FedDuA: Doubly Adaptive Federated Learning

Shokichi Takakura
LY Corporation

Seng Pei Liew
LY Corporation

Satoshi Hasegawa
LY Corporation

Abstract

Federated learning (FL) is a distributed learning framework where clients collaboratively train a global model without sharing their raw data. FedAvg is a popular algorithm for FL, but it often suffers from slow convergence due to the heterogeneity of local datasets and anisotropy in the parameter space. In this work, we formalize the central server optimization procedure through the lens of mirror descent and propose a novel framework, called *FedDuA*, which adaptively selects the global learning rate based on both inter-client and coordinate-wise heterogeneity in the local updates. We prove that our proposed doubly adaptive step-size rule is minimax optimal and provide a convergence analysis for convex objectives. Although the proposed method does not require additional communication or computational cost on clients, extensive numerical experiments show that our proposed framework outperforms baselines in various settings and is robust to the choice of hyperparameters.

1 INTRODUCTION

Federated Learning (FL) (Konečnỳ et al., 2016) is a distributed optimization framework where multiple clients collaboratively train a global model under the coordination of a central server. In FL, clients have their own local datasets and only send model updates to a central server and never share their raw data, which enhances the privacy of the data. In FL, there are two primary categories: *cross-silo* FL and *cross-device* FL (Kairouz et al., 2021). In this paper, we focus on cross-device FL, which is more challenging due to the limited computational resources and communication bandwidth on clients.

Federated Averaging (FedAvg) (McMahan et al., 2017) is one of the most popular algorithms for FL due to its simplicity, stateless properties, and communication efficiency. In FedAvg, clients perform multiple local training steps before sending the local updates to the server, which significantly reduces the communication cost compared to distributed SGD. However, FedAvg often suffers from slow convergence due to (1) *client heterogeneity* and (2) *gradient heterogeneity*. The former refers to non-i.i.d. data distribution across clients, which leads to so-called *client drift error* (Kairouz et al., 2021). The latter refers to the anisotropic nature of gradients, meaning that gradients have different scales or importance across different parameter dimensions, which often hinders the convergence of SGD (Zhang et al., 2020, 2024).

A line of work has dealt with client heterogeneity by introducing control variates or client-side momentum to reduce the client drift (Karimireddy et al., 2019, 2020a; Mishchenko et al., 2022; Cheng et al., 2024; Zaccone et al., 2025). While these methods are effective in *cross-silo* FL, they are not practical in *cross-device* FL since they require clients to be stateful or increase the communication and computational cost on clients. Thus, it is desirable to design an algorithm on a central server to overcome the client heterogeneity without changing the local training procedure. From this standpoint, Jhunjhunwala et al. (2023) have recently proposed FedExP, which accelerates FedAvg by selecting the global learning rate adaptively to the client heterogeneity.

To deal with the gradient heterogeneity, Reddi et al. (2021) have proposed a federated version of adaptive optimizers including FedAdagrad, FedAdam, FedYogi inspired by the success of adaptive methods in centralized optimization. These methods adaptively adjust the coordinate-wise learning rate based on the historical local updates. They have shown that coordinate-wise adaptivity can improve the performance of FL, especially for tasks with sparse gradients.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

Although adaptivity to both client and gradient heterogeneity is shown to be crucial to achieve fast convergence (Jhunjhunwala et al., 2023; Reddi et al., 2021), most existing works focus on either client or gradient heterogeneity. Thus, we pose the following question: *How can we incorporate two types of adaptivity without additional computational and communication cost at clients?*

To answer this question, we propose *FedDuA*, a novel framework which adaptively selects the global learning rate based on both the client and gradient heterogeneity. A key ingredient of our method is the mirror descent formulation of the global update procedure, which provides a unified view of both types of adaptivity. We numerically and theoretically show that dual adaptivity is essential to achieve better performance. In contrast to some existing methods (Karimireddy et al., 2020b; Qu et al., 2022), FedDuA does not require additional computational or communication cost on clients. See Table 1 for the comparison with existing methods. In addition, we would like to emphasize that FedDuA is *orthogonal* to the existing methods which improve the performance by modifying the local training procedure (Karimireddy et al., 2020a, 2019; Li et al., 2020; Mishchenko et al., 2022). Thus, FedDuA can be combined with them to further improve the performance as shown in our experiments.

Main contributions Our contribution can be summarized as follows:

- We propose FedDuA, a novel framework which adaptively selects the global learning rate based on both the inter-client and coordinate-wise heterogeneity, based on the mirror descent formulation of the global update procedure.
- We show that the update rule of FedDuA is minimax optimal under the approximate projection condition and show the benefit of adaptivity by providing the convergence analysis.
- We conduct extensive experiments on various datasets and show that FedDuA consistently outperforms existing adaptive methods. We also show that FedDuA can be combined with existing techniques to further improve the performance and is robust to the choice of hyperparameters due to its dual adaptivity.

1.1 Other Related Work

Adaptive Methods Adaptive methods like AdaGrad (Duchi et al., 2011), Adam (Kingma and Ba, 2015), and Yogi (Zaheer et al., 2018) have been widely used in centralized optimization. Inspired by the success in centralized optimization, several works have utilized adaptive methods in FL. For instance, Xie et al. (2019) have proposed AdaAlter, which replaces the local SGD with an adaptive optimizer such as AdaGrad. On the other hand, Reddi et al. (2021) have proposed to use such adaptive methods as a global optimizer in FL. Several works (Lee et al., 2024; Wang et al., 2021) have unified the above strategies and used adaptive optimizers at both the client and server. However, these methods are not adaptive to the heterogeneity among clients.

Mirror Descent Mirror descent is a generalization of gradient descent (Hazan et al., 2016) and has been adopted in various applications including online learning (Hazan et al., 2016), reinforcement learning (Tomar et al., 2022), and differentially private optimization (Odeyomi and Zaruba, 2021; Amid et al., 2022). In the context of FL, Yuan et al. (2021) have proposed FedMid and FedDualAvg with local mirror descent, but these methods are tailored for composite optimization problems and do not consider adaptive step-size selection, which is the focus of our work.

Notation For a vector $x \in \mathbb{R}^d$, $[x]_k$ denotes the k-th element of x, $\|x\|_p$ denotes the p-norm of x, and $\|x\|_G$ denotes $\sqrt{x^\top G x}$ for a positive semi-definite matrix $G \in \mathbb{R}^{d\times d}$. We use $\|x\|$ as a shorthand for $\|x\|_2$. For $s \in \mathbb{R}^d$, we write $\sqrt{s}$, s^{-1} and $s + a$ ($a \in \mathbb{R}$) as the element-wise square root, inverse and addition respectively.

2 PROBLEM FORMULATION AND PRELIMINARIES

In this paper, we consider the following FL problem:

$$\min_{w\in\mathbb{R}^d} F(w) := \frac{1}{M}\sum_{i=1}^{M} F_i(w),$$

where $F_i(w) = \mathbb{E}_{z\sim\mathcal{D}_i}[f_i(w,z)]$ is the loss function and $\mathcal{D}_i$ is the data distribution of the i-th client. The number of clients is denoted by M and the dimension of the parameter is denoted by d.

