such federated optimization problems when the analytical expression of the loss function is not available, that is, when gradient information is not available. Therefore, we employed derivative-free federated zeroth-order optimization in this paper which not rely on specific gradient information, but instead utilizes the changes in function values or model outputs to estimate the optimization direction. Furthermore, to enhance the performance of derivative-free zeroth-order optimization, we propose an effective adaptive algorithm that can dynamically adjust the learning rate and other hyperparameters based on the performance during the optimization process, aiming to accelerate convergence. We rigorously analyze the convergence of our approach, and the experimental findings demonstrate our method indeed can achieve faster convergence speed on the MNIST and CIFAR-10 datasets in cases where gradient information is not available.

Keywords: federated learning; zero-order optimization; adaptive; convergence

1. Introduction

With the rapid development of big data and artificial intelligence technologies, machine learning has become one of the key technologies in various fields. However, in practical applications, issues of data privacy and security have become a non-negligible challenge. In traditional machine learning, data usually needs to be sent to data centers for model training, but in the process of transmission, it faces the problems of data leakage and privacy protection. To address this issue, federated learning(FL) [1], an emerging distributed machine learning technology, has emerged. It achieves collaborative learning among multiple data sources, and while protecting data privacy by exchanging model parameters instead of raw data. Currently, FL is utilized across diverse domains such as autonomous driving [2], personalized recommendation systems [3], and medical informatics [4], among others.

However, existing federated learning methods also have some obvious shortcomings. For instance, some federated learning methods are unable to effectively prevent privacy leaks when faced with complex attack techniques, and the convergence speed of current federated learning algorithms remains a bottleneck. Additionally, designing more efficient optimization algorithms tailored to different types of models and tasks is also an urgent issue that needs to be addressed. Researching how to enhance the privacy protection capabilities of federated learning is of great significance in an environment where data security and privacy protection are increasingly critical. This not only safeguards users' personal information but also promotes the widespread application of federated learning technology across various domains. In practical applications, federated learning needs to meet the demands of various scenarios. By investigating the issue of slow convergence speed, it can better satisfy these practical application needs and enhance the practicality and value of federated learning.

 doi:10.20944/preprints202501.1650.v1

In recent years, the federated optimization for model training has increasingly captured the interest of both academic researchers and industry professionals. To achieve fast convergence rates and reduce communication loads, a wide range of algorithms have been proposed, including first-order methods (such as FedAvg [1], FedFOR [5], FedFomo [6]) and second-order methods (such as Fed-Sophia [7]). Most of these algorithms rely on gradient or Hessian information, and these methods are often only applicable to differentiable functions. However, when the objective function is difficult to compute gradients or the gradient computation cost is excessively high, these algorithms will not be able to solve such problems, such as in the case of black-box functions and in federated hyperparameter tuning [8], so we must look for other solutions. To alleviate the reliance on gradient or Hessian information, the federated zeroth-order optimization (FedZO) algorithm is proposed in [9] through the integration of zeroth-order optimization into the federated learning. The characteristic of derivative-free zeroth-order optimization is that it only queries the objective function values to tackle the problems of federated optimization. It approximates the full gradient and stochastic gradient of the function using function values, without using gradient and Hessian information. In each round of communication, several local updates are performed utilizes a two-point stochastic gradient estimator. Some recent studies such as [10] and [11] have also employed the zeroth-order idea. [10] proposed FedDisco algorithm, leveraging zeroth-order optimization techniques reduces communication overhead significantly. [11] designed a federated zeroth-order algorithm FedZeN, it focused on convex optimization and estimate the curvature of the global objective.

However, existing zeroth-order methods rely solely on function value information, lacking the utilization of gradient and other relevant information. So in complex federated learning scenarios, this method may result in the model failing to fully leverage the feature information of the data, leading to the need for more iterations to achieve convergence during the optimization process. This significantly increases computational time and affects model performance and convergence speed. If the convergence speed of zeroth-order optimization can be accelerated, it can play a better role in scenarios with limited resources such as large-scale distributed optimization and edge computing, improving resource utilization efficiency and thereby promoting the widespread deployment of optimization technologies in resource-sensitive applications.

The emergence of adaptive methods has effectively alleviated the problem of slow convergence. In deep learning, an excessively large learning rate may lead to unstable model training, resulting in overfitting or underfitting. Besides, traditional fixed learning rate methods often struggle to adapt to complex and diverse training data and model structures, potentially leading to slow convergence speed of the model. The adaptive method can dynamically adjust the learning rate based on feedback from the training process, thereby accelerating the convergence of the model. In addition, compared with stochastic gradient descent, this method can also escape saddle points more quickly [12]. Therefore, introducing it in practice represents a crucial approach to enhancing the performance of federated learning algorithms.

However, improper design of adaptive federated learning methods may lead to convergence issues [13]. Reddi et al. [14] first proposed a federated version of the adaptive optimizer, among them are FedAdagrad and FedYogi. However, their analysis is only valid under the condition that $\beta_1 = 0$ (where β_1 is the decay parameter), which cannot leverage the advantages of momentum. In the algorithm FedCAMS [15], it provides a complete proof, but it does not improve the convergence speed. Moreover, they usually require a global learning rate for initialization and adjustment, and the selection of this global learning rate may affect the performance of the algorithm. Existing adaptive methods still have shortcomings such as lagging response to complex environmental changes and high consumption of computational resources. Our goal is to design adaptive algorithms tailored to various real-world scenarios, characterized by efficiency, precision, and flexibility, so as to rapidly adapt to environmental changes, optimize resource allocation, and provide robust support for development across various fields.

 doi:10.20944/preprints202501.1650.v1

Based on the above, we combine the gradient-free optimization with adaptive methods, aiming to leverage the advantages of both to solve the problem of unavailable gradient information of the objective function in a finite-sum optimization problem, while achieving fast convergence and effectively improving the training efficiency and performance of the model.

Contributions The contributions of this paper are summarized as follows:

- By combining the zero-order optimization and adaptive gradient method, we have proposed a novel faster zero-order adaptive federated learning algorithm, called FAFedZO, it can eliminate the reliance on gradient information and accelerate convergence at the same time.
- We conducted a theoretical analysis of the proposed zeroth-order adaptive algorithm and provided a convergence analysis framework under some mild assumptions, demonstrating its convergence. Additionally, we have analyzed the computational complexity and convergence rate of the algorithm.
- Extensive comparative experimental results under both IID and non-IID settings have confirmed the efficacy of our proposed FAFedZO algorithm on the MNIST and CIFAR-10 datasets. when gradient information is not available, it can achieve faster convergence speed.

The structure of the rest of this paper is outlined below. Section 2 provides a summary of the related work. In Section 3, we present the formulation of the federated optimization problem and the algorithm framework of FAFedZO. Section 4 provides the convergence analysis of FAFedZO. Section 5 presents the outcomes of our experiments. Section 6 discusses the beneficial applications of our method in the real world. The paper concludes with Section 7.

2. Related Work

Federated Learning

There have been numerous studies on federated learning, and the pioneering work on federated learning began with [1], where proposed the algorithm called FedAvg. After FedAvg, numerous additional first-order schemes emerged, such as FedNova [16], FedProx [17], SCAFFOLD [18], FedSplit [19], and FedPD [20]. Among them, SCAFFOLD employs control variates to rectify "client drift" while maintaining the same sampling and communication complexity as FedAvg. FedProx introduced a penalty-based approach, which can reduced communication complexity to $O(\epsilon^{-2})$. To further minimize communication costs, various second-order optimization methods have been introduced, including GIANT [21] and FedDANE [22]. Additionally, there are also some FL algorithms based on momentum, including [23–25]. The work presented in [23] proposes a momentum fusion approach for synchronizing the server and local momentum buffers, however, it does not aim to decrease complexity. [24] proposed a momentum-based global update algorithm, Fed-GLOMO, which reduces variance on the server side using variance reduction techniques. [25] proposed the STEM algorithm, it employs momentum-assisted stochastic gradient directions for updates at both worker nodes and the central server.

Recent research [26,27] has also focused on the application of federated learning in satellite–terrestrial integrated networks. [26] proposes a federated split learning framework, which aims to leverage the computing capabilities of satellite and terrestrial networks for collaborative processing of sequential data while protecting data privacy. [27] presents an intrusion detection method that combines federated learning with conditional generative adversarial networks to address the challenges of data privacy and imbalanced datasets. However, there are still numerous federated optimization problems in reality that are challenging to solve, for instance when gradient information is unavailable or costly to acquire. Therefore, the study of gradient-free zeroth-order optimization is essential.

Zero-order Optimization

Early literatures that began to use the zero-order idea for estimation include [28–30]. Specifically, In [28], the authors developed a distributed zero-order algorithm utilizing gradient tracking techniques. In [29], the author provides the first generalization error analysis for black-box learning via derivative-free optimization, and demonstrates that under the assumption of Lipschitz and smoothing (unknown) losses, the ZoSS method attains a comparable generalization error boundary to Stochastic Gradient

 doi:10.20944/preprints202501.1650.v1

Descent (SGD). [30] proposed and analyzed zero-order stochastic approximation algorithms for non-convex and convex objective function, focusing on solving the problems of constrained optimization and high-dimensional settings. Recent key research achievements have also applied zero-order methods to various fields, such as [9–11] and [31]. In [9], a derivative-free algorithm FedZO is proposed. It is proven that this algorithm achieves a linear increase in speed in relation to the number of devices involved and local iteration times under non-convex settings. Under the framework of cross-device federated learning, [31] introduces a dual-communication zeroth-order method, which is the first technique to incorporate wireless channels into the algorithm. It employs a single-point gradient estimator, replacing long vectors with scalar values for communication while leveraging the characteristics of wireless channels. This represents a significant new achievement in the field of zeroth-order optimization, achieving a convergence rate of $O(\frac{1}{\sqrt[3]{K}})$ in non-convex settings. However, due to the nature of zeroth-order optimization, it can lead to slower convergence speeds. Therefore, an adaptive method is introduced below.

Adaptive Methods

Early proposed adaptive algorithms include [32,33]. The Adam method in [32] introduces decay coefficients and combines momentum optimization to better adapt to non-stationary data and large-scale datasets, thereby accelerating model convergence. The AdaGrad method in [33] can adaptively adjust the learning rate of each parameter that can makes sparse features to obtain larger learning rates, while frequently occurring features obtain smaller learning rates. Then researchers have extended these algorithms to the context of federated learning, like recent studies such as [32-35]. In [34], the authors designed an Adaptive Local Iteration Differential Privacy Federated Learning algorithm (ALI-DPFL) and demonstrated its superiority in resource-constrained scenarios. The adaptive methods in [35] and [36] effectively mitigate the issue of non-IID data, achieving critical progress in enabling better performance when dealing with non-IID data. In [35] proposed an adaptive fairness federated aggregation algorithm Ada-FFL, which dynamically adjusts fairness coefficients based on local model updates to ensure both convergence performance of the global model and fairness among clients. [36] introduced a novel framework called FedARF, aimed at enhancing federated learning performance through adaptively reconstructing local features during the training process. It adopts an adaptive feature fusion strategy, enabling the model to better adapt to the data distribution of each client, thereby accelerating the convergence speed on non-IID data. [37] introduced a novel federated learning algorithm named AFedAvg, which significantly reduced communication costs and accelerated convergence speed by combining adaptive communication frequency and gradient sparsity techniques. This algorithm dynamically selects the communication frequency for the next round based on the number of sparse parameters. The results show that it can achieve a communication compression ratio ranging from 2.4 times to 23.1 times, which marks a significant step forward in mitigating the obstacles posed by high communication volume to artificial intelligence. However, none of these studies have investigated the integration of adaptive methods within the context of zero-order optimization. Hence, our article addresses this by combining zero-order optimization with adaptive approaches to conduct our discussion.

3. Problem Formulation and Algorithm Design

In this part, we will introduce the federated optimization problem and the design of the zero-order adaptive federated optimization algorithm FAFedZO.

3.1. Federated Optimization Problem Formulation

We consider a federated learning task involving a central server and Q edge devices with an index of $\{1, 2, ...Q\}$. The central server aims to facilitate collaboration among these devices to address a specific optimization problem

 doi:10.20944/preprints202501.1650.v1

$$\min_{\Xi \in R^d} f(\Xi) \triangleq \frac{1}{Q} \sum_{i=1}^{Q} f_i(\Xi), \tag{1}$$

and

$$f_i(\Xi) \triangleq E_{\vartheta_i \sim D_i}[F_i(\Xi, \vartheta_i)], \tag{2}$$

here, $\Xi \in R^d$ represents a d-dimensional model parameter. For each edge device i, $f_i(\Xi)$ denotes its local loss function, while $f(\Xi)$ stands for the global loss function. In formula (1), $f_i(\Xi)$ evaluates the anticipated risk on the data distribution D_i on the edge device i, which is presented in formula (2). $\vartheta_i \sim D_i$ denotes the random variable ϑ_i is uniformly drawn from the distribution D_i, and $F_i(\Xi, \vartheta_i)$ denotes the loss function of ϑ_i at the parameter Ξ.

3.2. Algorithm Design of FAFedZO

We explored how to design a method that combines adaptive gradient with zero-order optimization in federated learning.

Firstly, we expand our algorithm description within the context of the FedAvg framework. We are focused here on solving problem (1) through zeroth-order optimization methods and propose a novel zero-order fast adaptive FL method (FAFedZO) employs a shared adaptive learning rate to address the issues. In particular, Algorithm 1 outlines the specifics of our FAFedZO method.

At the beginning, input the parameters and perform initialize. For all i, compute $\Xi_{1,i} = \Xi_{0,i} - \eta_0 n_{0,i}$, which represents the first model update based on the initial parameters and the initial gradient estimates.

Then, for each edge device i, perform the following steps: First, extract a mini-batch $\mathfrak{B}_{t,i}$ of size b from the local dataset D_i. Next, compute the stochastic gradient estimates $\hat{g}_{t,i}$ and $\hat{g}_{t-1,i}$ for the current model parameters $\Xi_{t,i}$ based on the mini-batch $\mathfrak{B}_{t,i}$. Here, we elaborate on the specific method for estimating the gradients.

To address the issue of unavailable gradient information and reduce the frequency of model exchange, we achieve this by using gradient estimators and performing stochastic zero-order updates in each communication round. In particular, at the t-th round, edge device i calculates a two-point stochastic gradient estimator [28] as outlined below

$$\widetilde{\nabla}_v^{\mu} F_i(\Xi_{t,i}, \vartheta_{t,i}) = \frac{d v_t}{\mu_t} \Big(F_i(\Xi_{t,i} + \mu_t v_t, \vartheta_{t,i}) - F_i(\Xi_{t,i}, \vartheta_{t,i}) \Big), \tag{3}$$

where $\Xi_{t,i}$ denote the local model of edge device i, while $\vartheta_{t,i}$ signifies a random variable drawn by edge device i according to its local data distribution D_i during the t-th round. v_t represents a randomly chosen d-dimensional direction, uniformly sampled from the unit sphere $\mathbb{S}^d$, while μ_t stands for a positive step size.

When computing gradient estimates, we do not directly use precise gradient information but instead approximate the gradient through random sampling and estimation. This zero-order optimization approach does not require the exact calculation of function derivatives, rather, it optimizes based on sampling and estimation of function values.

Afterward, we update local models according to the following stochastic zeroth-order update method

$$\Xi_{t+1,i} = \Xi_{t,i} - \eta_t \widetilde{\nabla}_v^{\mu} F_i(\Xi_{t,i}, \vartheta_{t,i}), t = 1, 2, \cdots, T$$

where η_t denotes the learning rate.

 doi:10.20944/preprints202501.1650.v1

In step 9 of Algorithm 1, we use $n_{t,i}$ to update the model, where $n_{t,i}$ is the momentum-based variance reduced gradient estimator, is the weighted sum of the current gradient estimate and previous gradient estimates, with the weights determined by χ_t. Its definition is as follows

$$n_{t,i} = \widetilde{\nabla}_v^{\mu} F_i(\Xi_{t,i}; \vartheta_{t,i}) + (1 - \chi_t)(n_{t-1,i} - \widetilde{\nabla}_v^{\mu} F_i(\Xi_{t-1,i}; \vartheta_{t,i})), \tag{4}$$

where the hyperparameter $\chi_t \in (0, 1)$, representing the decay factor.

In step 10 of Algorithm 1, we incorporate the coordinate-wise adaptive learning rate approach, akin to that utilized in Adam [32], which is defined as follows

$$\iota_{t,i} = \varrho \iota_{t-1,i} + (1 - \varrho)(\widetilde{\nabla}_v^{\mu} F_i(\Xi_{t,i}; \vartheta_{t,i}))^2, \tag{5}$$

where the hyperparameter $\varrho \in (0, 1)$, representing the decay factor.

Then, when the current iteration number t is an integer multiple of the local update number p (i.e., $\mod(t, p) = 0$), we perform step 12.

In step 12 of Algorithm 1, for $\iota_{t,i}$, we perform averaging and aggregation step and resulting in $\bar{\iota}_t$. Subsequently, we utilize the $\bar{\iota}_t$ to create an adaptive matrix $K_t = diag(\sqrt{\bar{\iota}_t} + \rho)$(where $diag$ denotes diagonal matrix), and where $\rho > 0$ (ρ represent tuning parameter).

