# Rethinking the Data Heterogeneity in Federated Learning

Jiayi Wang*, Shiqiang Wang†, Rong-Rong Chen*, Mingyue Ji*
*Department of Electrical & Computer Engineering, University of Utah
†IBM T. J. Watson Research Center
Email: *{jiayi.wang, rchen, mingyue.ji}@utah.edu, †wangshiq@us.ibm.com

2023 57th Asilomar Conference on Signals, Systems, and Computers | 979-8-3503-2574-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEEECONF59524.2023.10476979

***Abstract*—Dealing with data heterogeneity is a key challenge in the theoretical analysis of federated learning (FL) algorithms. In the literature, gradient divergence is often used as the sole metric for data heterogeneity. However, we observe that the gradient divergence cannot fully characterize the impact of the data heterogeneity in Federated Averaging (FedAvg) even for the quadratic objective functions. This limitation leads to an overestimate of the communication complexity. Motivated by this observation, we propose a new analysis framework based on the difference between the minima of the global objective function and the minima of the local objective functions. Using the new framework, we derive a tighter convergence upper bound for heterogeneous quadratic objective functions. The theoretical results reveal new insights into the impact of the data heterogeneity on the convergence of FedAvg and provide a deeper understanding of the two-stage learning rates. Experimental results using non-IID data partitions validate the theoretical findings.**

***Index Terms*—Federated Learning, Data Heterogeneity**

## I. Introduction

We consider the following federated optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} f(\boldsymbol{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}), \tag{1}$$

where $n$ is the number of workers and $f_i(\boldsymbol{x})$ is the local objective function of worker $i$. The most popular algorithm for solving this problem is Federated Averaging (FedAvg) [1]–[3], which can be found in Algorithm 1. In FedAvg, to reduce the communication cost, workers often perform $K > 1$ local iterations of stochastic gradient descent (SGD) on their own devices before sending the updated models to the global server. After receiving updated local models, the global server updates the global model by averaging all local models then sends the new global model back to all workers. The server and workers collaboratively perform $R$ communication rounds.

A key challenge in federated learning is the data heterogeneity, which severely restricts the usefulness of multiple local iterations [4], [5]. Since workers can only use local data to train the model, the local models often drift far away from the global model. Therefore, global aggregations are needed to mitigate the impact of data heterogeneity. In the literature, the gradient divergence in Assumption 1 is widely applied to characterize the impact of data heterogeneity. Based on Assumption 1, existing results [5], [6] show that we have to choose a small $K$ to deal with the large gradient divergence.

*Assumption 1 (Bounded Gradient Divergence):* There exists $\zeta > 0$ such that $\forall \boldsymbol{x} \in \mathbb{R}^d$,

$$\sup_{i\in[n],\boldsymbol{x}} \|\nabla f_i(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|^2 = \zeta^2. \tag{2}$$

However, in Section III, we show that the gradient divergence does not fully capture the impact of data heterogeneity on the convergence, which leads to a lack of understanding or even misunderstanding of the behavior of local updates. Specifically, this gap in understanding might result in an overestimate of the communication complexity. Furthermore, such misunderstandings could lead to improper choices of learning rates, potentially causing the divergence of FedAvg.

In this work, we propose a new framework for the analysis of FedAvg with quadratic objective functions, based on the difference between the minima of the global objective function and the minima of the local objective functions. The intuition behind the new framework is that *if the minima of the local objective functions and the global objective function are close to each other, then we can perform as many local updates as possible until the local model converges to its minima instead of doing a number of unnecessary global aggregations.* Our approach does not explicitly use the bounded gradient divergence assumption, as most existing works do, since the focus of our analysis is the heterogeneity on "destination", (the minima), instead of the "direction", (the gradients). Thus, our proposed framework can demonstrate convergence for heterogeneous objective functions even when the bounded gradient divergence assumption fails. Using the new framework, we derive a tighter convergence bound for quadratic objective functions, which matches the lower bound for the convex quadratic objective functions in [5]. The theoretical results provide new insights into the impact of data heterogeneity on the convergence of FedAvg and improve the understanding of the two-stage learning rates. Experimental results with non-IID data partitions further validate our theoretical findings.