FedAvg To solve the above optimization problem, we consider FedAvg (McMahan et al., 2017), which is a standard algorithm for FL. At each round t, the server sends the global model w_t to all clients. Then, each client performs τ-steps of local training using SGD to obtain the local updates $\{\Delta_i^t\}_{i=1}^M$ as follows:

$$\text{Run Local SGD: } w_{i,k+1}^t = w_{i,k}^t - \eta_l \nabla f_i(w_{i,k}^t, z_k)$$
$$\text{Compute Update: } \Delta_i^t = w_{i,\tau}^t - w_t,$$

Table 1: Comparison of adaptive methods in FL.

Algorithms	Adaptivity		
	Coordinate-wise	Inter-client	No extra client cost
AdaAlter (Xie et al., 2019)	✓		
Mime, SCAFFOLD (Karimireddy et al., 2020a,b)		✓	
FedOpt (Reddi et al., 2021)	✓		✓
FedExP (Jhunjhunwala et al., 2023)		✓	✓
FedDuA (ours)	✓	✓	✓

where $w_{i,0}^t = w_t$ and $z_k \sim \mathcal{D}_i$ $(k = 0, \dots, \tau - 1)$. When clients complete the local training, they send the local updates to the server and the server aggregates the local updates to obtain the global update $\bar{\Delta}_t = \frac{1}{M}\sum_{i=1}^M \Delta_i^t$. In FedAvg, the central server updates the global model as follows:

$$\text{Global Update (FedAvg): } w_{t+1} = w_t + \eta_g \bar{\Delta}_t,$$

where η_g is the global learning rate. Vanilla FedAvg uses $\eta_g = 1$, but in practice, $\eta_g > 1$ is often used to improve the convergence. FedAvg enjoys some favorable properties such as statelessness and communication efficiency, but it often suffers from slow convergence due to 1) the anisotropic nature of the local updates and 2) the heterogeneity of client data distributions. We call the former *gradient heterogeneity* and the latter *client heterogeneity*. Considering limited computational resources and communication bandwidth on clients, it is desirable to design an algorithm on a central server to overcome the above issues without changing the local training procedure.

FedOpt To deal with the gradient heterogeneity, Reddi et al. (2021) have introduced a general framework called FedOpt. In this framework, the aggregated update $\bar{\Delta}_t$ is regarded as a pseudo-gradient and adaptive methods such as AdaGrad and Adam are used as a global optimizer. General update rule of FedOpt is given by

$$\text{Global Update (FedOpt): } w_{t+1} = w_t + \eta_g G_t^{-1} v_t$$

where $G_t := \mathrm{diag}(\sqrt{s_t} + \epsilon)$ $(s_t \in \mathbb{R}^d, \epsilon > 0)$ is a time-dependent preconditioner and v_t is the aggregated update or momentum term. Here, $\epsilon > 0$ is a small constant for numerical stability. In FedAdagrad and FedAdam, s_t and v_t are defined as

$$\begin{aligned} \text{FedAdagrad: } & s_t = s_{t-1} + \bar{\Delta}_t^2,\ v_t = \bar{\Delta}_t. \\ \text{FedAdam: } & s_t = \beta_2 s_{t-1} + (1-\beta_2)\bar{\Delta}_t^2, \\ & v_t = \beta_1 v_{t-1} + (1-\beta_1)\bar{\Delta}_t, \end{aligned}$$

where $s_{-1}, v_{-1} = 0$ and $\beta_1, \beta_2 \in [0, 1)$ are hyperparameters.

FedExP To deal with the client heterogeneity, Jhunjhunwala et al. (2023) have proposed FedExP, which adaptively selects the global learning rate η_g in FedAvg based on the client heterogeneity. FedExP updates the global model as follows:

$$\text{Global Update (FedExP): } w_{t+1} = w_t + \eta_g^t \bar{\Delta}_t$$

with adaptive global learning rate $\eta_g^t = \frac{\frac{1}{2M}\sum_{i=1}^M \left\|\Delta_i^t\right\|^2}{\left\|\bar{\Delta}_t\right\|^2 + \epsilon_g}$. Here, ϵ_g is a small constant to avoid blow-up of the step size. If clients perform one step of local training with full-batch SGD and $\epsilon_g = 0$, $\Delta_i^t = \eta_l \nabla F_i(w_t)$ and η_g^t is reduced to $\frac{\frac{1}{2M}\sum_{i=1}^M \|\nabla F_i(w_t)\|^2}{\|\nabla F(w_t)\|^2}$, which is known as the measure of the heterogeneity among clients (Haddadpour and Mahdavi, 2019). Note that the above formula is tailored for FedAvg and not applicable to general FedOpt algorithms. This is because the learning rate in FedExP is derived based on norms of raw local and global updates, whereas FedOpt transforms these updates before applying them, making the original FedExP formulation incompatible.

3 GENERALIZED FORMULATION AND PROPOSED METHOD

Previous works have considered adaptivity to 1) gradient heterogeneity and 2) client heterogeneity separately. This paper unifies these two types of adaptivity through a generalized formulation of the global update procedure.

Mirror descent formulation As discussed in the original paper of Adagrad (Duchi et al., 2011), adaptive methods can be viewed as a generalized version of Mirror Descent. That is, the update rule of FedOpt can be written as

$$\begin{aligned} \text{Mirror Map: } & \theta_t = \nabla \psi_t(w_t), \\ \text{Dual Variable Update: } & \theta_{t+1} = \theta_t + \eta_g v_t, \\ \text{Inverse Mirror Map: } & w_{t+1} = \nabla \psi_t^{-1}(\theta_{t+1}) \end{aligned}$$

where $\psi_t(x) = \frac{1}{2} x^\top G_t x$ is a time-dependent distance-generating function and $\nabla \psi_t(x)$ is called the mirror

map. We define $\theta_t := \nabla\psi_t(w_t)$ as the dual variable and $\phi_t(\theta) := \max_w \langle w, \theta\rangle - \psi_t(w)$ as the convex conjugate of ψ_t. Note that the distance-generating function for adaptive methods is often chosen as the quadratic form but general smooth and strictly convex functions can be used in this formulation. In this paper, we focus on the quadratic case but also consider the general convex case and our proposed framework can be applied to them. For simplicity, we assume that ψ_t and ϕ_t are defined on $\mathbb{R}^d$, smooth and strongly convex, which is satisfied by the quadratic case with $G_t \succ 0$.

Given a smooth and strictly convex function $\psi_t(x)$, geometry of the parameter space is naturally induced by the Bregman divergence defined as

$$D_{\psi_t}(x \mid y) = \psi_t(x) - \psi_t(y) - \nabla\psi_t(y)^\top(x - y).$$

The Bregman divergence can be viewed as a generalization of the Euclidean (squared) distance. In fact, when $\psi_t(x) = \frac{1}{2}x^\top x$, the Bregman divergence reduces to the Euclidean distance.