At the step 13 in algorithm 1, perform averaging and aggregation on all nodes of $n_{t,i}$.

In step 14, the global model is updated based on the obtained K_t and $\bar{n}_t$.

Otherwise, proceed to steps 15 and 16 of the algorithm. Keep K_t as its previous value K_{t-1} and update the global model in the same manner as before.

In step 16 of algorithm 1, the identical K_t is employed for the local updates across various edge devices.

Here, model parameters, $n_{t,i}$, $\iota_{t,i}$ are aggregated by the global server every interval of p steps.

At the step 17 in algorithm 1, all edge devices calculate the model difference of their local models in the current round, that is, $\Delta_{t,i} = \Xi_{T,i} - \Xi_{0,i}$ ($i \in [Q]$, $[Q]$ represents all Q edge devices from 1 to Q), uploaded to the central server.

At the step 18, 19 in Algorithm 1, after receiving the local model difference, central server summarizes it as $\Delta_t = \frac{1}{Q} \sum_{i \in [Q]} \Delta_{t,i}$, and proceeds to update the global model, that is $\Xi_{t+1} = \Xi_t + \Delta_t$.

Finally, after all iterations are completed, the algorithm outputs a model $\bar{\Xi}$, which is uniformly randomly selected from all the global models $\{\bar{\Xi}_t\}_{t=1}^{T}$ obtained during the iterations, as the final result.

In the algorithm, K_t is a diagonal matrix whose diagonal elements are related to $\bar{\iota}_t$, which is computed based on certain statistics from the local model update process. Specifically, $\iota_{t,i}$ is related to the square of the gradient estimates, and $\bar{\iota}_t$ is obtained by averaging $\iota_{t,i}$ across all devices. Subsequently, K_t is updated accordingly. This approach allows K_t to adaptively adjust based on the variations in gradient information during the local model update process.

During the model training process, the data distribution on different devices may vary, leading to differing updates in model parameters. By adaptively updating K_t, it is possible to influence the aggregation and update methods of the global model. For instance, when the gradient variation on a particular device is significant (i.e., $\bar{\iota}_t$ is large), the value of K_t will also change accordingly, thereby assigning different weights to the amount of updates from this device when aggregating local updates. This enables the algorithm to better adapt to the data distribution and model training situations on different devices, accelerating the convergence speed of the model and enhancing the final model performance.

 doi:10.20944/preprints202501.1650.v1

Algorithm 1 FAFedZO Algorithm

Input: the number of iterations T, decay factor χ_t, ϱ, learning rate η_t, the number of local updates p, mini batch size b and initial batch-size B;
initialize: Initialize: $\Xi_{0,i} = \bar{\Xi}_0 = \frac{1}{Q}\sum_{i=1}^{Q}\Xi_{0,i}$, $n_{0,i} = \bar{n}_0 = \frac{1}{Q}\sum_{i=1}^{Q}\hat{n}_{0,i}$ with $\hat{n}_{0,i} = \widetilde{\nabla}_v^{\mu}F(\Xi_{0,i}, \mathfrak{B}_{0,i})$ and $\iota_{0,i} = \bar{\iota}_0 = \frac{1}{Q}\sum_{i=1}^{Q}\hat{\iota}_{0,i}$ with $\hat{\iota}_{0,i} = (\widetilde{\nabla}_v^{\mu}F(\Xi_{0,i}, \mathfrak{B}_{0,i}))^2$ where $\mid \mathfrak{B}_{0,i} \mid = B$ from D_i for $i \in [Q]$. $K_0 = diag(\sqrt{\bar{\iota}_0} + \rho)$
$\Xi_{1,i} = \Xi_{0,i} - \eta_0 n_{0,i}$, for all $i \in [Q]$
for $t = 1, 2, \ldots, T$ **do**
  for *For edge device* $i \in [Q]$ **do**
    Extract mini-batch samples $\mathfrak{B}_{t,i} = \{\vartheta_i^j\}_{j=1}^{b}$, $\mid \mathfrak{B}_{t,i} \mid = b$ from D_i locally
    Calculate local stochastic gradient estimator $\hat{g}_{t,i} = \widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i}, \vartheta_{t,i})$, $\hat{g}_{t-1,i} = \widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i}, \vartheta_{t,i})$ according to (3)
    Perform local update $\Xi_{t+1,i} = \Xi_{t,i} - \eta_t\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i}, \vartheta_{t,i})$
    $n_{t,i} = \hat{g}_{t,i} + (1-\chi_t)(n_{t-1,i} - \hat{g}_{t-1,i})$
    $\iota_{t,i} = \varrho\iota_{t-1,i} + (1-\varrho)\hat{g}_{t,i}^2$
    if $\mod(t, p) = 0$ **then**
      $\iota_{t,i} = \bar{\iota}_t = \frac{1}{Q}\sum_{i=1}^{Q}\iota_{t,i}$ and $K_t = diag(\sqrt{\bar{\iota}_t} + \rho)$
      $n_{t,i} = \bar{n}_t = \frac{1}{Q}\sum_{i=1}^{Q}n_{t,i}$
      $\Xi_{t+1,i} = \bar{\Xi}_{t+1} = \frac{1}{Q}\sum_{i=1}^{Q}(\Xi_{t,i} - \eta_t K_t^{-1}n_{t,i})$
      else
      $K_t = K_{t-1}$
      $\Xi_{t+1,i} = \Xi_{t,i} - \eta_t K_t^{-1}n_{t,i}$
    end
    Calculate the local model update $\Delta_{t,i} = \Xi_{T,i} - \Xi_{0,i}$, then central server get the uploaded local model updates.
  end
  Aggregate local changes $\Delta_t = \frac{1}{Q}\sum_{i\in[Q]}\Delta_{t,i}$
  Update the global model $\Xi_{t+1} = \Xi_t + \Delta_t$
end
Output: $\bar{\Xi}$ chosen uniformly random from $\{\bar{\Xi}_t\}_{t=1}^{T}$.

4. Convergence Analysis of FAFedZO Method

For this part, the convergence of FAFedZO method will be discussed. To facilitate the theoretical analysis of the proposed algorithm, we need to make some assumptions as follows.

Assumption 1. *The global loss function $f(\Xi)$ in (1) has a lower bound, that is, there is a fixed value f^* that exists $f(\Xi) \geq f^* > -\infty$.*

The Assumption 1 means that regardless of the value of the optimization variable Ξ, the global loss function $f(\Xi)$ will not decrease indefinitely, and there exists a minimum value f^*, ensuring that the optimization problem is meaningful.

Assumption 2. *We suppose that function $F_i(\Xi, \vartheta_i)$, $f_i(\Xi)$, $f(\Xi)$ are all L-smooth, that is, for all $\Xi_1 \in R^d, \Xi_2 \in R^d$, we can conclude that*

$$\|\nabla f_i(\Xi_1) - \nabla f_i(\Xi_2)\| \leq L\|\Xi_1 - \Xi_2\|,$$
$$f(\Xi_1) \leq f(\Xi_2) + \langle \nabla f(\Xi_2), \Xi_1 - \Xi_2\rangle + \frac{L}{2}\|\Xi_1 - \Xi_2\|^2.$$

The Assumption 2 is common in the theoretical analysis of non-convex optimization, indicating that the gradient changes of $f_i(\Xi)$, $f(\Xi)$ are smooth and do not change suddenly, which is crucial for proving the convergence of algorithms.

 doi:10.20944/preprints202501.1650.v1

Assumption 3. *$\nabla F_i(\Xi, \vartheta_i)$ serves as a precise approximation of $\nabla f_i(\Xi)$ without bias, namely*

$$E_{\vartheta_i}[\nabla F_i(\Xi, \vartheta_i)] = \nabla f_i(\Xi), \ \ \forall \Xi \in R^d.$$

The Assumption 3 allows us to use sample gradients to approximate the true gradient.

Assumption 4. *The variance of the stochastic gradient is limited within a certain range, that is, there have a fixed value σ_g can satisfies*

$$E_{\vartheta_i}\|\nabla F_i(\Xi, \vartheta_i) - \nabla f_i(\Xi)\|^2 \leq \sigma_g^2, \ \ \forall \Xi \in R^d.$$

The Assumption 4 indicates that the variance of stochastic gradients will not be infinite, which is important for controlling noise and uncertainty in the optimization process.

Assumption 5. *The difference between each local loss function and the global loss function remains within a certain limit, that is, there have a fixed value σ_h can satisfies*

$$\|\nabla f(\Xi) - \nabla f_i(\Xi)\|^2 \leq \sigma_h^2, \ \ \forall \Xi \in R^d.$$

The Assumption 5 ensures that local optimization does not lead to a significant decrease in global performance.

Assumption 6. *The inter-node variance is bounded, namely*

$$\|\nabla f_i(\Xi) - \nabla f_j(\Xi)\|^2 \leq \zeta^2, \ \ \forall \Xi \in R^d.$$

The Assumption 6 indicates that the differences between loss functions on different nodes are limited, ensuring that the optimization processes on different nodes do not diverge too much.

Assumption 7. *In our algorithms, for all $t \geq 1$, the adaptive matrices K_t satisfies the condition that*

$$\lambda_{min}(K_t) \geq \rho > 0,$$

there ρ represents an appropriate positive value.

The Assumption 7 guarantees the step size in the optimization process does not become too small, thereby guaranteeing the convergence of the algorithm.

Assumption 8. *The gradients of function $f_i(\Xi)$ is G-bounded, that is, for all i, $\Xi \in R^d$, we have*

$$\|\nabla f_i(\Xi)\| \leq G.$$

The Assumption 8 means that the gradients of the loss functions on each node are not infinite, ensuring stability and controllability in the optimization process.
We define the ϵ-stationary point as follows:

Definition 1. *A point Ξ is called ϵ-stationary point if $\|\nabla f(\Xi)\| \leq \epsilon$. Generally, a stochastic algorithm is defined to achieve an ϵ-stationary point in T iterations if $\mathbb{E}\|\nabla f(\Xi_T)\| \leq \epsilon$.*

Then, we investigate the convergence characteristics of our novel method based on Assumptions 1-8. The Appendix A contains the comprehensive proofs of our results.

About the convergence of the FAFedZO algorithm, we present the main results of this paper as follows:

 doi:10.20944/preprints202501.1650.v1

Theorem 1. *Assuming that the sequence* $\{\bar{\Xi}_t\}_{t=1}^{T}$ *is produced by Algorithm 1. Based on the aforementioned Assumptions 1-8, given that* $\forall t \geq 0$, $\chi_{t+1} = c \cdot \eta_t^2$, $c = \frac{1}{12Lp\bar{h}^3\rho^2} + \frac{30L^2}{\rho^2} \leq \frac{60L^2}{\rho^2}$, $w_t = \max(\frac{3}{2}, 1728L^3p^3\bar{h}^3 - t)$, $\bar{h} = \frac{1}{L}$ *and set*

$$\eta_t = \frac{\rho\bar{h}}{(w_t + t)^{1/3}} \tag{6}$$

we can conclude that

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\bar{\Xi}_t)\| \leq P\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\mathcal{Q}_t]} \\ \leq P\Bigg[\left[\frac{12Lp}{\rho^2T} + \frac{L}{\rho^2T^{2/3}}\right]\mathbb{E}[f(\bar{\Xi}_0) - f^*] + \frac{6L^2\mu^2p^2}{\rho^2QT} \\ + \left[\frac{12^2\times 75p}{\rho^2T} + \frac{900}{\rho^2T^{2/3}}\right]\left[\frac{5L^2\mu^2}{3} + 3\zeta^2\right](\ln T + 1) \\ + \frac{L^2\mu^2p}{2\rho^2QT^{2/3}}\Bigg]^{1/2}\end{aligned} \tag{7}$$

where $P = 5\sqrt{2d(G^2 + \sigma_g^2) + 2\rho^2 + \frac{1}{2}d^2L^2\mu^2}$

Remark 1. *We utilize* ζ *as an indicator of data heterogeneity. The final results demonstrate that an increase in* ζ *(indicating greater data heterogeneity) leads to a slowdown in the training process. In addition, as the parameters* L *and* μ *increase, the boundary in the theorem conclusion will also increase, which will also result in a decrease in the training speed.*

Remark 2. *A suitable value of* ρ *ensures a balanced incorporation of adaptive information in the learning rate. In practice, we commonly select* ρ *to be within the order of* $O(1)$*, steering clear of excessively small or large values.*

Analysis of computational complexity and convergence speed
(computational complexity) For $\mathcal{Q}_t = \frac{1}{12\eta_t^2}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 + \frac{1}{4\rho^2}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\mathcal{Q}_t] &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{12\eta_t^2}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 + \frac{1}{4\rho^2}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2\right] \\ &\leq \left[\frac{12Lp}{\rho^2T} + \frac{L}{\rho^2T^{2/3}}\right]\mathbb{E}[f(\bar{\Xi}_0) - f^*] + \frac{6L^2\mu^2p^2}{\rho^2QT} + \frac{L^2\mu^2p}{2\rho^2QT^{2/3}} \\ &+ \left[\frac{12^2\times 75p}{\rho^2T} + \frac{900}{\rho^2T^{2/3}}\right]\left[\frac{5L^2\mu^2}{3} + 3\zeta^2\right](\ln T + 1)\end{aligned}$$

without loss of generality, we let $\frac{p}{Q} = O(1)$, and choose $p = T^{\frac{1}{3}}$. To make the right side of the inequality less than ϵ^2, we can get $T = O(\epsilon^{-3})$ and $\frac{T}{p} = T^{\frac{2}{3}} = \epsilon^{-2}$. Therefore, to satisfy the definition of an ϵ-stationary point, which is $\mathbb{E}\|\nabla f(\Xi_T)\| \leq \epsilon$ and $\mathbb{E}[\mathcal{Q}_t] \leq \epsilon^2$, we obtain the total sample cost as $O(\epsilon^{-3})$ and the communication round as $O(\epsilon^{-2})$.

(convergence speed) In the FedZO algorithm, when full devices participate, it can satisfy

$$\begin{aligned}\min_{t\in[T]}\mathbb{E}\left\|\nabla f\left(\boldsymbol{x}^t\right)\right\|^2 \leq & 4\frac{f\left(\boldsymbol{x}^0\right) - f_*}{HT\eta} + \eta\frac{24dL}{N}\left(\sigma_g^2 + 3\sigma_h^2\right) \\ & + \frac{dL^2\mu^2}{12} + 4L^2\mu^2,\end{aligned}$$

 doi:10.20944/preprints202501.1650.v1

when partial devices participate, it can satisfy

$$\min_{t\in[T]} \mathbb{E}\left\|\nabla f\left(\boldsymbol{x}^t\right)\right\|^2 \leq 4\frac{f\left(\boldsymbol{x}^0\right)-f_*}{HT\eta}+\eta\frac{32dL}{M}(\sigma_g^2+3\sigma_h^2) + \eta\frac{36HL\sigma_h^2}{M}+24\eta HL^3\mu^2+\frac{dL^2\mu^2}{12}+4L^2\mu^2.$$

Regardless of the situation, we can observe that the FedZO algorithm achieves the convergence rate of $O(\frac{1}{T})$. Furthermore, as indicated by Theorem 1, our algorithm can achieve the convergence rate of $O(\frac{1}{T^{2/3}})$. Therefore, our FAFedZO method theoretically possesses a faster convergence rate compared to general zero-order methods. This further validates the advantages of our approach.

5. Experimental Results

For this part, we are conducting comparative experiments on the MNIST and CIFAR-10 datasets, present some experimental outcomes to assess the performance of the proposed FAFedZO method in federated black-box attacks, thereby confirming the advantages of this algorithm.

Experimental environment and datasets

Experimental environment configuration: This study utilizes the FedZO framework as the experimental baseline to conduct research on federated learning algorithms. This framework is renowned for its exceptional scalability and operational convenience, supporting numerous prominent federated learning algorithms. For this study, Python version 3.10.13 is adopted as the programming language, and PyTorch serves as the development platform. All tests are conducted on a Windows 10 platform equipped with an NV GTX1650 GPU and CUDA version 12.4.99.

Dataset: By conducting comparative experimental studies on the MNIST and CIFAR-10 datasets respectively, we have validated the practicality of the FAFedZO algorithm while significantly enhancing the performance of the model.

MNIST (Modified National Institute of Standards and Technology) is a widely used benchmark dataset for evaluating and comparing the performance of handwritten digit recognition algorithms. It was adapted by Professor Yann LeCun and his colleagues at New York University in 1998 from an original dataset of the National Institute of Standards and Technology (NIST). This dataset comprises a collection of grayscale images of handwritten digits (0-9) with 60,000 training samples and 10,000 test samples, each image measuring 28×28pixels. Each image has undergone a preprocessing pipeline, including centering and normalization, aimed at enhancing the accuracy of classification algorithms. The MNIST dataset is extensively utilized for training and testing image processing and machine learning algorithms, particularly as a classic dataset for introductory deep learning and neural network research. An example of this dataset is shown in Figure 1.