## II. Related Works

There have been a considerable amount of works analyzing the convergence of FedAvg, for convex objective functions [7], [8], non-convex objective functions [9], [10], and their variants [5], [11], [12]. However, in these works, the convergence error caused by the gradient divergence is given by $\mathcal{O}(\gamma^2 K^2 \zeta^2)$, which means that when the gradient divergence is large, the

**Algorithm 1:** Federated Averaging (FedAvg)

**Input:** $\gamma$, $\bar{\boldsymbol{x}}^0$, $K$, $\eta$ (if using two-stage learning rates)
**Output:** Global aggregated model $\bar{\boldsymbol{x}}^R$
**for** $r = 0$ *to* $R-1$ **do**
  Distribute the current global model $\bar{\boldsymbol{x}}^r$ to workers;
  **for** *Each worker i, in parallel* **do**
    $\tau = 0$;
    **while** $\tau < K$ **do**
      Compute $\nabla f_i(\boldsymbol{x}_i^{r,\tau})$ using the local dataset;
      $\boldsymbol{x}_i^{r,\tau+1} \leftarrow \boldsymbol{x}_i^{r,\tau} - \gamma \nabla f_i(\boldsymbol{x}_i^{r,\tau})$;
      $\tau \leftarrow \tau + 1$;
    Send $\boldsymbol{x}_i^{r,K}$ to the server;
  **if** *Using two-stage learning rates* **then**
    Update the global model
    $\bar{\boldsymbol{x}}^{r+1} \leftarrow \bar{\boldsymbol{x}}^r + \eta(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^{r,K} - \bar{\boldsymbol{x}}^r)$;
  **else**
    Update the global model $\bar{\boldsymbol{x}}^{r+1} \leftarrow \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^{r,K}$;

convergence error grows fast with $K$. Recently, a framework with two-stage learning rates, $\eta$ and $\gamma$, was proposed to improve the convergence performance under severe non-IID conditions [12]. Some work [13], [14] focus on the analysis of FedAvg with two-stage learning rates. However, the convergence bounds in these works imply that only when fixing the product of $\eta$ and $\gamma$, then letting $\gamma$ become as small as possible, the convergence error is minimized, which, as we will show in Section IV-B, can lead to the divergence.

## III. Motivation: A Single Aggregation Can Be Sufficient

In this section, we present an example involving heterogeneous quadratic objective functions, where FedAvg can converge to the global minima $\boldsymbol{x}^*$, with only one aggregation, while the gradient divergence can be arbitrarily large. We attribute this phenomenon to the limitations of the gradient divergence metric in capturing the difference between the minima of the local objective function $\boldsymbol{x}_i^*$ and the minima of the global objective function $\boldsymbol{x}^*$. These observations motivate us to shift our focus in the convergence analysis towards the heterogeneity on minima of objective functions, rather than relying solely on the gradient divergence.

We consider the following quadratic example. The global objective function is given by

$$f(\boldsymbol{x}) = c + \boldsymbol{b}^T\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}, \boldsymbol{A} \succ 0, \tag{3}$$

where $\boldsymbol{A} \succ 0$ means $\boldsymbol{A}$ is positive definite. The local objective function of worker $i$ is given by

$$f_i(\boldsymbol{x}) = c_i + \boldsymbol{b}_i^T\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}, \forall i \in [n]. \tag{4}$$

Both the server and workers share the same Hessian matrix $\boldsymbol{A}$, while other coefficients $c_i$, $\boldsymbol{b}_i$ can be different. According to Assumption 1, in this case, the gradient divergence is

$$\zeta^2 = \sup_i \|\boldsymbol{b} - \boldsymbol{b}_i\|^2. \tag{5}$$

When the local objective functions are highly heterogeneous, the gradient divergence can be arbitrarily large. In the literature, a large gradient divergence implies that frequent global aggregations are needed [10], [15]. However, only one aggregation is sufficient for this example. To see this, let workers perform a sufficiently large number $K$ of local iterations until it converges to the minima of the local objective function $\boldsymbol{x}_i^*$. Then no matter which $\bar{\boldsymbol{x}}^0$ is given, we always have $\boldsymbol{x}_i^{0,K} = \boldsymbol{x}_i^*$. According to Algorithm 1, after the aggregation we have

$$\begin{aligned}\bar{\boldsymbol{x}}^1 &= \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^{0,K} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^* = \frac{1}{n}\sum_{i=1}^n \left(-\boldsymbol{A}^{-1}\boldsymbol{b}_i\right) \\ &= -\boldsymbol{A}^{-1}\boldsymbol{b} = \boldsymbol{x}^*,\end{aligned} \tag{6}$$

which means that after one aggregation, $\bar{\boldsymbol{x}}^1$ converges to the global minima $\boldsymbol{x}^*$.