Proposed Method A natural question is then how to choose the global learning rate η_g to minimize the distance between w_{t+1} and an optimal solution w^*. In Jhunjhunwala et al. (2023), the distance is measured by Euclidean distance in the parameter space. However, this choice of distance is not always appropriate especially in modern machine learning tasks since the geometry of the parameter space is often anisotropic (Zhang et al., 2024; Tomihari and Sato, 2025). Thanks to our mirror-descent formulation, we can use the Bregman divergence to measure the distance between two points. That is, we consider the following minimization problem over η_g:

$$\min_{\eta_g} D_{\psi_t}(w^* \mid w^{t+1}) = \min_{\eta_g} D_{\phi_t}(\theta_t + \eta_g v_t \mid \theta^*).$$

Here, the equality follows from the duality (Nielsen et al., 2007).

As discussed in Jhunjhunwala et al. (2023), FedAvg can be interpreted as a generalized Projection onto Convex Sets (POCS) algorithm in overparameterized convex optimization problems, where the set of minimizers for each local objective S_i is convex and the objective is to find a common minimizer $w^* \in \cap_{i=1}^M S_i$. That is, clients perform approximate projection onto S_i through local SGD and the server aggregates them to find a shared minimizer w^*. Inspired by the above relation, we consider the *(strong) approximate projection condition* (A.P.C.):

Assumption 3.1. Let $w^* = \arg\min F(w)$ be the optimal solution of the global problem.

$$\text{A.P.C.} \quad \frac{1}{M}\sum_{i=1}^M \left\|w_t + \Delta_i^t - w^*\right\|^2 \le \|w_t - w^*\|^2, \quad (1)$$

$$\text{strong A.P.C.} \quad \langle \bar{\Delta}_t, w^* - w_t\rangle \ge \frac{1}{M}\sum_{i=1}^M \left\|\Delta_i^t\right\|^2 \quad (2)$$

Intuitively, A.P.C. means that the local models $w_t + \Delta_i^t$ after local training are closer to the optimal solution w^* on average. A.P.C. can be reduced to $\langle \bar{\Delta}_t, w^* - w_t\rangle \ge \frac{1}{2M}\sum_{i=1}^M \|\Delta_i^t\|^2$. Thus, strong A.P.C. is indeed a stronger condition than A.P.C. As shown in Lemma 1 and Section C.4.1 of Jhunjhunwala et al. (2023), A.P.C. is satisfied in the case of overparameterized convex optimization case, where local objectives share a common minimizer w^*, and strong A.P.C. is satisfied if the local training is the exact projection onto the set of optimal solutions of the local objective. Overparameterization is often satisfied in modern neural network models, where the number of parameters is much larger than the number of training data (Jacot et al., 2018). Note that we consider the above condition to motivate our proposed method and we prove later the convergence guarantee under much milder conditions.

Although local training at clients and approximate projection condition are agnostic to the choice of ψ_t, we can derive the non-trivial lower bound on the optimal step size for general choice of ψ_t for both cases with and without momentum.

Theorem 3.2 (Lower Bound on the Optimal Step Size)**.** *Let $v_t = \bar{\Delta}_t$, $h_t(\eta) := \frac{\mathrm{d}}{\mathrm{d}\eta}\phi_t(\theta_t + \eta v_t) - \langle v_t, w_t\rangle$, and $m_t = \frac{1}{2M}\sum_{i=1}^M \|\Delta_i^t\|^2$. Assume that A.P.C. in Assumption 3.1 holds and $v_t \neq 0$. Then, $\eta_g^* := \arg\min D_{\phi_t}(\theta_t + \eta_g\bar{\Delta}_t \mid \theta^*)$ is uniquely defined and satisfies*

$$\eta_g^* \ge h_t^{-1}(m_t).$$

Theorem 3.3 (Lower Bound on the Optimal Step Size with Momentum)**.** *For $s = 0, \dots, t$, let $v_s = (1-\beta_1)\bar{\Delta}_s + \beta_1 v_{s-1}$ ($v_{-1} = 0$), $h_s(\eta) := \frac{\mathrm{d}}{\mathrm{d}\eta}\phi_s(\theta_s + \eta v_s) - \langle v_s, w_s\rangle$, and $m_s = \frac{1-\beta_1}{2M}\sum_{i=1}^M \|\Delta_i^s\|^2 + \frac{\beta_1}{2}m_{s-1}$ ($m_{-1} = 0$). Assume that strong A.P.C. in Assumption 3.1 holds and $v_t \neq 0$. We further assume that at round $s = 0, \dots, t-1$, w_s is updated as Eq. (3) with $\eta_g^s \le h_s^{-1}(m_s)$. Then, $\eta_g^* := \arg\min D_{\phi_t}(\theta_t + \eta_g v_t \mid \theta^*)$ is uniquely defined and satisfies*

$$\eta_g^* \ge h_t^{-1}(m_t).$$

See Appendix A and B for the proof. Here, h_t^{-1} is well-defined since ϕ_t is assumed to be strongly convex and $v_t \neq 0$. Similar analysis can be found in Section 3.2

and Lemma 11 of Jhunjhunwala et al. (2023) but they only consider FedAvg(M) and l^2-norm. Our contribution is to extend this analysis to more general mirror descent framework. To deal with the non-linearity of the global update in the parameter space, we utilize the dual form of the Bregman divergence to derive the lower bound. Note that it is essential to measure the distance with the Bregman divergence since we cannot derive a non-trivial lower bound on the optimal step size if we use the Euclidean distance, which is *not* compatible with the global update procedure. The mirror descent formulation allows us to use an appropriate distance measure for a given global optimizer.

Motivated by the above result, we propose *FedDuA*, which uses the lower bound on the optimal step size as the global learning rate.

$$\begin{aligned} \text{Mirror Map: } & \theta_t = \nabla \psi_t(w_t), \\ \text{FedDuA Update: } & \theta_{t+1} = \theta_t + \eta_g^t v_t, \\ \text{Inverse Mirror Map: } & w_{t+1} = \nabla \psi_t^{-1}(\theta_{t+1}) \end{aligned} \tag{3}$$

where $\eta_g^t := h_t^{-1}(m_t)$. For quadratic case $\psi_t(x) = \frac{1}{2}x^\top G_t x$, the lower bound is calculated as $\frac{m_t}{\|v_t\|^2_{G_t^{-1}}}$ in both cases, with and without momentum. In practice, we can use $\frac{m_t}{\|v_t\|^2_{G_t^{-1}+\epsilon_g}}$ with small constant $\epsilon_g > 0$ for stability. We provide the detailed algorithm in Algorithm 1. As shown in Algorithm 1, FedDuA is compatible with partial participation of clients, where only a subset of clients participate in each round. For simplicity, we only provide the algorithms with Adagrad and Adam as the global optimizer, which we call FedDuAdagrad and FedDuAdam respectively. Note that the FedDuA framework is very versatile, allowing for the creation of new algorithms by combining it with various global optimizers through modification of ψ_t. In addition, while we use SGD as a local optimizer here, we can use other local optimizers such as SCAFFOLD and Adam for local training.