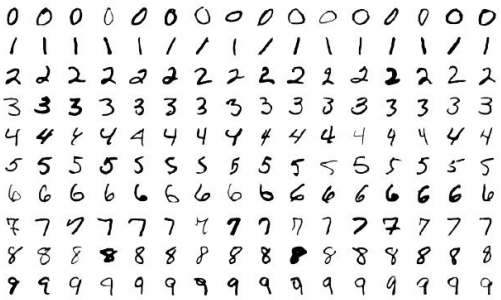


Figure 1. Example of the MNIST dataset.

CIFAR-10 (Canadian Institute For Advanced Research-10) is a commonly used dataset in computer vision, particularly for image classification tasks. It is consist of 60,000 32x32 color images across 10 categories and each category contains 6,000 images, and the dataset is divided into 50,000 training and 10,000 test images. Due to its smaller image size and fewer categories, it is often used for rapid

 doi:10.20944/preprints202501.1650.v1

verification, prototype development, as well as learning and understanding the basics of various computer vision algorithms. It is also employed as a benchmark dataset for deep learning models to assess their performance and generalization ability on image classification tasks.

Experimental result analysis

Here, we present the results of simulation experiments to assess the performance of the FAFedZO method in the context of black-box attack strategies.

Given the characteristics of black-box scenarios, optimizing black-box attacks falls into the realm of zero-order optimization. We investigate black-box attacks on a trained deep neural network(DNN) classification model. The purpose of our experiment is to train an interference image with the same size as the image in the dataset, which makes it difficult for the human eyes to recognize the difference between the original image and the adversarial image after adding the interference image, but it can induce the classification model to make a wrong judgment. We want to achieve a higher attack success rate with as little disturbance as possible, so we consider the loss function as shown below:

$$\begin{aligned}&\psi(\Xi,\vartheta)\\&=\underbrace{\max\left\{\varphi_{y_\vartheta}\left(\frac{1}{2}tanh\left(tanh^{-1}2\vartheta+\Xi\right)\right)-\max_{j\neq y_\vartheta}\left\{\varphi_j\left(\frac{1}{2}tanh\left(tanh^{-1}2\vartheta+\Xi\right)\right)\right\},0\right\}}_{I}\\&+c\underbrace{\left\|\frac{1}{2}tanh\left(tanh^{-1}2\vartheta+\Xi\right)-\vartheta\right\|^2}_{II}\end{aligned}\tag{8}$$

Here Ξ represents the interference image to be trained, ϑ represents the original image in the dataset, and y_ϑ represents the label corresponding to the image ϑ (for example, the label of the image "deer" is 4) , $\varphi_j(\vartheta)$ represents the confidence that the classification model recognizes image ϑ as label j. I in (8) formula measures the probability of attack failure (marked as attack loss), II represents the image distortion caused by disturbance, and c is the balance coefficient. In this way our goal can be achieved by minimizing $\psi(\Xi,\vartheta)$. Next we use the (8) formula to construct the local loss function.

We divide the samples in the dataset into Q groups randomly and unevenly without repetition, and then distribute them to each edge device, where Q is the total count of edge devices we preset. Then for all edge devices we define its local loss function as:

$$f_i(\Xi)\triangleq E_{\vartheta_i\sim D_i}[\psi(\Xi,\vartheta_i)],i\in\{1,2,...,Q\}\tag{9}$$

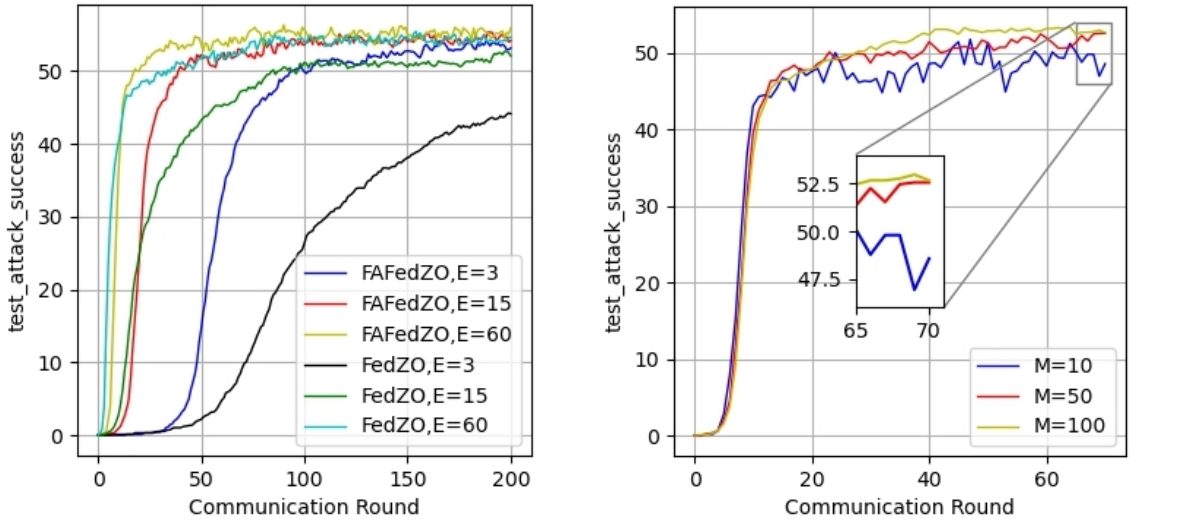
Where D_i represents the dataset at the ith edge device. In this way, the federal black-box attack problem on the DNN classification model can be formulated as a federated optimization problem: $f(\Xi)\triangleq\frac{1}{Q}\sum_{i=1}^{Q}f_i(\Xi)$. Next, we use the FAFedZO algorithm proposed in this paper to solve this problem.

We select balance parameter $c=1$. And for the remaining parameters, we select them as $(b_1,b_2,\eta,\mu)=(25,20,0.1,0.001)$.

In Figure 2, we demonstrate the change in the accuracy of the federated black-box attack as the number of communication rounds increases. Figure (a) illustrates the impact of the account of local updates $E\in\{3,15,60\}$ on attack accuracy of the proposed FAFedZO algorithm when the total number of edge devices $Q=50$ and the number of participating edge devices $M=30$. We also compare it with the FedZO algorithm. It can be observed that the larger the E value, the higher the attack accuracy of the algorithm. Furthermore, the attack accuracy of FAFedZO is significantly higher than that of FedZO. When $E=60$, they achieve comparable accuracy, proving that the algorithm proposed in this paper outperforms the original FedZO algorithm. Figure (b) shows how the convergence performance

 doi:10.20944/preprints202501.1650.v1

of the FAFedZO method is affected by the number of participating edge devices $M \in \{10, 50, 100\}$ when $Q = 100$ and $E = 60$. It can be seen that by adjusting the M value, the FAFedZO algorithm can effectively enhance the attack accuracy.

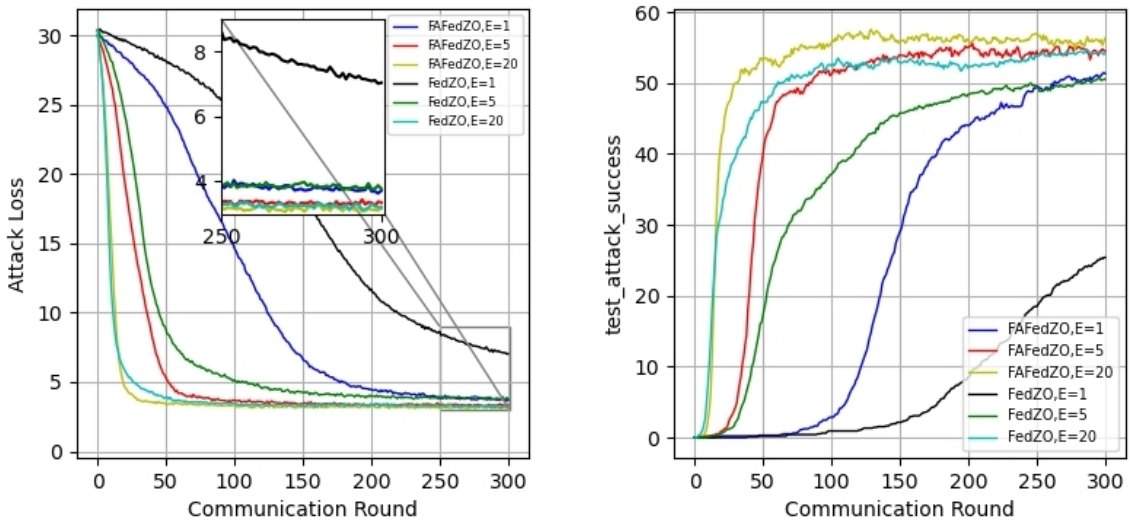


(**a**) Influence of varying the number of local updates that select $(Q, M) = (50, 30)$.

(**b**) Influence of the number of participating edge devices that select $(Q, E) = (100, 60)$

Figure 2. Federated black-box attack accuracy against the number of communication rounds.

In Figure 3, we present the relationships between attack loss and communication rounds as well as attack accuracy and communication rounds, respectively. The influence of the number of local updates $E \in \{1, 5, 20\}$ on the convergence performance of the algorithm is studied under the condition of $Q = 50$ and $M = 50$, i.e., all devices participate. It is evident that the FAFedZO scheme is capable of significantly decreasing the attack loss and improve attack accuracy compared to the FedZO algorithm under different E values. Moreover, as the value of E increases, the convergence speed of the FAFedZO algorithm also accelerates, with lower attack loss and higher attack accuracy, both of which tend to stabilize faster.



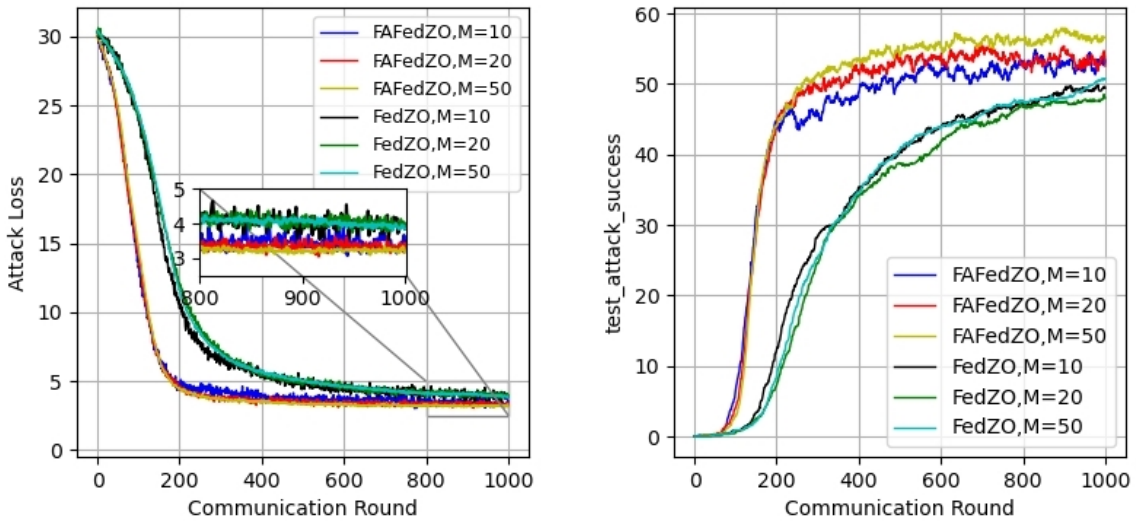
(**a**) Attack loss.

(**b**) Testing accuracy.

Figure 3. Influence of the number of local updates that select $Q = 50$ and $M = 50$.

 doi:10.20944/preprints202501.1650.v1

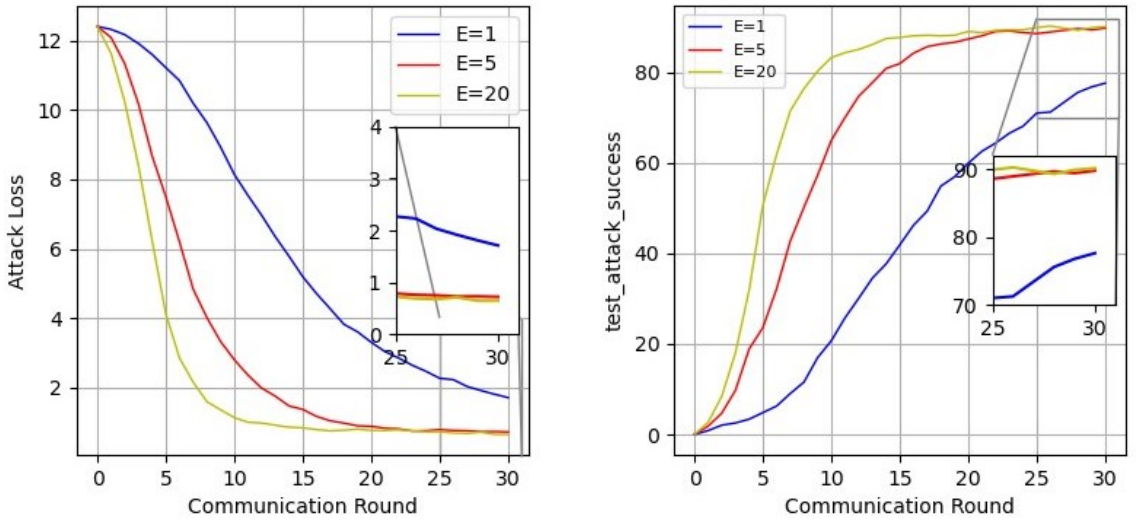
Figure 4 also shows the relationship between attack loss, attack accuracy, and communication rounds. However, we study how the number of participating edge devices $M \in \{10, 20, 50\}$ influences the convergence performance of the scheme under the condition of $Q = 50$ and $E = 1$. It is easy to see that the FAFedZO algorithm is far superior to the FedZO algorithm in terms of both attack loss and accuracy. Even when the FedZO algorithm takes the optimal value of M among 10, 20, and 50, its performance is still weaker than the worst value of the FAFedZO algorithm. In addition, as M increases, the attack loss value decreases, and the accuracy increases, which is in line with our expectations.



(**a**) Attack loss. (**b**) Testing accuracy.

Figure 4. Influence of the number of participating edge devices that select $Q = 50$ and $E = 1$.

The aforementioned experiments were all conducted on the MNIST dataset, and similar conclusions were also drawn from experiments performed on the CIFAR-10 dataset. As shown in Figure 5, under the conditions of $Q = 50$ and $M = 50$, the impact curves of the number of local updates $E \in \{1, 5, 20\}$ on attack loss and accuracy are presented. It is evident that by adjusting the E value, the performance of the FAFedZO algorithm can also be significantly enhanced.

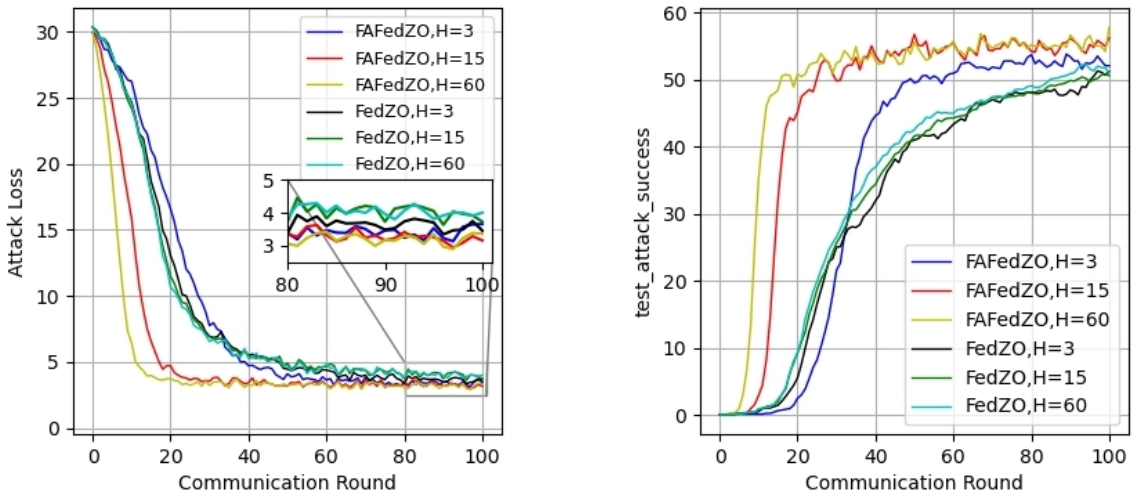


(**a**) Attack loss. (**b**) Testing accuracy.

Figure 5. Influence of the number of local updates that select $Q = 50$ and $M = 50$.

 doi:10.20944/preprints202501.1650.v1

In addition to investigating the impact of the number of local updates and the number of participating edge devices on the algorithm's performance, we also studied the influence of varying the number of random directions on the proposed algorithm. As shown in Figure 6, on the MNIST dataset, under the conditions of $Q = 50$ and $M = 10$, we present the curves illustrating the impact of the number of directions $H \in \{3, 15, 60\}$ on attack loss and attack accuracy. It can be observed that we have similar conclusions to the above figures, which further confirms the superiority of the FAFedZO algorithm.