From the above example, we can observe that although the heterogeneity on the **direction** of the gradient descent is large, which is shown by the large $\zeta$, there is no heterogeneity on the **destination**, since the averaged minima of the local objective functions $\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i^*$, is equal to $\boldsymbol{x}^*$. Therefore, one aggregation for all $\boldsymbol{x}_i^*$'s is sufficient. Motivated by these observations, we develop a convergence analysis for quadratic objective functions based on the heterogeneity on the minima in the following section.

## IV. Theoretical Results

In this section, we provide the new theoretical analysis of FedAvg for heterogeneous quadratic objective functions based on the heterogeneity on the minima of the objective functions. By the theoretical results, we show that the data heterogeneity does not only decrease the convergence speed but also causes the gap between the global model and the global minima $\boldsymbol{x}^*$, which dominates the convergence error. By extending the framework to the analysis for two-stage learning rates, we show that the optimal $\gamma$ is not close to zero while keep the product of $\gamma$ and $\eta$, which corrects the misunderstand of the two-stage learning rates in the literature.

### A. Convergence Analysis for Quadratic Functions

*1) Preliminaries:* In this section, we introduce the objective functions and the important relationship between $\boldsymbol{x}^*$ and the local minima $\boldsymbol{x}_i^*$. The local objective function is given by

$$f_i(\boldsymbol{x}) = c_i + \boldsymbol{b}_i^T\boldsymbol{x} + \frac{1}{2}\boldsymbol{x}^T\boldsymbol{A}_i\boldsymbol{x}, \boldsymbol{A}_i \succ 0, \forall i, \tag{7}$$

where $\boldsymbol{A}_i$'s can be different among workers. Since the Hessian matrix $\boldsymbol{A}_i$ is positive definite, $f_i(\boldsymbol{x})$ is strongly convex. By (1), for the global objective function, we have

$$\boldsymbol{A} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{A}_i \succ 0,\ \boldsymbol{b} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{b}_i,\ c = \frac{1}{n}\sum_{i=1}^n c_i. \tag{8}$$

For quadratic functions, the minima $\boldsymbol{x}_i^*$ of $f_i(\boldsymbol{x})$ and the minima $\boldsymbol{x}^*$ of $f(\boldsymbol{x})$ are respectively given by

$$\boldsymbol{x}_i^* = \boldsymbol{A}_i^{-1}\boldsymbol{b}_i,\ \boldsymbol{x}^* = \boldsymbol{A}^{-1}\boldsymbol{b}. \tag{9}$$

We can rewrite $\boldsymbol{x}^*$ as a weighted average of $\boldsymbol{x}_i^*$,

$$\boldsymbol{x}^* = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{A}^{-1}\boldsymbol{A}_i\boldsymbol{x}_i^*, \tag{10}$$

where

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{A}^{-1}\boldsymbol{A}_i = \boldsymbol{A}^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_i = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}. \tag{11}$$

It is shown by (10) that $\boldsymbol{x}^*$ is a weighted averaged of $\boldsymbol{x}_i^*$'s, where the weight of $\boldsymbol{x}_i^*$ is $\frac{1}{n}\boldsymbol{A}^{-1}\boldsymbol{A}_i$. In the example of Section III, since $\boldsymbol{A}_i = \boldsymbol{A}$, the weight is given by $\frac{1}{n}$.

*2) Convergence Analysis:* In this section, we provide the convergence analysis for heterogeneous quadratic objectives with one learning rate $\gamma$. We start with deriving the changes on the global model after one round as shown in Lemma 1.