Algorithm 1 FedDuA

Require: Initial model w_0, local learning rate η_l, number of local steps τ, small constants ϵ, ϵ_g

Set $s_{-1} = 0 \in \mathbb{R}^d$, $v_{-1} = 0 \in \mathbb{R}^d, m_{-1} = 0 \in \mathbb{R}$
for $t = 0, 1, \ldots, T-1$ **do**
  Server sends w_t to all clients
  for each client i in parallel **do**
    Perform τ steps of local SGD to compute Δ_i^t
    Send Δ_i^t to the server
  end for
  Sample a set S_t of participating clients
  Server aggregates updates: $\bar{\Delta}_t = \frac{1}{|S_t|}\sum_{i\in S_t}\Delta_i^t$
  Update s_t, v_t, and m_t
  FedDuAdagrad:
    $s_t = s_{t-1} + \bar{\Delta}_t^2$, $v_t = \bar{\Delta}_t$, $m_t = \frac{1}{2|S_t|}\sum_{i\in S_t}\|\Delta_i^t\|^2$
  FedDuAdam:
    $s_t = \beta_2 s_{t-1} + (1-\beta_2)\bar{\Delta}_t^2$, $v_t = \beta_1 v_{t-1} + (1-\beta_1)\bar{\Delta}_t$, $m_t = \frac{\beta_1}{2}m_{t-1} + \frac{1-\beta_1}{2|S_t|}\sum_{i\in S_t}\|\Delta_i^t\|^2$
  Compute preconditioner $G_t = \mathrm{diag}(\sqrt{s_t} + \epsilon)$
  Compute global learning rate $\eta_g^t = \frac{m_t}{\|v_t\|^2_{G_t^{-1}+\epsilon_g}}$
  Update global model: $w_{t+1} = w_t + \eta_g^t G_t^{-1} v_t$
end for

4 THEORETICAL ANALYSIS

In this section, we conduct detailed theoretical analysis of FedDuA and show that dual adaptivity is essential to achieve better performance.

4.1 Minimax Optimality

Here, we show that our proposed global update procedure is minimax optimal under the approximate projection condition.

Theorem 4.1 (Minimax Optimality)**.** *For a given global model w_t and local updates $\Delta_i^t \in \mathbb{R}^d$, define $H := \{w^* \mid A.P.C.\ is\ satisfied\}$ and worst-case distance difference $V(w) := \sup_{w^*\in H}[D_{\psi_t}(w^* \mid w) - D_{\psi_t}(w^* \mid w_t)]$. Then, for any ψ_t and $\{\Delta_i^t\}_{i=1}^M$ such that H is not empty, there exists a unique minimizer w_{t+1}^* of $V(w)$ and it matches w_{t+1} of FedDuA defined as in Eq. (3). On the other hand, if the global update procedure is different from that of FedDuA, it is suboptimal in the sense that $V(w_{t+1}) > V(w_{t+1}^*)$.*

See Appendix C for the proof. Theorem 4.1 shows our proposed global update rule performs best in the worst-case scenario. On the other hand, existing methods with partial adaptivity such as FedOpt and FedExP are suboptimal if their update differs from that of FedDuA, since the minimizer of $V(w)$ is unique. Thus, the dual adaptivity is provably necessary to update the global model optimally.

4.2 Convergence Analysis

Here, we prove the convergence guarantee of FedDuA with general distance-generating functions under the following standard assumptions.

Assumption 4.2 (L-smoothness and Bounded Data Heterogeneity at the optimum)**.** Local loss function $F_i(w)$ is differentiable and L-smooth. That is, for all $w, w' \in \mathbb{R}^d$, $\|\nabla F_i(w) - \nabla F_i(w')\|_2 \le L\|w - w'\|_2$. In addition, the norm of the local gradient at the optimum w^* is bounded as $\frac{1}{M}\sum_{i=1}^M \|\nabla F_i(w^*)\|^2 \le \sigma_*^2$

Theorem 4.3. *Assume that Assumption 4.2 holds, $\{F_i\}_{i=1}^M$ are convex, and clients use full-batch SGD and participate in every round. Then, if $\eta_l \le \frac{1}{6\tau L}$, $\{w^t\}_{t=1}^T$ generated by FedDuA satisfies*

$$F(\bar{w}_T) - F(w^*) = \underbrace{O\left(\frac{D_{\psi_0}(w^* \mid w_0) + \sum_{t=1}^{T-1} \delta_t}{\sum_{t=0}^{T-1} \eta_g^t \eta_l \tau}\right)}_{:=T_1} + \underbrace{O(\eta_l \tau \sigma_*^2)}_{:=T_2} + \underbrace{O(\eta_l^2 \tau(\tau-1) L \sigma_*^2)}_{:=T_3},$$

where $\bar{w}_T = \frac{\sum_{t=0}^{T-1} \eta_g^t w_t}{\sum_{t=0}^{T-1} \eta_g^t}$ and $\delta_t = D_{\psi_t}(w^* \mid w_t) - D_{\psi_{t-1}}(w^* \mid w_t)$.

See Appendix D for the proof. Technically, the difficulty of the proof lies in the fact that the global learning rate η_g^t does not have a closed form solution in general, which is in contrast to the case of FedExP. To overcome this issue, we leverage the duality and the convexity of Bregman divergence and bound the improvement at each round. Our theoretical analysis excludes the stochasticity by assuming full-batch SGD and full participation of clients since theoretical analysis becomes more complicated if the global learning rate is stochastic, as discussed in Jhunjhunwala et al. (2023). However, these assumptions are made solely for theoretical analysis and we empirically demonstrate that FedDuA performs effectively with stochastic local updates and partial client participation. While we focus on convex objectives here since the analysis provides clear insights into the benefit of dual adaptivity, we also provide the convergence guarantee for non-convex objectives in Appendix G.

Comparison with FedExP Letting $\psi_t(w) = \frac{1}{2}\|w\|^2$ recovers the result of FedExP with $T_1 = \frac{\|w_0 - w^*\|^2}{\sum_{t=0}^{T-1} \eta_g^t \eta_l \tau}$ in Jhunjhunwala et al. (2023) since the telescoping sum $\sum_{t=1}^{T-1}(D_{\psi_t}(w^* \mid w_t) - D_{\psi_{t-1}}(w^* \mid w_t))$ vanishes as $\psi_t = \psi_{t-1}$. Interestingly, the bias terms T_2 and T_3 introduced by the heterogeneity are independent of the choice of ψ_t. This is because FedDuA adaptively selects η_g^t based on the geometry induced by ψ_t. The choice of ψ_t only affects the initialization error term T_1. As shown in the next section, appropriate choice of ψ_t reduces the numerator and improves the convergence rate compared to FedExP.

Comparison with FedAvg The convergence rate of FedAvg is given by $T_1^{\text{fedavg}} + T_3$, where $T_1^{\text{fedavg}} = O\left(\frac{\|w_0 - w^*\|^2}{T \eta_l \tau}\right)$ (Khaled et al., 2020). As in the comparison with FedExP, T_1 of FedDuA can be smaller than that of FedAvg with appropriate choice of ψ_t. On the other hand, FedDuA yields an additional bias term T_2 due to the adaptive global learning rate. Note that it can be controlled by selecting a client learning rate η_l appropriately (e.g., decreasing η_l over rounds), and it vanishes in the interpolation regime where $\sigma_*^2 = 0$.