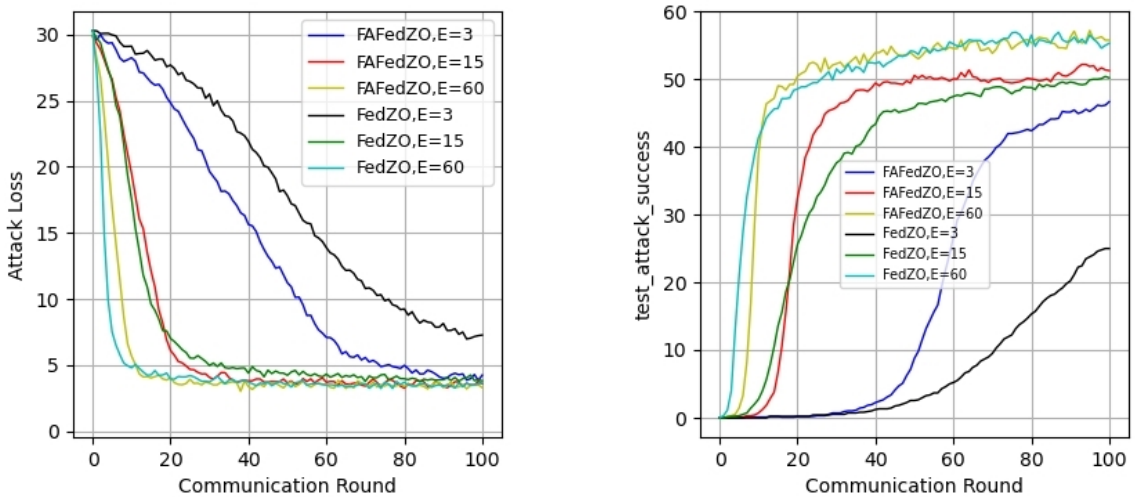


(**a**) Attack loss. (**b**) Testing accuracy.

Figure 6. Influence of the number of directions that select $Q = 50$ and $M = 10$.

In addition, we also discuss the performance of our algorithm under the setting of non-independent and identically distributed (non-iid) data.

Figure 7 presents the impact of the number of local updates on the attack loss and accuracy of the algorithm under the non-iid setting. It can be observed that, compared with Figure 2, the attack accuracy in this case is lower than that under the iid setting, which is attributed to the influence of the non-iid setting and aligns with our expectations. Furthermore, the results also indicate that under the non-iid setting, the performance of the FAFedZO algorithm remains superior to that of the FedZO algorithm, and both achieve comparable performance levels when $E = 60$.

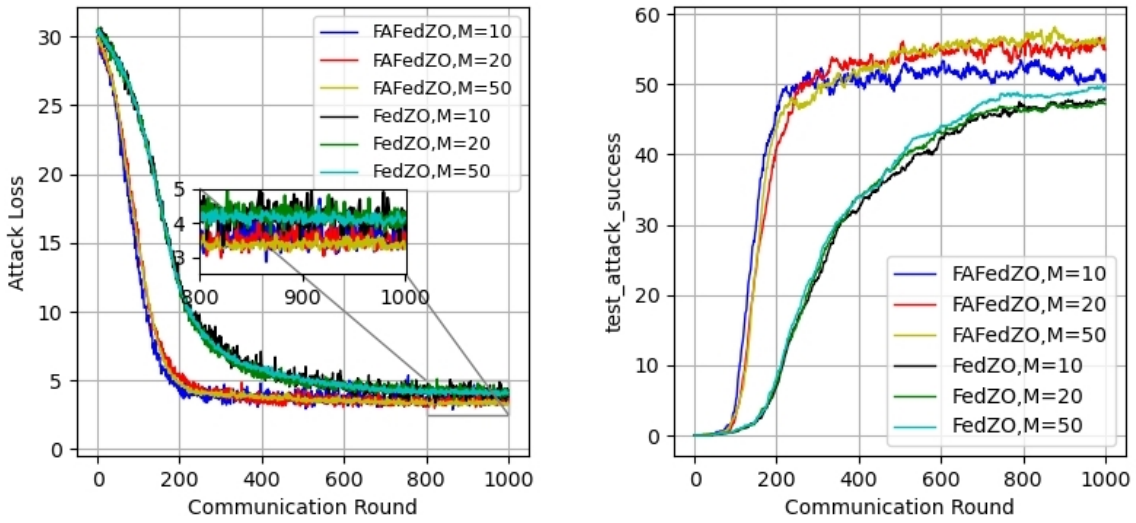


(**a**) Attack loss. (**b**) Testing accuracy.

Figure 7. The impact of different numbers of local updates when selecting $Q = 50$ and $M = 30$ in a non-IID setting.

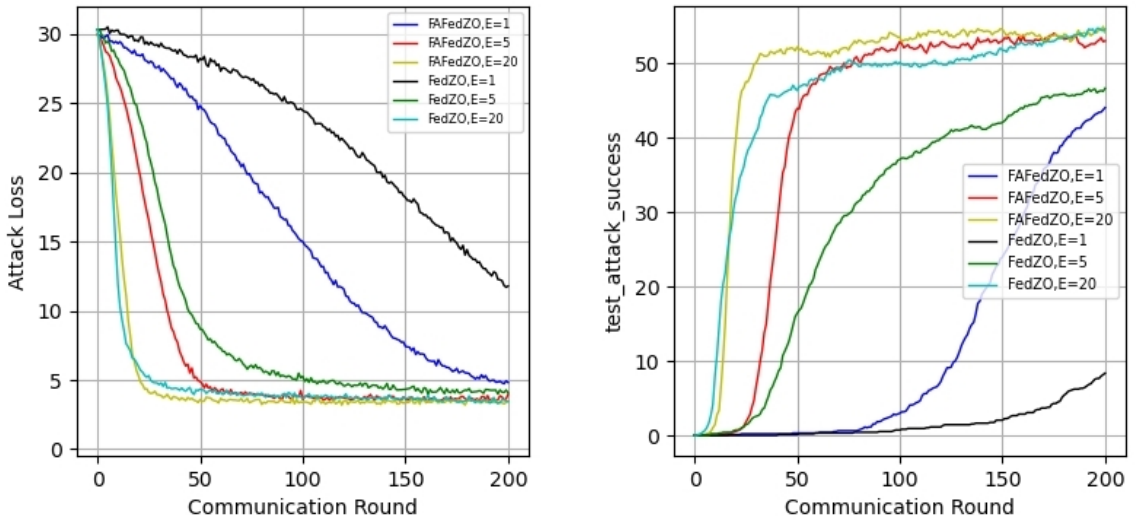
 doi:10.20944/preprints202501.1650.v1

The conclusions we previously mentioned are also supported by Figures 8 and 9. Therefore, in summary, the FAFedZO algorithm demonstrates superior performance in both iid and non-iid environments.



(**a**) Attack loss. (**b**) Testing accuracy.

Figure 8. The impact of the number of participating edge devices when selecting $Q = 50$ and $E = 1$ under the non-IID setting.



(**a**) Attack loss. (**b**) Testing accuracy.

Figure 9. The impact of different numbers of local updates when selecting $Q = 50$ and $M = 50$ under the non-IID setting.

6. Discussion

Our approach combines zeroth-order optimization with adaptive techniques, yielding significant benefits in various real-world applications. For instance, in healthcare, medical data contains sensitive patient information and varies across different regions and hospitals. Utilizing this method allows the model to automatically adapt to these variations during training. Additionally, zeroth-order optimization can be leveraged to explore better solutions, enhancing the model's generalization capability across diverse medical scenarios and improving its assistance in disease diagnosis. In the financial sector, financial institutions need to accurately identify risks such as fraudulent transactions. By employing this method, risk characteristics can be adaptively captured from vast amounts of financial transaction data, and zeroth-order optimization can be used for efficient optimization, thereby improving the accuracy of risk identification and preventing financial risks. In the realm of IoT

 doi:10.20944/preprints202501.1650.v1

devices, where resources are limited, zeroth-order optimization eliminates the need for gradient computation, reducing computational load. Adaptive methods can adjust training strategies based on device resources and data characteristics, enabling federated learning to operate efficiently on IoT devices and enhancing their data processing capabilities.

7. Conclusion

In this paper, we proposed FAFedZO, a zero-order federated optimization method based on an adaptive approach with a shared adaptive learning rate. In the non-convex setting, when gradient information is not available, FAFedZO can estimates the first-order gradient information through function queries to optimize the objective function. We have combined the advantages of zero-order optimization and adaptability. Zero-order estimation can solve the problem that the objective function is not available, thus the gradient information is not usable. Adaptability can accelerate the performance of the algorithm. Furthermore, our theoretical analysis of the algorithm demonstrates the convergence of FAFedZO. Finally, we conducted extensive experiments to verify the effectiveness of the proposed algorithm again on the MNIST and CIFAR-10 datasets, respectively.

Author Contributions: All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Key Technologies Research and Development Program of Henan Province under Grant No. 242102210102.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare there is no conflicts of interest.

Appendix A

Here, we present a thorough examination of our algorithm's convergence properties. To facilitate the analysis, we introduce the notation $g_{t,i}$ that $g_{t,i} = \nabla f_i(\Xi_{t,i})$ in the subsequent sections. Define s_t as $s_t = \lfloor t/p \rfloor p$ and use $\otimes$ as Kronecker product symbol. Below are some basic lemmas required for our analysis.

At first, for the zeroth-order gradient $\widetilde{\nabla}_v^{\mu} F_i(\Xi_{t,i}, \vartheta_{t,i})$, based on the characteristics of the gradient estimator outlined in [[38], Lemma 4.2], we can deduce that

$$\mathbb{E}\left[\widetilde{\nabla}_v^{\mu} F_i(\Xi_{t,i}, \vartheta_{t,i})\right] = \nabla f_i^{\mu}(\Xi_{t,i}) \tag{A1}$$

Then, we can get

$$\begin{aligned}
&\mathbb{E}\left\|\widetilde{\nabla}_v^{\mu} F_i(\Xi_{t,i}, \vartheta_{t,i}) - \nabla f_i^{\mu}(\Xi_{t,i})\right\|^2 \\
\leq &\mathbb{E}\left\|\widetilde{\nabla}_v^{\mu} F_i(\Xi_{t,i}, \vartheta_{t,i})\right\|^2 \\
\leq &2d\mathbb{E}\|\nabla F_i(\Xi_{t,i}, \vartheta_{t,i})\|^2 + \frac{1}{2}d^2L^2\mu^2 \\
= &2d\mathbb{E}\|\nabla F_i(\Xi_{t,i}, \vartheta_{t,i}) - \nabla f_i(\Xi_{t,i}) + \nabla f_i(\Xi_{t,i})\|^2 + \frac{1}{2}d^2L^2\mu^2 \\
= &2d\mathbb{E}\|\nabla F_i(\Xi_{t,i}, \vartheta_{t,i}) - \nabla f_i(\Xi_{t,i})\|^2 + 2d\mathbb{E}\|\nabla f_i(\Xi_{t,i})\|^2 + \frac{1}{2}d^2L^2\mu^2 \\
\leq &2d\sigma_g^2 + 2d\mathbb{E}\|\nabla f_i(\Xi_{t,i})\|^2 + \frac{1}{2}d^2L^2\mu^2 \\
\leq &2d(G^2 + \sigma_g^2) + \frac{1}{2}d^2L^2\mu^2
\end{aligned} \tag{A2}$$

 doi:10.20944/preprints202501.1650.v1

where the second equality is due to Assumption 3, the establishment of these four inequalities is because $\mathbb{E}\|z - \mathbb{E}z\|^2 \leq \mathbb{E}\|z\|^2$, [[38], Lemma 4.1] and Assumption 4, Assumption 8.

Thus, we can get

$$\begin{aligned}&\mathbb{E}\left\|\widetilde{\nabla}_v^{\mu} F_i(\Xi_{t,i}, \vartheta_{t,i}) - \nabla f_i^{\mu}(\Xi_{t,i})\right\|^2\\ \leq& 2d(G^2 + \sigma_g^2) + \frac{1}{2}d^2L^2\mu^2\end{aligned}$$

At the same time, we can also obtain

$$\begin{aligned}&\mathbb{E}\left\|\widetilde{\nabla}_v^{\mu} F_i(\Xi_{t,i}, \vartheta_{t,i})\right\|^2\\ \leq& 2d(G^2 + \sigma_g^2) + \frac{1}{2}d^2L^2\mu^2\end{aligned} \tag{A3}$$

The following Lemma A1 is a small conclusion throughout the proof.

Lemma A1 ([39]). *Given $\Xi_t \in \mathbb{R}^{Qd}$ and $\bar{\Xi}_t \in \mathbb{R}^d$, there $\mathbf{1} \in \mathbb{R}^Q$ represents the vector composed entirely of ones, we can get conclusion that*

$$\|\Xi - \mathbf{1} \otimes \bar{\Xi}\|^2 \leq \|\Xi\|^2 \tag{A4}$$

Then we continue to delve into the upper bound of the adaptive matrices K_t.

Lemma A2. *Suppose the adaptive matrices sequence $\{K_t\}_{t=1}^T$ is derived from Algorithm. On the Basis of Assumptions 1 to 8, we can conclude that*

$$\mathbb{E}\|K_t\|^2 \leq 2d(G^2 + \sigma_g^2) + 2\rho^2 + \frac{1}{2}d^2L^2\mu^2 \tag{A5}$$

Proof. Firstly, we know that

$$\mathbb{E}\|K_t\|^2 = \mathbb{E}\|K_{s_t}\|^2 = \mathbb{E}\|\sqrt{\bar{\iota}_{s_t}} + \rho\|_\infty^2 \leq 2\mathbb{E}\|\bar{\iota}_{s_t}\|_\infty + 2\rho^2 \tag{A6}$$

there $K_t = \mathrm{diag}(\sqrt{\bar{\iota}_t} + \rho)$ is a diagnoal matrix. And following the definition of ι_{s_t}, we can deduce that

$$\begin{aligned}\mathbb{E}\|\bar{\iota}_{s_t}\|_\infty &= \mathbb{E}\|\varrho\bar{\iota}_{s_t-1} + (1-\varrho)\frac{1}{Q}\sum_{i=1}^{Q}[\widetilde{\nabla}_v^{\mu} F_i(\Xi_{s_t-1,i}, \vartheta_{s_t-1,i})]^2\|_\infty\\ &\leq \varrho\mathbb{E}\|\bar{\iota}_{s_t-1}\|_\infty + (1-\varrho)\frac{1}{Q}\sum_{i=1}^{Q}\mathbb{E}\|[\widetilde{\nabla}_v^{\mu} F_i(\Xi_{s_t-1,i}, \vartheta_{s_t-1,i})]^2\|_\infty\\ &\leq \varrho\mathbb{E}\|\bar{\iota}_{s_t-1}\|_\infty + (1-\varrho)\frac{1}{Q}\sum_{i=1}^{Q}\mathbb{E}\|\widetilde{\nabla}_v^{\mu} F_i(\Xi_{s_t-1,i}, \vartheta_{s_t-1,i})\|_2^2\\ &\leq \varrho\mathbb{E}\|\bar{\iota}_{s_t-1}\|_\infty + (1-\varrho)\frac{1}{Q}\sum_{i=1}^{Q}[2d(G^2 + \sigma_g^2) + \frac{1}{2}d^2L^2\mu^2]\\ &= \varrho\mathbb{E}\|\bar{\iota}_{s_t-1}\|_\infty + (1-\varrho)[2d(G^2 + \sigma_g^2) + \frac{1}{2}d^2L^2\mu^2]\\ &\triangleq \varrho\mathbb{E}\|\bar{\iota}_{s_t-1}\|_\infty + (1-\varrho)C\end{aligned} \tag{A7}$$

 doi:10.20944/preprints202501.1650.v1

Therefore, because of $\varrho \in (0,1)$, taking the recursive expansion we can get

$$\begin{aligned}\mathbb{E}\|\bar{\iota}_{s_t}\|_\infty &\leq (1-\varrho)C + \varrho(1-\varrho)C + \varrho^2(1-\varrho)C + \cdots + \varrho^{s_t-1}(1-\varrho)C \\ &= \frac{1-\varrho^{s_t}}{1-\varrho}(1-\varrho)C \\ &\leq \frac{1}{2}C \end{aligned} \tag{A8}$$

So we finally got

$$\begin{aligned}\mathbb{E}\|K_t\|^2 &\leq 2\mathbb{E}\|\bar{\iota}_{s_t}\|_\infty + 2\rho^2 \\ &\leq 2\rho^2 + C \\ &= 2d(G^2 + \sigma_g^2) + 2\rho^2 + \frac{1}{2}d^2L^2\mu^2\end{aligned}$$

Lemma A2 is proved. □

The following Lemma A3 measures the boundary of variance between gradients.