*Lemma 1 (In One Round):* Given $\bar{\boldsymbol{x}}^r$, with $\gamma \in (0, \frac{1}{\max_i \lambda_i})$, after one round we have

$$\bar{\boldsymbol{x}}^{r+1} = \left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{I}-\gamma\boldsymbol{A}_i)^K\right]\bar{\boldsymbol{x}}^r + \frac{1}{n}\sum_{i=1}^{n}\left[\boldsymbol{I}-(\boldsymbol{I}-\gamma\boldsymbol{A}_i)^K\right]\boldsymbol{x}_i^*, \tag{12}$$

where $\lambda_i$ denotes the largest eigenvalue of $\boldsymbol{A}_i$.

To obtain the the global model after $r$ rounds, we apply the recursion to (12). Then we have the following lemma.

*Lemma 2 (After r Rounds):* Given $\bar{\boldsymbol{x}}^0$, with $\gamma \in (0, \frac{1}{\max_i \lambda_i})$, after $r$ rounds we have

$$\bar{\boldsymbol{x}}^r = \boldsymbol{D}_K^r\bar{\boldsymbol{x}}^0 + (\boldsymbol{I}-\boldsymbol{D}_K^r)\sum_{i=1}^{n}\boldsymbol{W}_i\boldsymbol{x}_i^*, \tag{13}$$

where

$$\boldsymbol{D}_K^r = \left[\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{I}-\gamma\boldsymbol{A}_i)^K\right]^r, \tag{14}$$

$$\boldsymbol{W}_i = \frac{1}{n}\left[\boldsymbol{I}-\frac{1}{n}\sum_{j=1}^{n}(\boldsymbol{I}-\gamma\boldsymbol{A}_j)^K\right]^{-1}\left[\boldsymbol{I}-(\boldsymbol{I}-\gamma\boldsymbol{A}_i)^K\right]. \tag{15}$$

We define $\hat{\boldsymbol{x}} := \sum_{i=1}^{n}\boldsymbol{W}_i\boldsymbol{x}_i^*$. It is worth noting that $\sum_{i=1}^{n}\boldsymbol{W}_i = \boldsymbol{I}$. Therefore, $\sum_{i=1}^{n}\boldsymbol{W}_i\boldsymbol{x}_i^*$ can be seen as a weighted average of $\boldsymbol{x}_i^*$'s with the weight matrix $\boldsymbol{W}_i$, which is determined by $\gamma$ and $K$. When $r$ increases, we have

$$\lim_{r\to\infty}\bar{\boldsymbol{x}}^r = \sum_{i=1}^{n}\boldsymbol{W}_i\boldsymbol{x}_i^* = \hat{\boldsymbol{x}}, \tag{16}$$

which means that given $\gamma$ and $K$, the global model $\bar{\boldsymbol{x}}^r$ will converge to $\hat{\boldsymbol{x}}$, which is called as the **convergence point**.

However, this implies that the global model might not be able to converge to the global minima $\boldsymbol{x}^*$. Only when $K = 1$, we have $\hat{\boldsymbol{x}} = \boldsymbol{x}^*$, which means the global model can converge to $\boldsymbol{x}^*$. As $K \to \infty$, we will have $\hat{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^*$, which means that the global model converges to $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^*$. We note that the difference between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}^*$ is bounded since

$$0 = \left\|\sum_{i=1}^{n}\boldsymbol{A}^{-1}\boldsymbol{A}_i\boldsymbol{x}_i^* - \boldsymbol{x}^*\right\| \le \left\|\hat{\boldsymbol{x}}-\boldsymbol{x}^*\right\| \le \left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^* - \boldsymbol{x}^*\right\|. \tag{17}$$

This property can help us understand the theoretical results in Theorem 1. Based on Lemma 2, we have the following theoretical results.

*Theorem 1:* For Algorithm 1, if the objective functions are defined as (7), with a constant learning rate $\gamma \in (0, \frac{1}{\lambda_{\max}})$, given $K$, after $R$ rounds we have

$$\begin{aligned}
&\|\bar{\boldsymbol{x}}^R - \boldsymbol{x}^*\|^2 \\
\le& 2\|\boldsymbol{D}_K^R\|^2\|(\bar{\boldsymbol{x}}^0-\boldsymbol{x}^*)\|^2 + 2\|\boldsymbol{I}-\boldsymbol{D}_K^R\|^2\Big\|\sum_{i=1}^{n}\boldsymbol{W}_i\boldsymbol{x}_i^* - \boldsymbol{x}^*\Big\|^2 \\
\le& \underbrace{2(1-\gamma\lambda_{\min})^{2KR}\|\bar{\boldsymbol{x}}^0-\boldsymbol{x}^*\|^2}_{\text{error caused by initialization}} \\
&+ \underbrace{2(1-(1-\gamma\lambda_{\max})^{KR})^2\left\|\hat{\boldsymbol{x}}-\boldsymbol{x}^*\right\|^2}_{\text{error caused by the convergence point}},
\end{aligned} \tag{18}$$