4.2.1 Benefit of Adaptivity

As a corollary of Theorem 4.3, we can derive the convergence rate of FedDuAdagrad.

Corollary 4.4. *Assume the same conditions as Theorem 4.3 and $\sup_{t=0}^T \|w_t - w^*\|_\infty \le D$. Then, $\{w^t\}_{t=1}^T$ generated by FedDuAdagrad satisfies*

$$F(\bar{w}_T) - F(w^*) = O\left(\frac{D^2 \operatorname{tr}(G_{T-1})}{\sum_{t=0}^{T-1} \eta_g^t \eta_l \tau}\right) + O(\eta_l \tau \sigma_*^2) + O(\eta_l^2 \tau(\tau-1) L \sigma_*^2)$$

See Appendix E for the proof.

To see the benefit of the adaptivity, let us consider the case where the local update is anisotropic, which is often the case in modern machine learning tasks (Faghri et al., 2020; Tomihari and Sato, 2025). Specifically, we assume $\left|[\bar{\Delta}_t]_k\right| = \Theta(a_t \cdot k^{-\beta})$ for some $a_t > 0, \beta > 1$. That is, the magnitude of the k-th element of the local update decays polynomially. Then, if $\epsilon, \epsilon_g = 0$, the initialization error term T_1 can be evaluated as

$$T_1 = \frac{D^2 \operatorname{tr}(G_{T-1})}{\sum_t \eta_g^t \eta_l \tau} = O\left(\frac{D^2 \sqrt{\sum_{t=0}^{T-1} a_t^2}}{\eta_l \tau \cdot \sum_{t=0}^{T-1} \sqrt{\sum_{s=0}^{t} a_s^2}}\right).$$

As w_t approaches the optimal solution, it is reasonable to assume that the magnitude of the averaged local update a_t decreases. Then, if a_t is monotonically decreasing, we obtain $T_1 = O(D^2/(T\eta_l\tau))$. See Appendix F for the detailed derivation. Thus, the convergence rate of FedDuAdagrad is independent of the dimension d. This is in contrast to the case of FedExP and FedAvg, where the convergence rate $O\left(\frac{\|w_0 - w^*\|^2}{\sum_{t=0}^{T-1} \eta_g^t \eta_l \tau}\right) = O\left(\frac{dD^2}{\sum_{t=0}^{T-1} \eta_g^t \eta_l \tau}\right)$ is linearly dependent on d since $\|w_0 - w_*\|^2 = O(dD^2)$. Thus, if $d \gg 1$, FedDuAdagrad is expected to converge faster than FedExP and FedAvg.

5 NUMERICAL EXPERIMENTS

5.1 Experimental Setup

We evaluate the performance of FedDuA on synthetic and real-world datasets. We consider a distributed overparameterized linear regression problem for synthetic datasets, and image classification and NLP tasks for real-world datasets. We compare FedDuA with

Table 2: Average validation accuracy (%) of the last iterate over 5 different random seeds. Results within 0.5% of the best result for each dataset are bolded.

Fed... dataset	DuAdam (ours)	DuAdagrad (ours)	ExP	ExPM	Adam	Adagrad	AvgM	Avg
CIFAR100	**62.9**	51.2	48.4	59.8	58.4	40.9	56.5	39.5
CIFAR10	**86.6**	82.1	80.7	**86.4**	81.9	74.0	80.6	73.1
FEMNIST	**78.0**	**78.3**	77.5	75.8	77.2	71.6	76.0	76.6
shakespeare	50.9	**52.6**	50.6	48.9	50.6	51.0	48.4	49.1

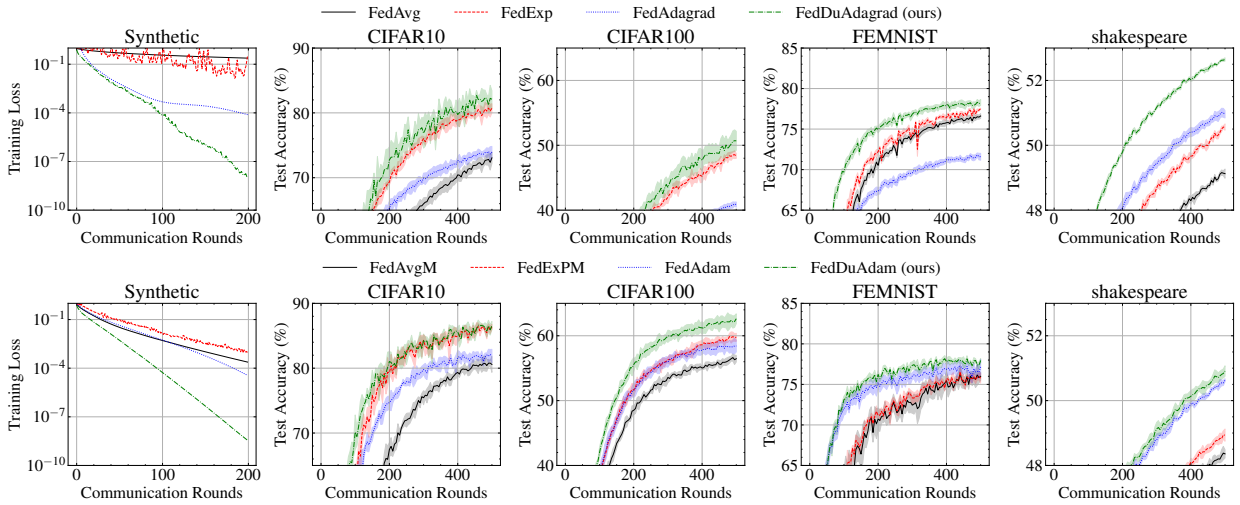

Figure 1: Test accuracy for FedDuA and baselines without server momentum (upper) and with server momentum (lower). Our proposed methods (green dashdot) consistently outperform baselines.

the following baselines: FedAvg, FedExP, FedOpt (FedAdagrad and FedAdam), and their momentum variants (FedAvgM, FedExPM). These algorithms do not require additional computational or communication cost on clients as our proposed method. As discussed in Jhunjhunwala et al. (2023), adaptive learning rate can cause oscillating behavior in the performance. Thus, we adopt averaging the last two iterates strategy as in Jhunjhunwala et al. (2023). For a fair comparison, we perform grid-search over η_g, η_l for FedAvg and FedOpt, and ϵ_g, η_l for FedExP and FedDuA. We fix $\beta_1 = 0.9$ and $\beta_2 = 0.99$ for Fed(Du)Adam following Reddi et al. (2021) and set $\epsilon = 10^{-9}$ for FedOpt and FedDuA if not specified. We fix the number of participating clients at each round to 20, minibatch size for local SGD to 50, and the number of local updates to $\tau = 20$. The results are averaged over 5 random seeds and the shade represents the standard deviation. See Appendix H for the detailed experimental setup.