Lemma A3. *For $i \in [Q]$, $[Q]$ represents all Q edge devices from 1 to Q, we can conclude that*

$$\mathbb{E}\|\widetilde{\nabla}_v^\mu F_i(\Xi_{t,i}; \mathfrak{B}_{t,i}) - g_{t,i}\|^2 \leq L^2\mu^2 \tag{A9}$$

$$\mathbb{E}\|g_t - \mathbf{1}\otimes\bar{g}_t\|^2 \leq 6L^2\mathbb{E}\|\Xi_t - \mathbf{1}\otimes\bar{\Xi}_t\|^2 + 3Q\zeta^2 \tag{A10}$$

Proof. For inequality (A9), we have

$$\begin{aligned}&\mathbb{E}\|\widetilde{\nabla}_v^\mu F_i(\Xi_{t,i}; \mathfrak{B}_{t,i}) - g_{t,i}\|^2 \\ =&\mathbb{E}\|\frac{1}{b}\sum_{\vartheta_{t,i}\in\mathfrak{B}_{t,i}}(\nabla f_i^\mu(\Xi_{t,i};\vartheta_{t,i}) - g_{t,i})\|^2 \\ =&\frac{1}{b^2}\mathbb{E}\|\sum_{\vartheta_{t,i}\in\mathfrak{B}_{t,i}}(\nabla f_i^\mu(\Xi_{t,i};\vartheta_{t,i}) - g_{t,i})\|^2 \\ \leq&\frac{1}{b}\sum_{\vartheta_{t,i}\in\mathfrak{B}_{t,i}}\mathbb{E}\|\nabla f_i^\mu(\Xi_{t,i};\vartheta_{t,i}) - g_{t,i}\|^2 \\ \leq&L^2\mu^2\end{aligned} \tag{A11}$$

where the first inequality is because $\left\|\sum_{i=1}^q a_i\right\|^2 \leq q\sum_{i=1}^q\|a_i\|^2$, the last inequality follows by [[28], Lemma 5.2].

 doi:10.20944/preprints202501.1650.v1

For inequality (A10), we have

$$
\begin{aligned}
&\mathbb{E}\|g_t - \mathbf{1}\otimes \bar{g}_t\|^2 \\
=&\sum_{i=1}^{Q}\mathbb{E}\|g_{t,i} - \bar{g}_t\|^2 \\
\leq& 3\sum_{i=1}^{Q}\mathbb{E}[\|g_{t,i} - \nabla f_i(\bar{\Xi}_t)\|^2 + \frac{1}{Q}\sum_{j=1}^{Q}\|\nabla f_j(\bar{\Xi}_t) - g_{t,j}\|^2 + \frac{1}{Q}\sum_{j=1}^{Q}\|\nabla f_i(\bar{\Xi}_t) - \nabla f_j(\bar{\Xi}_t)\|^2] \\
=&3\sum_{i=1}^{Q}\mathbb{E}\|g_{t,i} - \nabla f_i(\bar{\Xi}_t)\|^2 + 3\sum_{i=1}^{Q}\frac{1}{Q}\sum_{j=1}^{Q}\mathbb{E}\|\nabla f_j(\bar{\Xi}_t) - g_{t,j}\|^2 \\
&+3\sum_{i=1}^{Q}\frac{1}{Q}\sum_{j=1}^{Q}\mathbb{E}\|\nabla f_i(\bar{\Xi}_t) - \nabla f_j(\bar{\Xi}_t)\|^2 \\
\leq& 3L^2\sum_{i=1}^{Q}\mathbb{E}\|\Xi_{t,i} - \bar{\Xi}_t\|^2 + 3L^2\sum_{j=1}^{Q}\mathbb{E}\|\bar{\Xi}_t - \Xi_{t,j}\|^2 \\
&+3\sum_{i=1}^{Q}\frac{1}{Q}\sum_{j=1}^{Q}\mathbb{E}\|\nabla f_i(\bar{\Xi}_t) - \nabla f_j(\bar{\Xi}_t)\|^2 \\
\leq& 6L^2\mathbb{E}\|\Xi_t - \mathbf{1}\otimes\bar{\Xi}_t\|^2 + 3Q\zeta^2
\end{aligned}
\tag{A12}
$$

the last two inequalities here are attributed to Assumption 2, 6.
Thus lemma A3 is proved. □

Below, we study the bounds under different values of $\sum_{i=1}^{Q}\|\Xi_{t,i} - \bar{\Xi}_t\|^2$, which is important for the proof of the subsequent theorems.

Lemma A4. *Given that $t \in [\lfloor t/p\rfloor p, \lfloor t/p\rfloor(p+1)]$, Ξ_t is derived from the Algorithm, we can conclude that*

(1) if $t = s_t = \lfloor t/p\rfloor p$, then we can get

$$\sum_{i=1}^{Q}\left\|\Xi_{s_t,i} - \bar{\Xi}_{s_t}\right\|^2 = 0 \tag{A13}$$

(2) if $t \geq \lfloor t/p\rfloor p$, then we can get

$$\sum_{i=1}^{Q}\|\Xi_{t,i} - \bar{\Xi}_t\|^2 \leq (p-1)\sum_{s=s_t}^{t-1}\eta_s^2\sum_{i=1}^{Q}\|K_s^{-1}(n_{s,i} - \bar{n}_s)\|^2 \tag{A14}$$

Proof.

(1) $t = s_t = \lfloor t/p\rfloor p$, then we can conclude that mod $(t, p) = 0$, so there is $\Xi_{t+1,i} = \bar{\Xi}_{t+1}$, thus we have obtained the final result.

(2) we have

$$\Xi_{t,i} = \Xi_{s_t,i} - \sum_{s=s_t}^{t-1}\eta_s K_s^{-1} n_{s,i} \qquad \bar{\Xi}_t = \bar{\Xi}_{s_t} - \sum_{s=s_t}^{t-1}\eta_s K_s^{-1}\bar{n}_s$$

 doi:10.20944/preprints202501.1650.v1

thus

$$\begin{aligned}\sum_{i=1}^{Q}\|\Xi_{t,i}-\bar{\Xi}_t\|^2 &= \sum_{i=1}^{Q}\left\|\Xi_{s_t,i}-\bar{\Xi}_{s_t}-\left(\sum_{s=s_t}^{t-1}\eta_s K_s^{-1}n_{s,i}-\sum_{s=s_t}^{t-1}\eta_s K_s^{-1}\bar{n}_s\right)\right\|^2\\ &= \sum_{i=1}^{Q}\left\|\sum_{s=s_t}^{t-1}\eta_s K_s^{-1}[n_{s,i}-\bar{n}_s]\right\|^2\\ &\le (t-s_t)\sum_{s=s_t}^{t-1}\eta_s^2\sum_{i=1}^{Q}\|K_s^{-1}(n_{s,i}-\bar{n}_s)\|^2\\ &\le (p-1)\sum_{s=s_t}^{t-1}\eta_s^2\sum_{i=1}^{Q}\|K_s^{-1}(n_{s,i}-\bar{n}_s)\|^2 \end{aligned} \tag{A15}$$

Lemma A4 is proved. □

The following lemmas A5-A7 are crucial conclusions that will be used in proving the final theorem.

Lemma A5. *Assuming the sequence $\{\Xi_t\}_0^T$ is generated by the Algorithm, then we can get*

$$\begin{aligned}\mathbb{E}f(\bar{\Xi}_{t+1}) \le \mathbb{E}f(\bar{\Xi}_t) &- (\frac{3\rho}{4\eta_t}-\frac{L}{2})\mathbb{E}\|\bar{\Xi}_{t+1}-\bar{\Xi}_t\|^2-\frac{\eta_t}{4\rho}\mathbb{E}\|\nabla f(\bar{\Xi}_t)-\bar{n}_t\|^2+\frac{5\eta_t}{2\rho}\mathbb{E}\|\tilde{g}_t-\bar{n}_t\|^2\\ &+\frac{5\eta_t L^2}{2\rho Q}\mathbb{E}\|\Xi_t-\mathbf{1}\otimes\bar{\Xi}_t\|^2\end{aligned} \tag{A16}$$

Proof.

$$\begin{aligned}f(\bar{\Xi}_{t+1}) &\le f(\bar{\Xi}_t)+\langle\nabla f(\bar{\Xi}_t),\bar{\Xi}_{t+1}-\bar{\Xi}_t\rangle+\frac{L}{2}\|\bar{\Xi}_{t+1}-\bar{\Xi}_t\|^2\\ &= f(\bar{\Xi}_t)+\underbrace{\langle\nabla f(\bar{\Xi}_t)-\bar{n}_t,\bar{\Xi}_{t+1}-\bar{\Xi}_t\rangle}_{(1)}+\underbrace{\langle\bar{n}_t,\bar{\Xi}_{t+1}-\bar{\Xi}_t\rangle}_{(2)}+\frac{L}{2}\|\bar{\Xi}_{t+1}-\bar{\Xi}_t\|^2\end{aligned} \tag{A17}$$

Regarding (1), we can obtain

$$\begin{aligned}(1) &= \langle\nabla f(\bar{\Xi}_t)-\bar{n}_t,\bar{\Xi}_{t+1}-\bar{\Xi}_t\rangle\\ &\le \|\nabla f(\bar{\Xi}_t)-\bar{n}_t\|\|\bar{\Xi}_{t+1}-\bar{\Xi}_t\|\\ &\le \frac{\eta_t}{\rho}\|\nabla f(\bar{\Xi}_t)-\bar{n}_t\|^2+\frac{\rho}{4\eta_t}\|\bar{\Xi}_{t+1}-\bar{\Xi}_t\|^2\end{aligned} \tag{A18}$$

Regarding term (2), $K_t=\mathrm{diag}(\sqrt{\bar{\iota}_{s_t}}+\rho)$, $\bar{\Xi}_{t+1}=\bar{\Xi}_t-\eta_t K_t^{-1}\bar{n}_t$. We can obtain based on the given definition of K_t and assumption 7 that

$$\begin{aligned}\langle\bar{n}_t,\frac{1}{\eta_t}(\bar{\Xi}_t-\bar{\Xi}_{t+1})\rangle &= \langle K_t\frac{1}{\eta_t}(\bar{\Xi}_t-\bar{\Xi}_{t+1}),\frac{1}{\eta_t}(\bar{\Xi}_t-\bar{\Xi}_{t+1})\rangle\\ &\ge \rho\|\frac{1}{\eta_t}(\bar{\Xi}_t-\bar{\Xi}_{t+1})\|^2\end{aligned} \tag{A19}$$

So

$$\begin{aligned}(2) &= \langle\bar{n}_t,\bar{\Xi}_{t+1}-\bar{\Xi}_t\rangle\\ &\le -\eta_t\rho\|\frac{1}{\eta_t}(\bar{\Xi}_t-\bar{\Xi}_{t+1})\|^2\\ &= -\frac{\rho}{\eta_t}\|\bar{\Xi}_{t+1}-\bar{\Xi}_t\|^2\end{aligned} \tag{A20}$$

 doi:10.20944/preprints202501.1650.v1

Bring (A20), (A18) into (A17), we can get

$$\begin{aligned} f(\bar{\Xi}_{t+1}) &\leq f(\bar{\Xi}_t) + \frac{\eta_t}{\rho}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2 + \frac{\rho}{4\eta_t}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 - \frac{\rho}{\eta_t}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 \\ &+ \frac{L}{2}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 \\ &\leq f(\bar{\Xi}_t) - \frac{\eta_t}{4\rho}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2 + \frac{5\eta_t}{4\rho}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2 - (\frac{3\rho}{4\eta_t} - \frac{L}{2})\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 \\ &\leq f(\bar{\Xi}_t) - (\frac{3\rho}{4\eta_t} - \frac{L}{2})\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 - \frac{\eta_t}{4\rho}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2 + \frac{5\eta_t}{2\rho}\|\bar{g}_t - \bar{n}_t\|^2 \\ &+ \frac{5\eta_t}{2\rho}\|\nabla f(\bar{\Xi}_t) - \bar{g}_t\|^2 \end{aligned} \tag{A21}$$

We considering the last term in (A21), taking expectation on both sides have

$$\begin{aligned} \mathbb{E}\|\nabla f(\bar{\Xi}_t) - \bar{g}_t\|^2 &= \mathbb{E}\|\frac{1}{Q}\sum_{i=1}^{Q}(\nabla f_i(\bar{\Xi}_t) - g_{t,i})\|^2 \\ &\leq \frac{1}{Q^2}Q\sum_{i=1}^{Q}\mathbb{E}\|\nabla f_i(\bar{\Xi}_t) - g_{t,i}\|^2 \\ &\leq \frac{L^2}{Q}\sum_{i=1}^{Q}\mathbb{E}\|\Xi_{t,i} - \bar{\Xi}_t\|^2 \\ &= \frac{L^2}{Q}\mathbb{E}\|\Xi_t - \mathbf{1}\otimes\bar{\Xi}_t\|^2 \end{aligned} \tag{A22}$$

By substituting it into (A21), then taking the expectation of both sides, we can get this conclusion. Lemma A5 is proved. □

Lemma A6. *Suppose that n_t are produced by the Algorithm, we subsequently obtain*

$$\mathbb{E}\|\bar{n}_t - \bar{g}_t\|^2 \leq 2(1-\chi_t)^2\mathbb{E}\|\bar{n}_{t-1} - \bar{g}_{t-1}\|^2 + \frac{4(1-\chi_t)^2L^2}{Q}\mathbb{E}\|\Xi_t - \Xi_{t-1}\|^2 + 4\chi_t^2L^2\mu^2 \tag{A23}$$

Proof. We know that

$$\bar{n}_t = \frac{1}{Q}\sum_{i=1}^{Q}[\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i};\mathfrak{B}_{t,i}) + (1-\chi_t)(\bar{n}_{t-1} - \widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i}))]$$

 doi:10.20944/preprints202501.1650.v1

So we can get

$$
\begin{aligned}
&\mathbb{E}\|\bar{n}_t - \bar{g}_t\|^2 \\
=&\mathbb{E}\|\frac{1}{Q}\sum_{i=1}^{Q}[\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i};\mathfrak{B}_{t,i}) + (1-\chi_t)(\bar{n}_{t-1} - \widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i}))] - \bar{g}_t\|^2 \\
=&\mathbb{E}\|\frac{1}{Q}\sum_{i=1}^{Q}[(\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i};\mathfrak{B}_{t,i}) - g_{t,i}) - (1-\chi_t)(\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i}) - g_{t-1,i})] \\
&+ (1-\chi_t)(\bar{n}_{t-1} - \bar{g}_{t-1})\|^2 \\
=&\mathbb{E}\|\frac{1}{b}\sum_{\vartheta_{t,i}\in\mathfrak{B}_{t,i}}\frac{1}{Q}\sum_{i=1}^{Q}[(\nabla f_i^{\mu}(\Xi_{t,i};\vartheta_{t,i}) - g_{t,i}) - (1-\chi_t)(\nabla f_i^{\mu}(\Xi_{t-1,i};\vartheta_{t,i}) \\
&- g_{t-1,i})] + (1-\chi_t)(\bar{n}_{t-1} - \bar{g}_{t-1})\|^2 \\
\leq&2(1-\chi_t)^2\mathbb{E}\|\bar{n}_{t-1} - \bar{g}_{t-1}\|^2 + \frac{2}{b^2Q^2}bQ\sum_{\vartheta_{t,i}\in\mathfrak{B}_{t,i}}\sum_{i=1}^{Q}\mathbb{E}\|(\nabla f_i^{\mu}(\Xi_{t,i};\vartheta_{t,i}) - g_{t,i}) \\
&- (1-\chi_t)(\nabla f_i^{\mu}(\Xi_{t-1,i};\vartheta_{t,i}) - g_{t-1,i})\|^2 \\
\leq&2(1-\chi_t)^2\mathbb{E}\|\bar{n}_{t-1} - \bar{g}_{t-1}\|^2 + \frac{2}{bQ}\sum_{\vartheta_{t,i}\in\mathfrak{B}_{t,i}}\sum_{i=1}^{Q}\mathbb{E}\|(1-\chi_t)[(\nabla f_i^{\mu}(\Xi_{t,i};\vartheta_{t,i}) - g_{t,i}) \\
&- (\nabla f_i^{\mu}(\Xi_{t-1,i};\vartheta_{t,i}) - g_{t-1,i})] + \chi_t(\nabla f_i^{\mu}(\Xi_{t,i};\vartheta_{t,i}) - g_{t,i})\|^2 \\
\leq&2(1-\chi_t)^2\mathbb{E}\|\bar{n}_{t-1} - \bar{g}_{t-1}\|^2 + \frac{4(1-\chi_t)^2}{bQ}\sum_{\vartheta_{t,i}\in\mathfrak{B}_{t,i}}\sum_{i=1}^{Q}\mathbb{E}\|\nabla f_i^{\mu}(\Xi_{t,i};\vartheta_{t,i}) \\
&- \nabla f_i^{\mu}(\Xi_{t-1,i};\vartheta_{t,i})\|^2 + \frac{4\chi_t^2}{bQ}\sum_{\vartheta_{t,i}\in\mathfrak{B}_{t,i}}\sum_{i=1}^{Q}\mathbb{E}\|\nabla f_i^{\mu}(\Xi_{t,i};\vartheta_{t,i}) - g_{t,i}\|^2 \\
\leq&2(1-\chi_t)^2\mathbb{E}\|\bar{n}_{t-1} - \bar{g}_{t-1}\|^2 + \frac{4(1-\chi_t)^2L^2}{Q}\sum_{i=1}^{Q}\mathbb{E}\|\Xi_{t,i} - \Xi_{t-1,i}\|^2 + 4\chi_t^2L^2\mu^2 \\
=&2(1-\chi_t)^2\mathbb{E}\|\bar{n}_{t-1} - \bar{g}_{t-1}\|^2 + \frac{4(1-\chi_t)^2L^2}{Q}\mathbb{E}\|\Xi_t - \Xi_{t-1}\|^2 + 4\chi_t^2L^2\mu^2
\end{aligned}
\tag{A24}
$$

where the last inequality is because the L-smoothness and Lemma A3.
Lemma A6 is proved. □