where $\lambda_{\max} = \max_i \lambda_i$, $\lambda_{\min} = \min_i \lambda_i^{\min}$ and $\lambda_i^{\min}$ is the minimum eigenvalue of $\boldsymbol{A}_i$.

*Remark 1 (Explanation for the Convergence Upper Bound):* The convergence bound in (18) is composed of two parts, the error caused by initialization and the error caused by the convergence point. Since $\gamma\lambda_{\min} < 1$, increasing either $K$ or $R$ can reduce the error caused by initialization. In contrast, the error caused by the convergence point increases as $R$ or $I$ increases. However, the impact of $R$ and $K$ is not exactly the same for the error caused by the convergence point. While the coefficient $1-(1-\gamma\lambda_{\max})^{KR}$ increases as $R$ increases, increasing $K$ does not only increase the coefficient, but also increases the gap between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}^*$. This is because that as shown in (17), when $K$ increases, $\|\hat{\boldsymbol{x}}-\boldsymbol{x}^*\|$ increases. In addition, when $K = 1$, for any $R \ge 1$, the error caused by the convergence point becomes zero. In contrast, when $K > 1$, we cannot find a $R > 1$ such that the error caused by the convergence point is zero. In addition, for $K > 1$, it has been shown in (17) that the error caused by the convergence point is upper bounded by a constant $\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^* - \boldsymbol{x}^*\|^2$. This is a new finding, since in previous results, the divergence term $O(\gamma^2K^2\zeta^2)$ in the upper bound grows unbounded with $K$.

*Remark 2 (New Insights for Data Heterogeneity):* By Theorem 1, the impact of data heterogeneity can be seen in two aspects. First, the data heterogeneity causes the gap between the convergence point and the minima. When data are IID, by the definition of $\hat{\boldsymbol{x}}$, we have $\hat{\boldsymbol{x}} = \boldsymbol{x}^*$, which means that the error caused by the convergence point is zero regardless of $K$. However, when data are non-IID, with $K > 1$ and a constant learning rate, the error caused by the convergence point is nonzero. Second, larger data heterogeneity may lead to a bigger error caused by initialization. When data are IID, we have $\lambda_{\min} = \bar{\lambda}_{\min}$, where $\bar{\lambda}_{\min}$ denotes the largest eigenvalue of the global Hessian matrix $\boldsymbol{A}$. In contrast, when data are non-IID, we have $1-\gamma\lambda_{\min} > 1-\gamma\bar{\lambda}_{\min}$.

Now we show how the choice of learning rate affects the upper bound in Theorem 1.

*Lemma 3:* Given the number of local iterations $K$,

$$\lim_{\gamma\to 0}\|\hat{\boldsymbol{x}}-\boldsymbol{x}^*\|^2 = 0. \tag{19}$$

Lemma 3 implies that a small learning rate can reduce the gap between $\boldsymbol{x}^*$ and the convergence point $\hat{\boldsymbol{x}}$. Based on Lemma 3, we explicitly rewrite the convergence upper bound as a function of $\gamma$, $R$ and $K$ as follows.

*Corollary 1:* For the convergence upper bound in Theorem 1, when $\gamma$ is sufficiently small, we have

$$\|\bar{\boldsymbol{x}}^R - \boldsymbol{x}^*\|^2 \leq 2e^{-2\gamma\lambda_{\min}KR}\|\bar{\boldsymbol{x}}^0 - \boldsymbol{x}^*\|^2 + 2C\gamma^2, \tag{20}$$

where $C$ is a constant and when $\boldsymbol{A}_j = \boldsymbol{A}, \forall j$, $C = 0$. Then the learning rate can be chosen as

$$\gamma = \frac{1}{(KR)^q}, q \in (0, 1). \tag{21}$$

For example, by choosing $q = \frac{1}{2}$, we will obtain $\|\bar{\boldsymbol{x}}^R - \boldsymbol{x}^*\|^2 = O(\frac{1}{KR})$, since the second term in (20) is the dominant term. We note that in [5, Theorem II], the lower bound for the convex quadratic objective function is $\Omega(\frac{1}{R^2})$. We can see that if $q$ approaches 1, $\gamma$ will go to $\frac{1}{KR}$, and the convergence rate will converge to $O(\frac{1}{R^2})$, which matches the lower bound with respect to $R$.