Synthetic Datasets For the synthetic experiment, we generate $M = 20$ clients with $|\mathcal{D}_i| = 30$ samples. The synthetic datasets are generated following a similar procedure as in Jhunjhunwala et al. (2023) and Li et al. (2020), but we consider anisotropic data distribution, which corresponds to the fact that data often lies on a low-dimensional structure (Ansuini et al., 2019). Specifically, we generate the input data $x \in \mathbb{R}^d$ ($d = 1000$) as $x \sim \mathcal{N}(0, \Sigma)$, where Σ is a diagonal matrix with its k-th diagonal element $\Sigma_{kk} = k^{-\beta}$ for $\beta = 1.1$.

Real-world Datasets For CIFAR10/100, we partition the data into $M = 100$ clients by following a Dirichlet distribution with parameter $\alpha = 0.3$ and use ResNet-18. FEMNIST is naturally partitioned into $3,550$ clients based on the writer of the digit or character (Caldas et al., 2018). We subsample 100 clients for train and test to reduce the computational cost, and use the same CNN architecture as in Zhu et al. (2022). Shakespeare dataset is also naturally partitioned into $1,129$ clients based on the speaking role (Caldas et al., 2018) and we subsample 100 clients and use a LSTM for next character prediction.

5.2 Results and Discussion

FedDuA consistently outperforms baselines Fig. 1 and Table 2 show the performance of FedDuA and baselines on synthetic and real-world datasets. Overall, FedDuA consistently outperforms existing

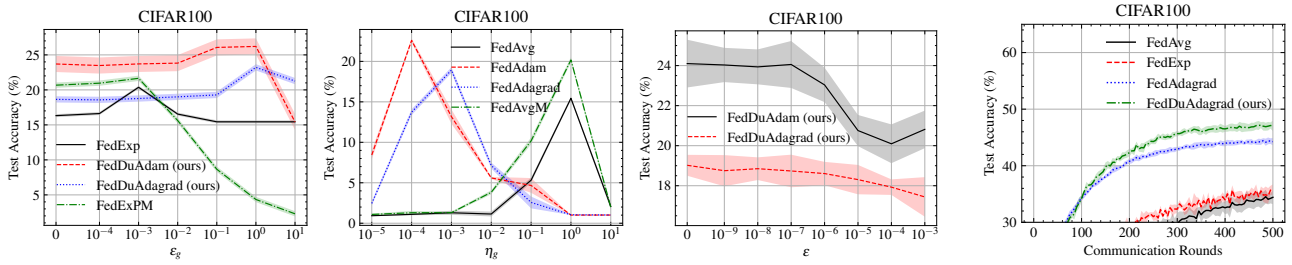

Figure 2: Test accuracy averaged over the last 5 iterates with different hyperparameters ϵ_g, η_g and ϵ. FedDuA is less sensitive to the choice of hyperparameters.

Figure 3: Training of ViT on CIFAR-100.

methods across all datasets. These results clearly show that dual adaptivity is a key to achieve fast convergence in FL. In CIFAR10/100, momentum-based methods perform better than non-momentum methods while non-momentum methods perform better or comparable in the other datasets. See Appendix I.1 for long-term behavior of the algorithms.

FedDuA is Robust to the choice of hyperparameters Hyperparameter-tuning is often time-consuming and expensive in FL. Adaptive methods are expected to be robust to the choice of hyperparameters. To see this, we run each algorithm with different choice of ϵ_g, η_g for 50 rounds. We tune η_l for each hyperparameter setting. As shown in Fig. 2, the performance of FedDuA is not sensitive to the choice of ϵ_g. Furthermore, $\epsilon_g = 0$ is sufficient to achieve good performance. This indicates that FedDuA is robust to the choice of hyperparameters and even hyperparameter-free by setting $\eta_g^t := m_t / \|v_t\|^2_{G_t^{-1}}$. On the other hand, FedOpt does not perform well with not well-tuned hyperparameters, which requires careful tuning of η_g. We also conduct an ablation study on the choice of ϵ for FedDuA. Note that ϵ determines the degree of adaptivity on gradient heterogeneity. We see that FedDuA is also robust to the choice of ϵ but increasing ϵ degrades the performance gradually since it reduces the gradient adaptivity. We find that $\epsilon = 0$ works well but we use $\epsilon = 10^{-9}$ in other experiments for the sake of numerical stability.

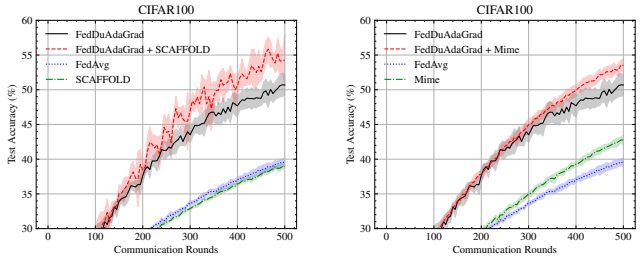

Figure 4: Comparison with SCAFFOLD and Mime.

FedDuA is Complementary to Mime and SCAFFOLD Our approach is orthogonal to existing methods that modify the local update procedure, such as SCAFFOLD (Karimireddy et al., 2020b) and Mime (Karimireddy et al., 2020a). Therefore, FedDuA can be naturally combined with these methods to further improve performance. As shown in Fig. 4, integrating FedDuA with SCAFFOLD or Mime-type local updates yields better results than vanilla FedDuA, whereas vanilla SCAFFOLD and Mime alone fail to surpass FedDuA. For Mime, we adopt SGDm as the base optimizer and set the momentum parameter to 0.9, following the original paper. These results demonstrate that our method is complementary to existing approaches for mitigating client heterogeneity.

Coordinate-wise adaptivity is essential in training of Transformers As discussed in previous works (Zhang et al., 2024; Tomihari and Sato, 2025), adaptive methods such as Adam outperform SGD in centralized training of Transformers. To see the effect of adaptivity in FL, we train a Vision Transformer (ViT) (Dosovitskiy et al., 2021) on CIFAR-100. Fig. 3 shows that FedExP does not work well in training of ViT while it performs comparably in training of ResNet. This is in contrast to FedDuAdagrad, which performs well in both cases. The result implies that coordinate-wise adaptivity is also effective in FL training of Transformers.

6 CONCLUSION

In this paper, we addressed the question of how to design a global update procedure for FL, which is adaptive to both client and gradient heterogeneity. For this purpose, we proposed FedDuA, a doubly adaptive step-size rule for general mirror descent-type algorithms, by formulating the global update procedure through the lens of mirror descent. We proved that our proposed step-size is minimax optimal under approximate projection condition and provided the convergence analysis. Extensive numerical experiments show

that FedDuA outperforms existing adaptive methods in various settings without requiring additional computational and communication cost on clients. Furthermore, FedDuA is robust to the choice of hyperparameters and can be combined with existing methods such as Mime to further improve the performance.

References

Amid, E., Ganesh, A., Mathews, R., Ramaswamy, S., Song, S., Steinke, T., Suriyakumar, V. M., Thakkar, O., and Thakurta, A. (2022). Public data-assisted mirror descent for private model training. In *International Conference on Machine Learning*, pages 517–535. PMLR.

Ansuini, A., Laio, A., Macke, J. H., and Zoccolan, D. (2019). Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečnỳ, J., McMahan, H. B., Smith, V., and Talwalkar, A. (2018). Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*.