Lemma A7. *Suppose that n_t are produced by the Algorithm, we subsequently obtain*

$$
\frac{30\rho}{72Q}\sum_{t=s_t}^{\bar{s}}\eta_t\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i} - \bar{n}_t)\|^2 \leq \frac{\rho}{4}\sum_{t=s_t}^{\bar{s}}\frac{1}{\eta_t}\mathbb{E}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 + \left[\frac{\rho\mu^2c^2}{4Q} + \frac{3\rho\zeta^2c^2}{4L^2}\right]\sum_{t=s_t}^{\bar{s}}\eta_t^3 \tag{A25}
$$

 doi:10.20944/preprints202501.1650.v1

Proof.

$$
\begin{aligned}
&\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i}-\bar{n}_t)\|^2\\
\leq&\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}[[\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i};\mathfrak{B}_{t,i})+(1-\chi_t)(n_{t-1,i}-\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i}))]\\
&-\frac{1}{Q}\sum_{i=1}^{Q}[\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i};\mathfrak{B}_{t,i})+(1-\chi_t)(n_{t-1,i}-\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i}))]]\|^2\\
=&\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}[(1-\chi_t)(n_{t-1,i}-\bar{n}_{t-1})+[\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i};\mathfrak{B}_{t,i})-\frac{1}{Q}\sum_{i=1}^{Q}\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i};\mathfrak{B}_{t,i})\\
&-(1-\chi_t)(\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})-\frac{1}{Q}\sum_{i=1}^{Q}\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i}))]]\|^2\\
\leq&(1+\gamma)(1-\chi_t)^2\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t-1,i}-\bar{n}_{t-1})\|^2+(1+\frac{1}{\gamma})\frac{1}{\rho^2}\sum_{i=1}^{Q}\mathbb{E}\|[\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i};\mathfrak{B}_{t,i})\\
&-\frac{1}{Q}\sum_{i=1}^{Q}\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i};\mathfrak{B}_{t,i})]-(1-\chi_t)[\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})\\
&-\frac{1}{Q}\sum_{i=1}^{Q}\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})]\|^2
\end{aligned}
\tag{A26}
$$

there (A26) arises is because the fact that $K_t \succ \rho I_d$. Continuing to process the latter term in (A26) yields

$$
\begin{aligned}
&\sum_{i=1}^{Q}\mathbb{E}\|\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i};\mathfrak{B}_{t,i})-\frac{1}{Q}\sum_{i=1}^{Q}\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i};\mathfrak{B}_{t,i})\\
&\qquad-(1-\chi_t)[\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})-\frac{1}{Q}\sum_{i=1}^{Q}\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})]\|^2\\
\leq&2\sum_{i=1}^{Q}\mathbb{E}\|[\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i};\mathfrak{B}_{t,i})-\frac{1}{Q}\sum_{i=1}^{Q}\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i};\mathfrak{B}_{t,i})]\\
&-[\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})-\frac{1}{Q}\sum_{i=1}^{Q}\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})]\|^2\\
&+2\chi_t^2\sum_{i=1}^{Q}\mathbb{E}\|\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})-\frac{1}{Q}\sum_{i=1}^{Q}\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})\|^2\\
\leq&2\sum_{i=1}^{Q}\mathbb{E}\|\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t,i};\mathfrak{B}_{t,i})-\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})\|^2\\
&+2\chi_t^2\sum_{i=1}^{Q}\mathbb{E}\|\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})-\frac{1}{Q}\sum_{i=1}^{Q}\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})\|^2\\
\leq&2\sum_{i=1}^{Q}L^2\mathbb{E}\|\Xi_{t,i}-\Xi_{t-1,i}\|^2\\
&+2\chi_t^2\sum_{i=1}^{Q}\mathbb{E}\|\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})-\frac{1}{Q}\sum_{i=1}^{Q}\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})\|^2\\
=&2L^2\mathbb{E}\|\Xi_t-\Xi_{t-1}\|^2\\
&+2\chi_t^2\sum_{i=1}^{Q}\mathbb{E}\|\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})-\frac{1}{Q}\sum_{i=1}^{Q}\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})\|^2
\end{aligned}
\tag{A27}
$$

 doi:10.20944/preprints202501.1650.v1

the last two inequalities here are attributed to Lemma A1 and the L-smoothness. Regarding the last term, we can get

$$\begin{aligned}
&\sum_{i=1}^{Q}\mathbb{E}\|\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})-\frac{1}{Q}\sum_{j=1}^{Q}\widetilde{\nabla}_v^{\mu}F_j(\Xi_{t-1,j};\mathfrak{B}_{t,j})\|^2\\
=&\sum_{i=1}^{Q}\mathbb{E}\|[\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})-g_{t-1,i}]-\frac{1}{Q}\sum_{j=1}^{Q}[\widetilde{\nabla}_v^{\mu}F_j(\Xi_{t-1,j};\mathfrak{B}_{t,j})-g_{t-1,j}]\\
&+[g_{t-1,i}-\bar{g}_{t-1}]\|^2\\
\leq&2\sum_{i=1}^{Q}\mathbb{E}\|\widetilde{\nabla}_v^{\mu}F_i(\Xi_{t-1,i};\mathfrak{B}_{t,i})-g_{t-1,i}\|^2+2\sum_{i=1}^{Q}\mathbb{E}\|g_{t-1,i}-\bar{g}_{t-1}\|^2\\
\leq&2QL^2\mu^2+6Q\zeta^2+12L^2\mathbb{E}\|\Xi_{t-1}-\mathbf{1}\otimes\bar{\Xi}_{t-1}\|^2
\end{aligned}\tag{A28}$$

the last two inequalities here are attributed to Lemma A1 and Lemma A3. Then, by integrating the aforementioned inequalities (A28), (A27), (A26) with the definition of K_t, we can deduce that when $\mathrm{mod}(t,p)\neq 0$, the following formula can be reached.

$$\begin{aligned}
&\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i}-\bar{n}_t)\|^2\\
\leq&(1-\chi_t)^2(1+\gamma)\sum_{i=1}^{Q}\mathbb{E}\|K_{t-1}^{-1}(n_{t-1,i}-\bar{n}_{t-1})\|^2+\frac{2L^2}{\rho^2}(1+\frac{1}{\gamma})\mathbb{E}\|\Xi_t-\Xi_{t-1}\|^2\\
+&\frac{4L^2\mu^2}{\rho^2}(1+\frac{1}{\gamma})\chi_t^2+\frac{12Q}{\rho^2}\zeta^2(1+\frac{1}{\gamma})\chi_t^2+24L^2(1+\frac{1}{\gamma})\frac{\chi_t^2}{\rho^2}\mathbb{E}\|\Xi_{t-1}-\mathbf{1}\otimes\bar{\Xi}_{t-1}\|^2\\
\leq&(1-\chi_t)^2(1+\gamma)\sum_{i=1}^{Q}\mathbb{E}\|K_{t-1}^{-1}(n_{t-1,i}-\bar{n}_{t-1})\|^2+\frac{4L^2\mu^2}{\rho^2}(1+\frac{1}{\gamma})\chi_t^2+\frac{12Q}{\rho^2}\zeta^2(1+\frac{1}{\gamma})\chi_t^2\\
+&\frac{2L^2}{\rho^2}(1+\frac{1}{\gamma})\sum_{i=1}^{Q}\mathbb{E}\|\Xi_{t,i}-\Xi_{t-1,i}\|^2+24L^2(1+\frac{1}{\gamma})\frac{\chi_t^2}{\rho^2}\sum_{i=1}^{Q}\mathbb{E}\|\Xi_{t-1,i}-\bar{\Xi}_{t-1}\|^2\\
\leq&(1-\chi_t)^2(1+\gamma)\sum_{i=1}^{Q}\mathbb{E}\|K_{t-1}^{-1}(n_{t-1,i}-\bar{n}_{t-1})\|^2+\frac{4L^2\mu^2}{\rho^2}(1+\frac{1}{\gamma})\chi_t^2+\frac{12Q}{\rho^2}\zeta^2(1+\frac{1}{\gamma})\chi_t^2\\
+&\frac{2L^2}{\rho^2}(1+\frac{1}{\gamma})\sum_{i=1}^{Q}\mathbb{E}\|\eta_{t-1}K_{t-1}^{-1}n_{t-1,i}\|^2\\
+&24L^2(1+\frac{1}{\gamma})\frac{\chi_t^2}{\rho^2}(p-1)\sum_{s=s_t}^{t-1}\eta_s^2\sum_{i=1}^{Q}\mathbb{E}\|K_s^{-1}(n_{s,i}-\bar{n}_s)\|^2\\
\leq&(1-\chi_t)^2(1+\gamma)\sum_{i=1}^{Q}\mathbb{E}\|K_{t-1}^{-1}(n_{t-1,i}-\bar{n}_{t-1})\|^2+\frac{4L^2\mu^2}{\rho^2}(1+\frac{1}{\gamma})\chi_t^2+\frac{12Q}{\rho^2}\zeta^2(1+\frac{1}{\gamma})\chi_t^2\\
+&24L^2(1+\frac{1}{\gamma})\frac{\chi_t^2}{\rho^2}(p-1)\sum_{s=s_t}^{t-1}\eta_s^2\sum_{i=1}^{Q}\mathbb{E}\|K_s^{-1}(n_{s,i}-\bar{n}_s)\|^2\\
+&\frac{4L^2}{\rho^2}(1+\frac{1}{\gamma})\sum_{i=1}^{Q}\mathbb{E}\Big[\|\eta_{t-1}K_{t-1}^{-1}(n_{t-1,i}-\bar{n}_{t-1})\|^2+\|\eta_{t-1}K_{t-1}^{-1}\bar{n}_{t-1}\|^2\Big]
\end{aligned}\tag{A29}$$

 doi:10.20944/preprints202501.1650.v1

where we utilize the Lemma A4. Then combine like terms we have

$$\begin{aligned}
&\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i}-\bar{n}_t)\|^2\\
\leq&[(1-\chi_t)^2(1+\gamma)+\frac{4L^2}{\rho^2}(1+\frac{1}{\gamma})\eta_{t-1}^2]\sum_{i=1}^{Q}\mathbb{E}\|K_{t-1}^{-1}(n_{t-1,i}-\bar{n}_{t-1})\|^2\\
+&\frac{4QL^2}{\rho^2}(1+\frac{1}{\gamma})\eta_{t-1}^2\mathbb{E}\|K_{t-1}^{-1}\bar{n}_{t-1}\|^2+\frac{4L^2\mu^2}{\rho^2}(1+\frac{1}{\gamma})\chi_t^2+\frac{12Q}{\rho^2}\zeta^2(1+\frac{1}{\gamma})\chi_t^2\\
+&24L^2(1+\frac{1}{\gamma})\frac{\chi_t^2}{\rho^2}(p-1)\sum_{s=s_t}^{t-1}\eta_s^2\sum_{i=1}^{Q}\mathbb{E}\|K_s^{-1}(n_{s,i}-\bar{n}_s)\|^2
\end{aligned}\tag{A30}$$

We select $\gamma=\frac{1}{p}$, $\eta_t\leq\frac{\rho}{12Lp}$ and we know that $\chi_t\in(0,1)$, then

$$\begin{aligned}
(1-\chi_t)^2(1+\gamma)+\frac{4L^2}{\rho^2}(1+\frac{1}{\gamma})\eta_{t-1}^2&\leq 1+\frac{1}{p}+\frac{4L^2}{\rho^2}(1+p)\eta_{t-1}^2\\
&\leq 1+\frac{1}{p}+\frac{p+1}{36p^2}\\
&\leq 1+\frac{19}{18p}
\end{aligned}\tag{A31}$$

Substitute (A31) into (A30), then we can get

$$\begin{aligned}
&\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i}-\bar{n}_t)\|^2\\
\leq&(1+\frac{19}{18p})\sum_{i=1}^{Q}\mathbb{E}\|K_{t-1}^{-1}(n_{t-1,i}-\bar{n}_{t-1})\|^2+\frac{4QL^2}{\rho^2}(1+\frac{1}{\gamma})\eta_{t-1}^2\mathbb{E}\|K_{t-1}^{-1}\bar{n}_{t-1}\|^2\\
+&\frac{4L^2\mu^2}{\rho^2}(1+\frac{1}{\gamma})\chi_t^2+\frac{12Q}{\rho^2}\zeta^2(1+\frac{1}{\gamma})\chi_t^2\\
+&24L^2(1+\frac{1}{\gamma})\frac{\chi_t^2}{\rho^2}(p-1)\sum_{s=s_t}^{t-1}\eta_s^2\sum_{i=1}^{Q}\mathbb{E}\|K_s^{-1}(n_{s,i}-\bar{n}_s)\|^2\\
\leq&(1+\frac{19}{18p})\sum_{i=1}^{Q}\mathbb{E}\|K_{t-1}^{-1}(n_{t-1,i}-\bar{n}_{t-1})\|^2+\frac{2QL}{3\rho}\eta_{t-1}\mathbb{E}\|K_{t-1}^{-1}\bar{n}_{t-1}\|^2+\frac{2L\mu^2c^2}{3\rho}\eta_{t-1}^3\\
+&\frac{2Q\zeta^2c^2}{L\rho}\eta_{t-1}^3+48\frac{L^2p^2c^2\eta_{t-1}^4}{\rho^2}\sum_{s=s_t}^{t-1}\eta_s^2\sum_{i=1}^{Q}\mathbb{E}\|K_s^{-1}(n_{s,i}-\bar{n}_s)\|^2
\end{aligned}\tag{A32}$$

where we considering $\chi_t=c\eta_{t-1}^2$ and $1+p\leq p+p=2p$.

 doi:10.20944/preprints202501.1650.v1

On the other hand, if $\mathrm{mod}(t,p)=0$ that is $t=s_t$, it follows that $\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i}-\bar{n}_t)\|^2=0$. So taking the recursive expansion to (A32), we have

$$\begin{aligned}
&\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i}-\bar{n}_t)\|^2\\
\leq&\frac{2QL}{3\rho}\sum_{s=s_t}^{t-1}(1+\frac{19}{18p})^{t-1-s}\eta_s\mathbb{E}\|K_s^{-1}\bar{n}_s\|^2+\left[\frac{2L\mu^2c^2}{3\rho}+\frac{2Q\zeta^2c^2}{L\rho}\right]\sum_{s=s_t}^{t-1}(1+\frac{19}{18p})^{t-1-s}\eta_s^3\\
+&\frac{48L^2p^2c^2}{\rho^2}\sum_{s=s_t}^{t-1}(1+\frac{19}{18p})^{t-1-s}\eta_s^4\sum_{\bar{s}=s_t}^{s}\eta_{\bar{s}}^2\sum_{i=1}^{Q}\mathbb{E}\|K_{\bar{s}}^{-1}(n_{\bar{s},i}-\bar{n}_{\bar{s}})\|^2\\
\leq&\frac{2QL}{3\rho}\sum_{s=s_t}^{t-1}(1+\frac{19}{18p})^{p}\eta_s\mathbb{E}\|K_s^{-1}\bar{n}_s\|^2+\left[\frac{2L\mu^2c^2}{3\rho}+\frac{2Q\zeta^2c^2}{L\rho}\right]\sum_{s=s_t}^{t-1}(1+\frac{19}{18p})^{p}\eta_s^3\\
+&\frac{48L^2p^3c^2}{\rho^2}(\frac{\rho}{12Lp})^5(1+\frac{19}{18p})^{p}\sum_{s=s_t}^{t}\eta_s\sum_{i=1}^{Q}\mathbb{E}\|K_s^{-1}(n_{s,i}-\bar{n}_s)\|^2\\
\leq&\frac{2QL}{\rho}\sum_{s=s_t}^{t}\eta_s\mathbb{E}\|K_s^{-1}\bar{n}_s\|^2+\left[\frac{2L\mu^2c^2}{\rho}+\frac{6Q\zeta^2c^2}{L\rho}\right]\sum_{s=s_t}^{t}\eta_s^3\\
+&\frac{144L^2p^3c^2}{\rho^2}(\frac{\rho}{12Lp})^5\sum_{s=s_t}^{t}\eta_s\sum_{i=1}^{Q}\mathbb{E}\|K_s^{-1}(n_{s,i}-\bar{n}_s)\|^2
\end{aligned}\tag{A33}$$

where we utilize that $t-1-s\leq p-1<p$ and $(1+\frac{19}{18p})^p\leq e^{\frac{19}{18}}\leq 3$. Then by multiplying both sides by $\sum_{t=s_t}^{\bar{s}}\eta_t$ we can get

$$\begin{aligned}
&\sum_{t=s_t}^{\bar{s}}\eta_t\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i}-\bar{n}_t)\|^2\\
\leq&\frac{2QL}{\rho}\sum_{t=s_t}^{\bar{s}}\eta_t\sum_{s=s_t}^{t}\eta_s\mathbb{E}\|K_s^{-1}\bar{n}_s\|^2+\left[\frac{2L\mu^2c^2}{\rho}+\frac{6Q\zeta^2c^2}{L\rho}\right]\sum_{t=s_t}^{\bar{s}}\eta_t\sum_{s=s_t}^{t}\eta_s^3\\
+&\frac{144L^2p^3c^2}{\rho^2}(\frac{\rho}{12Lp})^5\sum_{t=s_t}^{\bar{s}}\eta_t\sum_{s=s_t}^{t}\eta_s\sum_{i=1}^{Q}\mathbb{E}\|K_s^{-1}(n_{s,i}-\bar{n}_s)\|^2
\end{aligned}\tag{A34}$$