### B. Understanding the Two-Stage Learning Rates

In this section, we extend our analysis to FedAvg with two-stage learning rates. We found that the choices for $\eta$ and $\gamma$ implied by existing theoretical results [13] can lead to the divergence. To see this, we provide the convergence upper bound for non-convex local GD in [13, Theorem 1] as follows.

$$\min_{r\in[R]} \|\nabla f(\bar{\boldsymbol{x}}^r)\|^2 \leq \frac{f_0 - f_\star}{c\eta\gamma KR} + 15K^2\gamma^2L^2\zeta^2, \tag{22}$$

where $c$ is a constant, $L$ is the Lipschitz constant and $\eta\gamma \leq \frac{1}{KL}$. To minimize the convergence error, we should choose $\eta\gamma = \frac{1}{KL}$ which minimizes the first term. Then we should let $\gamma$ as small as possible since it makes the second term smaller.

However, as shown in Figure 1, if we keep the product of $\gamma$ and $\eta$, choosing a small $\gamma$ can lead to the divergence of FedAvg. This shows that in the literature, the impact of the two-stage learning rates is not well understood. To obtain a deeper understanding for the two-stage learning rates, we extend our analysis to FedAvg with two-stage learning rates. The results are shown in Theorem 2.

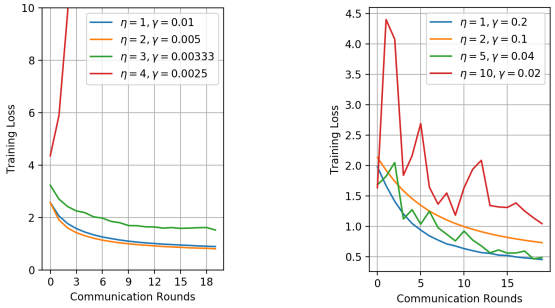
*Theorem 2:* For Algorithm 1 with two-stage learning rates, if the objective functions are defined as (7), with a constant local learning rate $\gamma \in (0, \frac{1}{\lambda_{\max}})$, a constant global learning rate $\eta \in (0, \frac{1}{1-\frac{1}{n}\sum_{i=1}^n(1-\gamma\lambda_i)^K})$, given $K$, after $R$ rounds,

$$\|\bar{\boldsymbol{x}}^R - \boldsymbol{x}^*\|^2 \leq \underbrace{2\|\boldsymbol{D'}_K^R\|^2\|\bar{\boldsymbol{x}}^0 - \boldsymbol{x}^*\|^2}_{\text{error caused by initialization}}$$
$$\underbrace{+ 2\|\boldsymbol{I} - \boldsymbol{D'}_K^R\|^2\|\sum_{i=1}^n \boldsymbol{W}_i\boldsymbol{x}_i^* - \boldsymbol{x}^*\|^2}_{\text{error caused by the convergence point}}, \tag{23}$$

where $\boldsymbol{D'}_K^R = \left(\boldsymbol{I} - \eta\left[\boldsymbol{I} - \frac{1}{n}\sum_{i=1}^n(\boldsymbol{I} - \gamma\boldsymbol{A}_i)^K\right]\right)^R$.

In this case, according to the range of $\gamma$ and $\eta$, we have

$$\|\boldsymbol{D'}_K^R\| \in (0, 1), \|\boldsymbol{I} - \boldsymbol{D'}_K^R\| \in (0, 1). \tag{24}$$




(a) Linear model. (b) Two-layer neural network.

Fig. 1: Empirical results for the two-stage learning rates. For each curve, we keep the product $\eta\gamma$ the same and we set $K = 10$. In (a), $\eta = 2, \gamma = 0.005$ converges fastest while $\eta = 4, \gamma = 0.0025$ cannot converge. In (b), $\eta = 1, \gamma = 0.2$ converges fastest while $\eta = 10, \gamma = 0.02$ is the worst.