Cheng, Z., Huang, X., Wu, P., and Yuan, K. (2024). Momentum benefits non-iid federated learning simply and provably. In *The Twelfth International Conference on Learning Representations*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Faghri, F., Duvenaud, D., Fleet, D. J., and Ba, J. (2020). A study of gradient variance in deep learning. *arXiv preprint arXiv:2007.04532*.

Haddadpour, F. and Mahdavi, M. (2019). On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*.

Hazan, E. et al. (2016). Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

Jhunjhunwala, D., Wang, S., and Joshi, G. (2023). FedExP: Speeding Up Federated Averaging via Extrapolation. In *International Conference on Learning Representations*.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021). Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210.

Karimireddy, S. P., Jaggi, M., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. (2020a). Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020b). Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pages 5132–5143.

Karimireddy, S. P., Rebjock, Q., Stich, S., and Jaggi, M. (2019). Error feedback fixes signsgd and other gradient compression schemes. In *International Conference on Machine Learning*, pages 3252–3261.

Khaled, A., Mishchenko, K., and Richtárik, P. (2020). Tighter theory for local sgd on identical and heterogeneous data. In *International conference on artificial intelligence and statistics*, pages 4519–4529. PMLR.

Kingma, D. P. and Ba, J. (2015). Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations*.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Lee, S. H., Sharma, S., Zaheer, M., and Li, T. (2024). Efficient Adaptive Federated Optimization. *arXiv preprint arXiv:2410.18117*.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. (2020). Federated optimization in heterogeneous networks. In *Proceedings of Machine learning and systems*, volume 2, pages 429–450.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, pages 1273–1282.

Mishchenko, K., Malinovsky, G., Stich, S., and Richtárik, P. (2022). Proxskip: Yes! local gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine Learning*, pages 15750–15769. PMLR.

Nielsen, F., Boissonnat, J.-D., and Nock, R. (2007). Bregman voronoi diagrams: Properties, algorithms and applications. *arXiv preprint arXiv:0709.2196*.

Odeyomi, O. T. and Zaruba, G. (2021). Privacy-preserving online mirror descent for federated learning with single-sided trust. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7. IEEE.

omiita (2024). ViT-CIFAR. `https://github.com/omihub777/ViT-CIFAR`. Commit hash: ab9043e.

Qu, Z., Li, X., Duan, R., Liu, Y., Tang, B., and Lu, Z. (2022). Generalized federated learning via sharpness aware minimization. In *International conference on machine learning*, pages 18250–18280. PMLR.

Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan, H. B. (2021). Adaptive Federated Optimization. In *International Conference on Learning Representations*.

Tomar, M., Shani, L., Efroni, Y., and Ghavamzadeh, M. (2022). Mirror Descent Policy Optimization. In *International Conference on Learning Representations*.

Tomihari, A. and Sato, I. (2025). Understanding Why Adam Outperforms SGD: Gradient Heterogeneity in Transformers. *arXiv preprint arXiv:2502.00213*.

Wang, J., Xu, Z., Garrett, Z., Charles, Z., Liu, L., and Joshi, G. (2021). Local adaptivity in federated learning: Convergence and consistency. *arXiv preprint arXiv:2106.02305*.

Xie, C., Koyejo, O., Gupta, I., and Lin, H. (2019). Local adaalter: Communication-efficient stochastic gradient descent with adaptive learning rates. *arXiv preprint arXiv:1911.09030*.

Yuan, H., Zaheer, M., and Reddi, S. (2021). Federated composite optimization. In *International Conference on Machine Learning*, pages 12253–12266. PMLR.

Zaccone, R., Karimireddy, S. P., Masone, C., and Ciccone, M. (2025). Communication-efficient heterogeneous federated learning with generalized heavy-ball momentum. *Transactions on Machine Learning Research*.

Zaheer, M., Reddi, S., Sachan, D., Kale, S., and Kumar, S. (2018). Adaptive methods for nonconvex optimization. *Advances in neural information processing systems*, 31.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. (2020). Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393.

Zhang, Y., Chen, C., Ding, T., Li, Z., Sun, R., and Luo, Z. (2024). Why transformers need adam: A hessian perspective. *Advances in Neural Information Processing Systems*, 37:131786–131823.

Zhu, C., Xu, Z., Chen, M., Konečný, J., Hard, A., and Goldstein, T. (2022). Diurnal or Nocturnal? Federated Learning of Multi-branch Networks from Periodically Shifting Distributions. In *International Conference on Learning Representations*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] We provide the convergence analysis in Section 4.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] The code is provided in the supplementary material.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes] The proofs are provided in the supplementary material.
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] See Appendix H.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Yes]

(b) The license information of the assets, if applicable. [Yes]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Not Applicable]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A Proof for Theorem 3.2

The approximate projection condition yields

$$\langle \bar{\Delta}_t, w^* - w_t \rangle \geq \frac{1}{2M} \sum_{i=1}^{M} \|\Delta_i^t\|^2 \geq 0.$$

For $\theta_{t+1} := \theta_t + \eta_g \bar{\Delta}_t$, we have

$$\begin{aligned} D_{\phi_t}(\theta_{t+1} \mid \theta^*) - D_{\phi_t}(\theta_t \mid \theta^*) &= \phi_t(\theta_{t+1}) - \phi_t(\theta_t) - \eta_g \langle \nabla \phi_t(\theta^*), \bar{\Delta}_t \rangle \\ &= \phi_t(\theta_t + \eta_g \bar{\Delta}_t) - \phi_t(\theta_t) - \eta_g \langle w^*, \bar{\Delta}_t \rangle. \end{aligned}$$

For the last equality, we used $\nabla \phi_t(\theta^*) = w^*$ from the duality. From the first-order optimality condition, we have

$$h_t(\eta_g) = \langle w^* - w_t, \bar{\Delta}_t \rangle,$$

where $h_t(\eta) = \frac{\mathrm{d}}{\mathrm{d}\eta} \phi_t(\theta_t + \eta \bar{\Delta}_t) - \langle w_t, \bar{\Delta}_t \rangle$. Since ϕ_t is assumed to be strongly convex, there exists a constant $\alpha_t > 0$ such that $\nabla^2 \phi_t(\theta) \succ \alpha_t I_d$, and we have

$$\frac{\mathrm{d}h_t}{\mathrm{d}\eta}(\eta) = \langle \nabla^2 \phi_t(\theta_t + \eta \bar{\Delta}_t) \bar{\Delta}_t, \bar{\Delta}_t \rangle > \alpha_t \left\| \bar{\Delta}_t \right\|^2.$$

This implies that h_t is strictly increasing and the inverse function h_t^{-1} exists on $\mathbb{R}$. Thus, η_g^* is uniquely determined by

$$\eta_g^* = h_t^{-1} \left(\langle w^* - w_t, \bar{\Delta}_t \rangle \right).$$

Using the monotonicity of h_t^{-1}, we have

$$\begin{aligned} \eta_g^* &= h_t^{-1}(\langle w^* - w_t, \bar{\Delta}_t \rangle) \\ &\geq h_t^{-1} \left(\frac{1}{2M} \sum_{i=1}^{M} \|\Delta_i^t\|^2 \right), \end{aligned}$$

which completes the proof.