Finally,

$$\begin{aligned}
&\sum_{t=s_t}^{\bar{s}}\eta_t\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i}-\bar{n}_t)\|^2\\
\leq&\frac{2QL}{\rho}(\sum_{t=s_t}^{\bar{s}}\eta_t)\sum_{t=s_t}^{\bar{s}}\eta_t\mathbb{E}\|K_t^{-1}\bar{n}_t\|^2+\left[\frac{2L\mu^2c^2}{\rho}+\frac{6Q\zeta^2c^2}{L\rho}\right](\sum_{t=s_t}^{\bar{s}}\eta_t)\sum_{t=s_t}^{\bar{s}}\eta_t^3\\
+&\frac{144L^2p^3c^2}{\rho^2}(\frac{\rho}{12Lp})^5(\sum_{t=s_t}^{\bar{s}}\eta_t)\sum_{t=s_t}^{\bar{s}}\eta_t\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i}-\bar{n}_t)\|^2\\
\leq&\frac{Q}{6}\sum_{t=s_t}^{\bar{s}}\eta_t\mathbb{E}\|K_t^{-1}\bar{n}_t\|^2+\left[\frac{\mu^2c^2}{6}+\frac{Q\zeta^2c^2}{2L^2}\right]\sum_{t=s_t}^{\bar{s}}\eta_t^3\\
+&\frac{144L^2p^4c^2}{\rho^2}(\frac{\rho}{12Lp})^6\sum_{t=s_t}^{\bar{s}}\eta_t\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i}-\bar{n}_t)\|^2
\end{aligned}\tag{A35}$$

where the last inequality is because $\eta_t\leq\frac{\rho}{12Lp}$ so that $\sum_{t=s_t}^{\bar{s}}\eta_t\leq(\bar{s}-s_t)\frac{\rho}{12Lp}\leq p\frac{\rho}{12Lp}=\frac{\rho}{12L}$.

 doi:10.20944/preprints202501.1650.v1

Therefore,

$$\begin{aligned}&[1-\frac{144L^2p^4c^2}{\rho^2}(\frac{\rho}{12Lp})^6]\sum_{t=s_t}^{\bar{s}}\eta_t\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i}-\bar{n}_t)\|^2\\ \leq&\frac{Q}{6}\sum_{t=s_t}^{\bar{s}}\eta_t\mathbb{E}\|K_t^{-1}\bar{n}_t\|^2+\left[\frac{\mu^2c^2}{6}+\frac{Q\zeta^2c^2}{2L^2}\right]\sum_{t=s_t}^{\bar{s}}\eta_t^3\end{aligned}\tag{A36}$$

Given that $c\leq\frac{60L^2}{\rho^2}$ and $1-\frac{144L^2p^4c^2}{\rho^2}(\frac{\rho}{12Lp})^6\geq\frac{20}{72}$. By multiply $\frac{3\rho}{2Q}$ on both side, we can get

$$\begin{aligned}&\frac{30\rho}{72Q}\sum_{t=s_t}^{\bar{s}}\eta_t\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i}-\bar{n}_t)\|^2\\ \leq&\frac{\rho}{4}\sum_{t=s_t}^{\bar{s}}\eta_t\mathbb{E}\|K_t^{-1}\bar{n}_t\|^2+\left[\frac{\rho\mu^2c^2}{4Q}+\frac{3\rho\zeta^2c^2}{4L^2}\right]\sum_{t=s_t}^{\bar{s}}\eta_t^3\\ =&\frac{\rho}{4}\sum_{t=s_t}^{\bar{s}}\frac{1}{\eta_t}\mathbb{E}\|\bar{\Xi}_{t+1}-\bar{\Xi}_t\|^2+\left[\frac{\rho\mu^2c^2}{4Q}+\frac{3\rho\zeta^2c^2}{4L^2}\right]\sum_{t=s_t}^{\bar{s}}\eta_t^3\end{aligned}\tag{A37}$$

Lemma A7 is proved. □

Proof of Theorem

Now let's prove the final theorem.

Proof. We set $\eta_t=\frac{\rho\bar{h}}{(w_t+t)^{1/3}}$, $\chi_{t+1}=c\cdot\eta_t^2$, $c=\frac{1}{12Lp\bar{h}^3\rho^2}+\frac{30L^2}{\rho^2}\leq\frac{60L^2}{\rho^2}$, $\bar{h}=\frac{1}{L}$ and $w_t=\max(\frac{3}{2},1728L^3p^3\bar{h}^3-t)$. So we can infer that

$$\begin{aligned}\eta_t&=\min\{\frac{\rho\frac{1}{L}}{(\frac{3}{2}+t)^{\frac{1}{3}}},\frac{\rho\frac{1}{L}}{\left(1728L^3p^3\frac{1}{L^3}-t+t\right)^{\frac{1}{3}}}\}\\&=\min\{\frac{\rho}{L(\frac{3}{2}+t)^{\frac{1}{3}}},\frac{\rho}{12Lp}\}\end{aligned}$$

It is clear that $\eta_t\leq\frac{\rho}{12Lp}$. And

$$\begin{aligned}2\eta_t^{-1}-\eta_{t-1}^{-1}&=\frac{2(w_t+t)^{1/3}}{\rho\bar{h}}-\frac{(w_{t-1}+t-1)^{1/3}}{\rho\bar{h}}\\&\leq\frac{2}{3\rho\bar{h}(w_t+(t-1))^{2/3}}\\&\leq\frac{2}{3\rho\bar{h}(w_t/3+t/3)^{2/3}}=\frac{2\cdot3^{2/3}}{3\rho\bar{h}(w_t+t)^{2/3}}\\&=\frac{2\cdot3^{2/3}}{3\rho^3\bar{h}^3}\cdot\frac{\rho^2\bar{h}^2}{(w_t+t)^{2/3}}=\frac{2\cdot3^{2/3}}{3\rho^3\bar{h}^3}\eta_t^2\\&\leq\frac{\eta_t}{6\rho^2\bar{h}^3Lp}\end{aligned}\tag{A38}$$

 doi:10.20944/preprints202501.1650.v1

where we leverage the concavity of $f(\Xi) = \Xi^{1/3}$, that is $(\Xi + \boldsymbol{y})^{1/3} \leq \Xi^{1/3} + \frac{\boldsymbol{y}}{3\Xi^{2/3}}$, the second inequality is valid because of $w_t \geq \frac{3}{2}$ and the last inequality is obtained based on $\eta_t \leq \frac{\rho}{12Lp}$.

$$
\begin{aligned}
&\frac{\mathbb{E}\|\bar{n}_{t+1} - \bar{g}_{t+1}\|^2}{\eta_t} - \frac{\mathbb{E}\|\bar{n}_t - \bar{g}_t\|^2}{\eta_{t-1}} \\
\leq &\left[\frac{2(1-\chi_{t+1})^2}{\eta_t} - \frac{1}{\eta_{t-1}}\right]\mathbb{E}\|\bar{n}_t - \bar{g}_t\|^2 + \frac{4(1-\chi_{t+1})^2L^2}{Q\eta_t}\mathbb{E}\|\Xi_{t+1} - \Xi_t\|^2 + \frac{4\chi_{t+1}^2L^2\mu^2}{\eta_t} \\
\leq &[2\eta_t^{-1} - \eta_{t-1}^{-1} - 2c\eta_t]\mathbb{E}\|\bar{n}_t - \bar{g}_t\|^2 + \frac{8(1-\chi_{t+1})^2L^2}{Q}\eta_t\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i} - \bar{n}_t)\|^2 \\
&+8(1-\chi_{t+1})^2L^2\eta_t\mathbb{E}\|K_t^{-1}\bar{n}_t\|^2 + \frac{4\chi_{t+1}^2L^2\mu^2}{\eta_t} \\
\leq &-\frac{60L^2}{\rho^2}\eta_t\mathbb{E}\|\bar{n}_t - \bar{g}_t\|^2 + \frac{8L^2}{Q}\eta_t\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i} - \bar{n}_t)\|^2 + 8L^2\eta_t\mathbb{E}\|K_t^{-1}\bar{n}_t\|^2 \\
&+4c^2\eta_t^3L^2\mu^2.
\end{aligned} \tag{A39}
$$

where the second inequality holds true because $(a+b)^2 \leq 2a^2 + 2b^2$ and $(1-\chi_{t+1})^2 \leq 1 - \chi_{t+1}$, $\chi_{t+1} = c \cdot \eta_t^2$ and the last inequality is obtained based on (A38). Therefore, we have

$$
\begin{aligned}
&\frac{\rho}{24L^2}[\frac{\mathbb{E}\|\bar{n}_{t+1} - \bar{g}_{t+1}\|^2}{\eta_t} - \frac{\mathbb{E}\|\bar{n}_t - \bar{g}_t\|^2}{\eta_{t-1}}] \\
\leq &-\frac{5}{2\rho}\eta_t\mathbb{E}\|\bar{n}_t - \bar{g}_t\|^2 + \frac{\rho}{3Q}\eta_t\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i} - \bar{n}_t)\|^2 + \frac{\rho}{3}\eta_t\mathbb{E}\left\|K_t^{-1}\bar{n}_t\right\|^2 + \frac{\rho c^2\eta_t^3\mu^2}{6}
\end{aligned} \tag{A40}
$$

Subsequently, we set

$$
\Gamma_t = f(\bar{\Xi}_t) + \frac{\rho}{24L^2}\frac{\|\bar{n}_t - \bar{g}_t\|^2}{\eta_{t-1}}
$$

$$
\begin{aligned}
\mathbb{E}[\Gamma_{t+1} - \Gamma_t] &= \mathbb{E}[f(\bar{\Xi}_{t+1}) - f(\bar{\Xi}_t) + \frac{\rho}{24L^2}(\frac{\|\bar{n}_{t+1} - \bar{g}_{t+1}\|^2}{\eta_t} - \frac{\|\bar{n}_t - \bar{g}_t\|^2}{\eta_{t-1}})] \\
&\leq -(\frac{3\rho}{4\eta_t} - \frac{L}{2})\mathbb{E}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 - \frac{\eta_t}{4\rho}\mathbb{E}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2 \\
&+ \frac{5\eta_tL^2(p-1)}{2\rho Q}\sum_{s=s_t}^{t}\eta_s^2\sum_{i=1}^{Q}\mathbb{E}\|K_s^{-1}(n_{s,i} - \bar{n}_s)\|^2 \\
&+ \frac{\rho}{3Q}\eta_t\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i} - \bar{n}_t)\|^2 + \frac{\rho}{3}\eta_t\mathbb{E}\|K_t^{-1}\bar{n}_t\|^2 + \frac{\rho c^2\eta_t^3\mu^2}{6} \\
&= -(\frac{3\rho}{4\eta_t} - \frac{L}{2})\mathbb{E}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 - \frac{\eta_t}{4\rho}\mathbb{E}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2 \\
&+ \frac{5\eta_tL^2(p-1)}{2\rho Q}\sum_{s=s_t}^{t}\eta_s^2\sum_{i=1}^{Q}\mathbb{E}\|K_s^{-1}(n_{s,i} - \bar{n}_s)\|^2 \\
&+ \frac{\rho}{3Q}\eta_t\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i} - \bar{n}_t)\|^2 + \frac{\rho}{3\eta_t}\mathbb{E}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 + \frac{\rho c^2\eta_t^3\mu^2}{6} \\
&\leq -\frac{\rho}{3\eta_t}\mathbb{E}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 - \frac{\eta_t}{4\rho}\mathbb{E}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2 \\
&+ \frac{5\eta_tL^2(p-1)}{2\rho Q}\sum_{s=s_t}^{t}\eta_s^2\sum_{i=1}^{Q}\mathbb{E}\|K_s^{-1}(n_{s,i} - \bar{n}_s)\|^2 \\
&+ \frac{\rho}{3Q}\eta_t\sum_{i=1}^{Q}\mathbb{E}\|K_t^{-1}(n_{t,i} - \bar{n}_t)\|^2 + \frac{\rho c^2\eta_t^3\mu^2}{6}
\end{aligned} \tag{A41}
$$

 doi:10.20944/preprints202501.1650.v1

where we utilize that Lemma A4, Lemma A5 and $\frac{L}{2} \leq \frac{\rho}{24\eta_t p} \leq \frac{\rho}{24\eta_t}$. By summing the results from $t = s_t$ to $\bar{s}$, where $\bar{s} \in [\lfloor t/p \rfloor p, (\lfloor t/p \rfloor + 1)p]$, then can obtain

$$\begin{aligned}
\mathbb{E}[\Gamma_{\bar{s}+1} - \Gamma_{s_t}] \leq & \sum_{t=s_t}^{\bar{s}} [-\frac{\rho}{3\eta_t}\mathbb{E}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 - \frac{\eta_t}{4\rho}\mathbb{E}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2] + \sum_{t=s_t}^{\bar{s}} \frac{\rho c^2 \eta_t^3 \mu^2}{6} \\
& + \sum_{t=s_t}^{\bar{s}} \frac{5\eta_t L^2 (p-1)}{2\rho Q} \sum_{s=s_t}^{t} \eta_s^2 \sum_{i=1}^{Q} \mathbb{E}\|K_s^{-1}(n_{s,i} - \bar{n}_s)\|^2 \\
& + \frac{\rho}{3Q} \sum_{t=s_t}^{\bar{s}} \eta_t \sum_{i=1}^{Q} \mathbb{E}\|K_t^{-1}(n_{t,i} - \bar{n}_t)\|^2 \\
\leq & \sum_{t=s_t}^{\bar{s}} [-\frac{\rho}{3\eta_t}\mathbb{E}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 - \frac{\eta_t}{4\rho}\mathbb{E}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2] + \sum_{t=s_t}^{\bar{s}} \frac{\rho c^2 \eta_t^3 \mu^2}{6} \\
& + \frac{5L^2(p-1)}{2\rho Q} (\sum_{t=s_t}^{\bar{s}} \eta_t) \sum_{t=s_t}^{\bar{s}} \eta_t^2 \sum_{i=1}^{Q} \mathbb{E}\|K_t^{-1}(n_{t,i} - \bar{n}_t)\|^2 \\
& + \frac{\rho}{3Q} \sum_{t=s_t}^{\bar{s}} \eta_t \sum_{i=1}^{Q} \mathbb{E}\|K_t^{-1}(n_{t,i} - \bar{n}_t)\|^2 \\
\leq & \sum_{t=s_t}^{\bar{s}} [-\frac{\rho}{3\eta_t}\mathbb{E}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 - \frac{\eta_t}{4\rho}\mathbb{E}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2] \\
& + \frac{\rho}{3Q} \sum_{t=s_t}^{\bar{s}} \eta_t \sum_{i=1}^{Q} \mathbb{E}\left\|K_t^{-1}(n_{t,i} - \bar{n}_t)\right\|^2 \\
& + \frac{5L^2(p-1)}{2\rho Q} \left(p \times \frac{\rho}{12Lp} \times \frac{\rho}{12Lp}\right) \sum_{t=s_t}^{\bar{s}} \eta_t \sum_{i=1}^{Q} \mathbb{E}\|K_t^{-1}(n_{t,i} - \bar{n}_t)\|^2 \\
& + \sum_{t=s_t}^{\bar{s}} \frac{\rho c^2 \eta_t^3 \mu^2}{6} \\
\leq & \sum_{t=s_t}^{\bar{s}} [-\frac{\rho}{3\eta_t}\mathbb{E}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 - \frac{\eta_t}{4\rho}\mathbb{E}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2] + \sum_{t=s_t}^{\bar{s}} \frac{\rho c^2 \eta_t^3 \mu^2}{6} \\
& + \frac{26\rho}{72Q} \sum_{t=s_t}^{\bar{s}} \eta_t \sum_{i=1}^{Q} \mathbb{E}\|K_t^{-1}(n_{t,i} - \bar{n}_t)\|^2 \\
\leq & \sum_{t=s_t}^{\bar{s}} [-\frac{\rho}{3\eta_t}\mathbb{E}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 - \frac{\eta_t}{4\rho}\mathbb{E}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2] + \sum_{t=s_t}^{\bar{s}} \frac{\rho c^2 \eta_t^3 \mu^2}{6} \\
& + \frac{\rho}{4} \sum_{t=s_t}^{\bar{s}} \frac{1}{\eta_t}\mathbb{E}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 + \left[\frac{\rho\mu^2 c^2}{4Q} + \frac{3\rho\zeta^2 c^2}{4L^2}\right] \sum_{t=s_t}^{\bar{s}} \eta_t^3
\end{aligned} \tag{A42}$$

here, by utilizing Lemma A7 and the fact that $\frac{26}{72} < \frac{30}{72}$, we can derive the last inequality. Subsequently, summing the terms from the start can obtain

$$\begin{aligned}
\mathbb{E}[\Gamma_T - \Gamma_0] \leq & \sum_{t=0}^{T-1} \left[-\frac{\rho}{12\eta_t}\mathbb{E}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 - \frac{\eta_t}{4\rho}\mathbb{E}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2\right] + \sum_{t=0}^{T-1} \frac{\rho c^2 \eta_t^3 \mu^2}{6} \\
& + \frac{\rho\mu^2 c^2}{4Q} \sum_{t=0}^{T-1} \eta_t^3 + \frac{3\rho\zeta^2 c^2}{4L^2} \sum_{t=0}^{T-1} \eta_t^3
\end{aligned} \tag{A43}$$

 doi:10.20944/preprints202501.1650.v1

Furthermore, we can obtain

$$\begin{aligned}
&\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{\rho}{12\eta_t}\|\bar{\Xi}_{t+1}-\bar{\Xi}_t\|^2+\frac{\eta_t}{4\rho}\|\nabla f(\bar{\Xi}_t)-\bar{n}_t\|^2\right]\\
\leq&\mathbb{E}[\Gamma_0-\Gamma_T]+\frac{5\rho\mu^2c^2}{12}\sum_{t=0}^{T-1}\eta_t^3+\frac{3\rho\zeta^2c^2}{4L^2}\sum_{t=0}^{T-1}\eta_t^3\\
\leq&\mathbb{E}[f(\bar{\Xi}_0)-f^*]+\frac{\rho}{24L^2}\frac{\|\bar{n}_0-\bar{g}_0\|^2}{\eta_0}\\
+&\frac{5\rho\mu^2c^2}{12}\sum_{t=0}^{T-1}\eta_t^3+\frac{3\rho\zeta^2c^2}{4L^2}\sum_{t=0}^{T-1}\eta_t^3
\end{aligned}\tag{A44}$$