*Remark 3 (Advantage of Two-stage Learning Rates):* First, we compare Theorem 1 and Theorem 2 to show the advantage of two-stage learning rates. It can be seen that the main difference is on the coefficients, where $\boldsymbol{D}_K^R$ is substituted by $\boldsymbol{D'}_K^R$ and $\boldsymbol{D'}_K^R$ depending on $\eta$. In Theorem 1, we have shown that when $\gamma$ decreases, the error caused by initialization increases while the error caused by the convergence point decreases. However, in Theorem 2, when $\gamma$ decreases, we can still choose a large $\eta$ such that both the error caused by initialization and the error caused by the convergence point can be smaller. Therefore, it can be seen that by choosing a large $\eta$ and a small $\gamma$, the two-stage learning rates can help improve the convergence rate. It is worth noting that $\eta$ is upper bounded and cannot be arbitrarily large when fixing the product of $\eta$ and $\gamma$, which is not shown in the literature. More explanations about the range of learning rates are provided in the following.

*Remark 4 (New Insights for Two-stage Learning Rates):* From Theorem 2, it can be seen that the upper bound of $\eta$ depends on different parameters. However, the impact of each parameter is different from that shown in the literature. First, in the literature [13], the upper bound of $\eta$ is given by $\eta \leq \frac{1}{\gamma KL}$, which is not affected by the data heterogeneity, while in Theorem 2, we show that the upper bound of $\eta$ is affected by the data heterogeneity. The reason is that since $\frac{1}{n}\sum_{i=1}^n(1-\gamma\lambda_i)^K) < 1 - \gamma\bar{\lambda}_{\max}$, where $\bar{\lambda}_{\max}$ is the largest eigenvalue of $\boldsymbol{A}$, a larger heterogeneity can lead to a smaller upper bound of $\eta$.

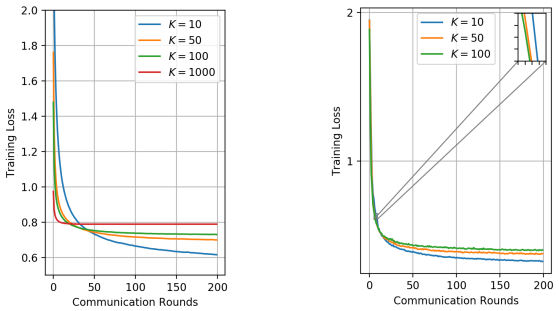
Second, the results in Theorem 2 imply that when $K > 1$, we cannot increase $\eta$ to a infinitely large value while fixing the product of $\eta$ and $\gamma$. When $K > 1$. we have a smaller upper bound for $\eta$ since $\frac{1}{1-\frac{1}{n}\sum_{i=1}^n(1-\gamma\lambda_i)^K} < \frac{1}{\gamma\frac{1}{n}\sum_{i=1}^n\lambda_i}$. Only when $K = 1$, we can have $\eta < \frac{1}{\gamma\frac{1}{n}\sum_{i=1}^n\lambda_i}$. This also explains why in Figure 1a, the curve with $\eta = 4, \gamma = 0.0025$ diverges. Since we fix the product of $\eta$ and $\gamma$ in this figure, $\eta = 4$ is greater than $\frac{1}{1-\frac{1}{n}\sum_{i=1}^n(1-\gamma\lambda_i)^K}$ for $K = 10$. Therefore, the convergence with $\eta = 4$ cannot be guaranteed.

Third, the convergence upper bound in Theorem 2 provides insights into how to choose the optimal $\eta$ and $\gamma$. When the

error caused by initialization is large, we can choose a large $\eta$ such that the coefficient $\|\boldsymbol{D}'^{R}_{K}\|$ is small. When the error caused by the convergence point is large, we need to choose a small $\eta$ and a small $\gamma$, such that both the coefficient $\|\boldsymbol{I}-\boldsymbol{D}'^{R}_{K}\|$ and the gap between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}^*$ become small.