Then consider that $\sum_{t=0}^{T-1}\eta_t^3=\sum_{t=0}^{T-1}\frac{\rho^3\bar{h}^3}{w_t+t}\leq\sum_{t=0}^{T-1}\frac{\rho^3\bar{h}^3}{1+t}\leq\rho^3\bar{h}^3(\ln T+1)$, since $w_t\geq\frac{3}{2}>1$. Taking Lemma A3 and dividing both sides of the above result by $\rho\eta_T T$ can get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{12\eta_t^2}\|\bar{\Xi}_{t+1}-\bar{\Xi}_t\|^2+\frac{1}{4\rho^2}\|\nabla f(\bar{\Xi}_t)-\bar{n}_t\|^2\right]\\
\leq&\frac{\mathbb{E}[f(\bar{\Xi}_0)-f^*]}{\eta_T T\rho}+\frac{\mu^2}{\eta_T T24Q\eta_0}+\frac{\rho^3}{\eta_T TL^2}\left[\frac{5L^2\mu^2}{12}+\frac{3\zeta^2}{4}\right]c^2\bar{h}^3(\ln T+1)
\end{aligned}\tag{A45}$$

Regarding the first term in (A45),

$$\frac{1}{\eta_T T}=\frac{(w_T+T)^{1/3}}{\rho\bar{h}T}\leq\frac{w_T^{1/3}}{\rho\bar{h}T}+\frac{1}{\rho\bar{h}T^{2/3}}\leq\frac{12Lq}{\rho T}+\frac{L}{\rho T^{2/3}}\tag{A46}$$

For the middle term, we have

$$\begin{aligned}
\frac{\mu^2}{\eta_T T24Q\eta_0}&\leq\left(\frac{12Lp}{\rho T}+\frac{L}{\rho T^{2/3}}\right)\times\frac{\mu^2}{24Q}\times\frac{w_0^{1/3}}{\rho\bar{h}}\\
&\leq\left(\frac{12Lp}{\rho T}+\frac{L}{\rho T^{2/3}}\right)\times\frac{\mu^2}{24Q}\times\frac{12Lp}{\rho}\\
&\leq\frac{6L^2\mu^2p^2}{\rho^2QT}+\frac{L^2\mu^2p}{2\rho^2QT^{2/3}}
\end{aligned}\tag{A47}$$

For the third term

$$\begin{aligned}
\frac{\rho^3c^2\bar{h}^3}{4\eta_T TL^2}&\leq\left(\frac{12Lp}{\rho T}+\frac{L}{\rho T^{2/3}}\right)\times\left(\frac{60L^2}{\rho^2}\right)^2\times\frac{\rho^3\bar{h}^3}{4L^2}\\
&=\left(\frac{12Lp}{\rho T}+\frac{L}{\rho T^{2/3}}\right)\times\frac{3600L^4}{\rho^4}\times\frac{\rho^3\frac{1}{L^3}}{4L^2}\\
&=\left(\frac{12Lp}{\rho T}+\frac{L}{\rho T^{2/3}}\right)\times\frac{900}{\rho L}\\
&=\frac{12^2\times75p}{\rho^2T}+\frac{900}{\rho^2T^{2/3}}
\end{aligned}\tag{A48}$$

 doi:10.20944/preprints202501.1650.v1

We let $\mathcal{Q}_t = \frac{1}{12\eta_t^2}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 + \frac{1}{4\rho^2}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2$

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\mathcal{Q}_t] &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\frac{1}{12\eta_t^2}\|\bar{\Xi}_{t+1} - \bar{\Xi}_t\|^2 + \frac{1}{4\rho^2}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2\right] \\ &\leq \left[\frac{12Lp}{\rho^2 T} + \frac{L}{\rho^2 T^{2/3}}\right]\mathbb{E}[f(\bar{\Xi}_0) - f^*] + \frac{6L^2\mu^2p^2}{\rho^2 QT} + \frac{L^2\mu^2 p}{2\rho^2 QT^{2/3}} \\ &+ \left[\frac{12^2\times 75p}{\rho^2 T} + \frac{900}{\rho^2 T^{2/3}}\right]\left[\frac{5L^2\mu^2}{3} + 3\zeta^2\right](\ln T + 1) \end{aligned} \tag{A49}$$

and if we choose $p = \left(\frac{T}{Q^2}\right)^{\frac{1}{3}}$, then $\frac{p}{T} = \frac{1}{(QT)^{\frac{2}{3}}}$, $\frac{p^2}{QT} = \frac{1}{Q^{\frac{7}{3}}T^{\frac{1}{3}}}$, $\frac{p}{QT^{\frac{2}{3}}} = \frac{1}{Q^{\frac{5}{3}}T^{\frac{1}{3}}}$. So we can infer that it is convergent.

Then with Jensen's inequality and $\|K_t\| \geq \rho$ can get

$$\begin{aligned}&\frac{1}{\eta_t}\|\bar{\Xi}_t - \bar{\Xi}_{t+1}\| + \frac{1}{\rho}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\| \\ =&\left\|K_t^{-1}\bar{n}_t\right\| + \frac{1}{\rho}\left\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\right\| \\ =&\frac{1}{\|K_t\|}\|K_t\|\left\|K_t^{-1}\bar{n}_t\right\| + \frac{1}{\rho}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\| \\ \geq&\frac{1}{\|K_t\|}\|\bar{n}_t\| + \frac{1}{\|K_t\|}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\| \\ \geq&\frac{1}{\|K_t\|}\|\nabla f(\bar{\Xi}_t)\| \end{aligned} \tag{A50}$$

Finally,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\bar{\Xi}_t)\| &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|K_t\|\left(\frac{1}{\eta_t}\|\bar{\Xi}_t - \bar{\Xi}_{t+1}\| + \frac{1}{\rho}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|\right)\right] \\ &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\frac{\lambda}{2}\|K_t\|^2 + \frac{1}{2\lambda}\left[\frac{1}{\eta_t}\|\bar{\Xi}_t - \bar{\Xi}_{t+1}\| + \frac{1}{\rho}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|\right]^2\right] \\ &= \frac{\lambda}{2}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|K_t\|^2 + \frac{1}{2\lambda}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{\eta_t}\|\bar{\Xi}_t - \bar{\Xi}_{t+1}\| + \frac{1}{\rho}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|\right]^2 \\ &\leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|K_t\|^2}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\frac{2}{\eta_t^2}\|\bar{\Xi}_t - \bar{\Xi}_{t+1}\|^2 + \frac{2}{\rho^2}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|^2\right]} \\ &\leq 5\sqrt{2d(G^2 + \sigma_g^2) + 2\rho^2 + \frac{1}{2}d^2L^2\mu^2}\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\mathcal{Q}_t]} \end{aligned} \tag{A51}$$

where we utilize Young's inequality and $(a+b)^2 \leq 2a^2 + 2b^2$, and $\lambda = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{\eta_t}\|\bar{\Xi}_t - \bar{\Xi}_{t+1}\| + \frac{1}{\rho}\|\nabla f(\bar{\Xi}_t) - \bar{n}_t\|\right]^2} / \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|K_t\|^2}$.

Since $\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\mathcal{Q}_t]$ is convergent as mentioned before, it can be known that $\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\bar{\Xi}_t)\|$ is also convergent, thus the theorem is proved. □

 doi:10.20944/preprints202501.1650.v1

References

1. McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the Artificial intelligence and statistics, 2017, pp. 1273–1282.
2. Shi, Y.; Yang, K.; Yang, Z.; Zhou, Y. Mobile edge artificial intelligence: Opportunities and challenges **2021**.
3. Yang, L.; Tan, B.; Zheng, V.W.; Chen, K.; Yang, Q. Federated recommendation systems. *Federated Learning: Privacy and Incentive* **2020**, pp. 225–239.
4. Yang, K.; Shi, Y.; Zhou, Y.; Yang, Z.; Fu, L.; Chen, W. Federated machine learning for intelligent IoT via reconfigurable intelligent surface. *IEEE network* **2020**, *34*, 16–22.
5. Tian, J.; Smith, J.S.; Kira, Z. Fedfor: Stateless heterogeneous federated learning with first-order regularization. *arXiv.2209.10537* **2022**.
6. Zhang, M.; Sapra, K.; Fidler, S.; Yeung, S.; Alvarez, J.M. Personalized federated learning with first order model optimization **2021**.
7. Elbakary, A.; Issaid, C.B.; Shehab, M.; Seddik, K.G.; ElBatt, T.A.; Bennis, M. Fed-Sophia: A Communication-Efficient Second-Order Federated Learning Algorithm. *arXiv.2406.06655* **2024**.
8. Dai, Z.; Low, B.K.H.; Jaillet, P. Federated Bayesian optimization via Thompson sampling. *Advances in Neural Information Processing Systems* **2020**, *33*, 9687–9699.
9. Fang, W.; Yu, Z.; Jiang, Y.; Shi, Y.; Jones, C.N.; Zhou, Y. Communication-efficient stochastic zeroth-order optimization for federated learning. *IEEE Transactions on Signal Processing* **2022**, *70*, 5058–5073.
10. Li, Z.; Ying, B.; Liu, Z.; Yang, H. Achieving Dimension-Free Communication in Federated Learning via Zeroth-Order Optimization. *arXiv.2405.15861* **2024**.
11. Maritan, A.; Dey, S.; Schenato, L. FedZeN: Quadratic convergence in zeroth-order federated learning via incremental Hessian estimation. In Proceedings of the 2024 European Control Conference, 2024, pp. 2320–2327.
12. Staib, M.; Reddi, S.; Kale, S.; Kumar, S.; Sra, S. Escaping saddle points with adaptive gradient methods. In Proceedings of the International Conference on Machine Learning, 2019, pp. 5956–5965.
13. Chen, X.; Li, X.; Li, P. Toward communication efficient adaptive gradient method. In Proceedings of the Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference, 2020, pp. 119–128.
14. Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečnỳ, J.; Kumar, S.; McMahan, H.B. Adaptive federated optimization. *arXiv:2003.00295* **2020**.
15. Wang, Y.; Lin, L.; Chen, J. Communication-efficient adaptive federated learning. In Proceedings of the International Conference on Machine Learning, 2022, pp. 22802–22838.
16. Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; Poor, H.V. A novel framework for the analysis and design of heterogeneous federated learning. *IEEE Transactions on Signal Processing* **2021**, *69*, 5234–5249.
17. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems* **2020**, *2*, 429–450.
18. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; Suresh, A.T. Scaffold: Stochastic controlled averaging for federated learning. In Proceedings of the International conference on machine learning, 2020, pp. 5132–5143.
19. Pathak, R.; Wainwright, M.J. FedSplit: An algorithmic framework for fast federated optimization. *Advances in neural information processing systems* **2020**, *33*, 7057–7066.
20. Zhang, X.; Hong, M.; Dhople, S.; Yin, W.; Liu, Y. Fedpd: A federated learning framework with adaptivity to non-iid data. *IEEE Transactions on Signal Processing* **2021**, *69*, 6055–6070.
21. Wang, S.; Roosta, F.; Xu, P.; Mahoney, M.W. Giant: Globally improved approximate newton method for distributed optimization. *Advances in Neural Information Processing Systems* **2018**, *31*.
22. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smithy, V. Feddane: A federated newton-type method. In Proceedings of the 2019 53rd Asilomar Conference on Signals, Systems, and Computers, 2019, pp. 1227–1231.
23. Xu, A.; Huang, H. Coordinating momenta for cross-silo federated learning. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 8735–8743.
24. Das, R.; Acharya, A.; Hashemi, A.; Sanghavi, S.; Dhillon, I.S.; Topcu, U. Faster non-convex federated learning via global and local momentum. In Proceedings of the Uncertainty in Artificial Intelligence, 2022, pp. 496–506.

25. Khanduri, P.; Sharma, P.; Yang, H.; Hong, M.; Liu, J.; Rajawat, K.; Varshney, P. Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. *Advances in Neural Information Processing Systems* **2021**, *34*, 6050–6061.
26. Jiang, W.; Han, H.; Zhang, Y.; Mu, J. Federated split learning for sequential data in satellite-terrestrial integrated networks. *Information Fusion* **2024**, *103*, 102141.
27. Jiang, W.; Han, H.; Zhang, Y.; Mu, J.; Shankar, A. Intrusion Detection with Federated Learning and Conditional Generative Adversarial Network in Satellite-Terrestrial Integrated Networks. *Mobile Networks and Applications* **2024**, pp. 1–14.
28. Tang, Y.; Zhang, J.; Li, N. Distributed zero-order algorithms for nonconvex multiagent optimization. *IEEE Transactions on Control of Network Systems* **2020**, *8*, 269–281.
29. Nikolakakis, K.; Haddadpour, F.; Kalogerias, D.; Karbasi, A. Black-box generalization: Stability of zeroth-order learning. *Advances in neural information processing systems* **2022**, *35*, 31525–31541.
30. Balasubramanian, K.; Ghadimi, S. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. *Advances in Neural Information Processing Systems* **2018**, *31*.
31. Mhanna, E.; Assaad, M. Rendering Wireless Environments Useful for Gradient Estimators: A Zero-Order Stochastic Federated Learning Method. *arXiv.2401.17460* **2024**.
32. Diederik, P.K. Adam: A method for stochastic optimization **2014**.
33. Duchi, J.; Hazan, E.; Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research* **2011**, *12*.
34. Ling, X.; Fu, J.; Wang, K.; Liu, H.; Chen, Z. Ali-dpfl: Differentially private federated learning with adaptive local iterations. In Proceedings of the 2024 IEEE 25th International Symposium on a World of Wireless, Mobile and Multimedia Networks, 2024, pp. 349–358.
35. Cong, Y.; Qiu, J.; Zhang, K.; Fang, Z.; Gao, C.; Su, S.; Tian, Z. Ada-FFL: Adaptive computing fairness federated learning. *CAAI Transactions on Intelligence Technology* **2024**, *9*, 573–584.
36. Huang, Y.; Zhu, S.; Chen, W.; Huang, Z. FedAFR: Enhancing Federated Learning with adaptive feature reconstruction. *Computer Communications* **2024**, *214*, 215–222.
37. Li, Y.; He, Z.; Gu, X.; Xu, H.; Ren, S. AFedAvg: Communication-efficient federated learning aggregation with adaptive communication frequency and gradient sparse. *Journal of Experimental & Theoretical Artificial Intelligence* **2024**, *36*, 47–69.
38. Gao, X.; Jiang, B.; Zhang, S. On the information-adaptive variants of the ADMM: an iteration complexity perspective. *Journal of Scientific Computing* **2018**, *76*, 327–363.
39. Wu, X.; Huang, F.; Hu, Z.; Huang, H. Faster adaptive federated learning. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2023, Vol. 37, pp. 10379–10387.