## V. Experiments

We provide the experimental results with the MNIST dataset, which consists of 10 classes. The number of workers is $n = 10$. The dataset is partitioned in a non-IID manner such that there is only one data class at each worker. We use linear regression with mean square error (MSE) loss (quadratic objective function) and a two-layer neural network with cross-entropy loss (non-convex objective function). The learning rates are set as $\gamma = 0.01$ and $\gamma = 0.1$ for linear regression and neural network, respectively.




(a) Linear model. (b) Two-layer neural network.

Fig. 2: Empirical results with different $K$.

Figure 2a shows the results with linear regression and MSE loss. It can be seen that the larger $K$ means the global model can converge faster to $\hat{\boldsymbol{x}}$. This is consistent with our convergence upper bound in Theorem 1 since the first term of (18) is an exponential function of $K$, a larger $K$ implies a faster decay of the first term. Figure 2a also shows that a smaller $K$ can lead to a smaller training loss when $R$ is sufficiently large. This is because that as shown in the second term of (18), a smaller $K$ can reduce the gap between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}$, which means the convergence point is closer to the global minima so the loss on the convergence point is smaller. It is worth noting that the difference between curves of $K = 10$ and $K = 50$ is far more than that between $K = 50$ and $K = 100$. This is because that as shown in Theorem 1, increasing $K$ makes the $\hat{\boldsymbol{x}}$ closer to $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^*$ and the difference between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}^*$ is bounded by $\|\boldsymbol{x}^* - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^*\|$, which will not increase as $K$ becomes larger. Figure 2b shows the results with a two-layer neural network and cross-entropy loss. It can be seen that the observations obtained from Figure 2a can also be applied to Figure 2b. This means that the insights shown by our theoretical results have the potential to be extended to the non-convex objective functions.

## VI. Conclusion

In this paper, a new framework for the analysis of FedAvg has been proposed. For heterogeneous quadratic objective functions, we have derived a new convergence upper bound, which shows that the data heterogeneity does not only lead to the gap between the convergence point $\hat{\boldsymbol{x}}$ and the global minima $\boldsymbol{x}^*$, but also decreases the decaying coefficient of the error caused by the initialization. We have extended the new framework to the analysis for the two-stage rates. The theoretical results reveal the insights behind the global learning rate $\eta$ and show that the optimal choice for the learning rates is not fixing the product of $\gamma$ and $I$ and let $\gamma$ become as small as possible as a common understanding shown the literature. The experiments have validated our theoretical results and showed that our results have the potential to be applied to the general non-convex objective functions. Future works include extending the analysis to the general non-convex objective functions and considering the stochastic gradients.


## Acknowledgement

This work was supported in part by NSF CAREER Award 2145835 and NSF Award 2312227.


## References


[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, pp. 1273–1282.

[2] S. U. Stich, "Local SGD converges fast and communicates little," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=S1g2JnRcFX

[3] T. Lin, S. U. Stich, K. K. Patel, and M. Jaggi, "Don't use large mini-batches, use local sgd," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=B1eyO1BFPr

[4] P. Kairouz, H. B. McMahan *et al.*, "Advances and open problems in federated learning," *arXiv preprint arXiv:1912.04977*, 2019.

[5] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5132–5143.

[6] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," 2020.

[7] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of fedavg on non-iid data," in *International Conference on Learning Representations*, 2020.

[8] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Adaptive federated learning in resource constrained edge computing systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1205–1221, 2019.

[9] Haddadpour, Farzin *et al.*, "Local sgd with periodic averaging: Tighter analysis and adaptive synchronization," in *Advances in Neural Information Processing Systems*, 2019.

[10] H. Yu, S. Yang, and S. Zhu, "Parallel restarted SGD with faster convergence and less communication: Demystifying why model averaging works for deep learning," in *AAAI*, Jan.–Feb. 2019.

[11] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *arXiv preprint arXiv:1908.07873*, 2019.

[12] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," 2020.

[13] H. Yang, M. Fang, and J. Liu, "Achieving linear speedup with partial worker participation in non-iid federated learning," in *International Conference on Learning Representations*, 2020.

[14] J. Wang, R. Das, G. Joshi, S. Kale, Z. Xu, and T. Zhang, "On the unreasonable effectiveness of federated averaging with heterogeneous data," 2022.

[15] H. Yu, R. Jin, and S. Yang, "On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization," in *ICML*, Jun. 2019, pp. 7184–7193.