B Proof for Theorem 3.3

The strong approximate projection condition yields

$$\langle \bar{\Delta}_t, w^* - w_t \rangle \geq \frac{1}{M} \sum_{i=1}^{M} \|\Delta_i^t\|^2 \geq 0.$$

For $\theta_{t+1} := \theta_t + \eta_g v_t$, we have

$$\begin{aligned} D_{\phi_t}(\theta_{t+1} \mid \theta^*) - D_{\phi_t}(\theta_t \mid \theta^*) &= \phi_t(\theta_{t+1}) - \phi_t(\theta_t) - \eta_g \langle \nabla \phi_t(\theta^*), v_t \rangle \\ &= \phi_t(\theta_t + \eta_g v_t) - \phi_t(\theta_t) - \eta_g \langle w^*, v_t \rangle. \end{aligned}$$

For the last equality, we used $\nabla\phi_t(\theta^*) = w^*$ from the duality. From the first-order optimality condition, we have

$$h_t(\eta_g) = \langle w^* - w_t, v_t\rangle.$$

and $\eta_g^* = h_t^{-1}(\langle w^* - w_t, v_t\rangle)$ as in the proof of Theorem 3.2. From the monotonicity of h_t^{-1}, it suffices to show

$$\langle w^* - w_t, v_t\rangle \geq 2m_t \geq m_t.$$

To prove this, we use induction on t.

Case $t = 0$ We have

$$\begin{aligned}\langle w^* - w_0, v_0\rangle &= (1-\beta_1)\langle w^* - w_0, \bar{\Delta}_0\rangle \\ &\geq (1-\beta_1)\frac{1}{M}\sum_{i=1}^{M}\left\|\Delta_i^0\right\|^2 = 2m_0.\end{aligned}$$

The inequality follows from strong A.P.C. Thus, we obtain the desired result.

Case $t \geq 1$ Assume that the inequality holds for $t-1$. Then, we have

$$\begin{aligned}\langle w^* - w_t, v_t\rangle &= (1-\beta_1)\langle w^* - w_t, \bar{\Delta}_t\rangle + \beta_1\langle w^* - w_t, v_{t-1}\rangle \\ &= (1-\beta_1)\langle w^* - w_t, \bar{\Delta}_t\rangle + \beta_1(\langle w^* - w_{t-1}, v_{t-1}\rangle - \underbrace{\langle w_t - w_{t-1}, v_{t-1}\rangle}_{:=R}).\end{aligned}$$

The term R can be evaluated as

$$\begin{aligned}R &= \langle\nabla\phi_{t-1}(\theta_{t-1} + \eta_g^{t-1}v_{t-1}), v_{t-1}\rangle - \langle w_{t-1}, v_{t-1}\rangle \\ &= h_{t-1}(\eta_g^{t-1}) \leq m_{t-1}\end{aligned}$$

since η_g^{t-1} is assumed to be smaller than $h_{t-1}^{-1}(m_{t-1})$. Substituting the above equality, we obtain

$$\begin{aligned}\langle w^* - w_t, v_t\rangle &= (1-\beta_1)\langle w^* - w_t, \bar{\Delta}_t\rangle + \beta_1(\langle w^* - w_{t-1}, v_{t-1}\rangle - m_{t-1}) \\ &\geq (1-\beta_1)\langle w^* - w_t, \bar{\Delta}_t\rangle + \beta_1(2m_{t-1} - m_{t-1}) \\ &\geq (1-\beta_1)\frac{1}{M}\sum_{i=1}^{M}\left\|\Delta_i^t\right\|^2 + \beta_1 m_{t-1} \\ &= 2m_t.\end{aligned}$$

Here, we used the induction hypothesis for the first inequality and strong A.P.C. for the second inequality. Then, we obtain the result by induction.

C Proof for Theorem 4.1

Given $w \in \mathbb{R}^d$, the worst-case distance difference is defined as

$$\begin{aligned}V(w) &= \sup_{w^*\in H} V(w, w^*) \\ &:= \sup_{w^*\in H} D_{\psi_t}(w^* \mid w) - D_{\psi_t}(w^* \mid w_t).\end{aligned}$$

From the definition of the Bregman divergence, we have

$$\begin{aligned}V(w, w^*) &= D_{\psi_t}(w^* \mid w) - D_{\psi_t}(w^* \mid w_t) \\ &= -\psi_t(w) - \nabla\psi_t(w)^\top(w^* - w) + \psi_t(w_t) + \nabla\psi_t(w_t)^\top(w^* - w_t) \\ &= \psi_t(w_t) - \psi_t(w) + (\nabla\psi_t(w_t) - \nabla\psi_t(w))^\top(w^* - w).\end{aligned}$$

Thus, $V(w, w^*)$ is affine in w^*.

Let us consider the Lagrangian

$$\mathcal{L}(w, w^*, \lambda) = V(w, w^*) - \lambda\left(\langle \bar{\Delta}^t, w^t - w^* \rangle + \frac{1}{2M}\sum_{i=1}^{M}\left\|\Delta_i^t\right\|^2\right)$$

Since $V(w, w^*)$ and the constraint are affine, strong duality holds and we have

$$\begin{aligned} V(w) &= \sup_{w^* \in \mathbb{R}^d} \inf_{\lambda \geq 0} \mathcal{L}(w, w^*, \lambda) \\ &= \inf_{\lambda \geq 0} \sup_{w^* \in \mathbb{R}^d} \mathcal{L}(w, w^*, \lambda), \end{aligned}$$

and

$$\begin{aligned} \inf_w V(w) &= \inf_w \inf_{\lambda \geq 0} \sup_{w^* \in \mathbb{R}^d} \mathcal{L}(w, w^*, \lambda) \\ &= \inf_{\lambda \geq 0} \inf_w \sup_{w^* \in \mathbb{R}^d} \mathcal{L}(w, w^*, \lambda). \end{aligned}$$

Since the Lagrangian is affine, $\sup_{w^* \in \mathbb{R}^d} \mathcal{L}(w, w^*, \lambda)$ is infinite unless $\theta_t - \theta + \lambda\bar{\Delta}_t = 0$ and if this condition holds, the value of $\sup_{w^* \in \mathbb{R}^d} \mathcal{L}(w, w^*, \lambda)$ is independent of w^*. Thus, we have

$$\sup_{w^* \in \mathbb{R}^d} \mathcal{L}(w, w^*, \lambda) = \begin{cases} V(w, w_t) + \frac{\lambda}{2M}\sum_{i=1}^{M}\|\Delta_i^t\|^2 & \text{if } \theta_t - \theta + \lambda\bar{\Delta}_t = 0, \\ \infty & \text{otherwise.} \end{cases}$$

Therefore, for a given $\lambda \geq 0$, $\sup_{w^*} \mathcal{L}(w, w^*, \lambda)$ is minimized at $w = w^\lambda := \nabla\phi_t(\theta_t + \lambda\bar{\Delta}_t)$ Therefore, we have

$$\begin{aligned} \inf_w \sup_{w^* \in \mathbb{R}^d} \mathcal{L}(w, w^*, \lambda) &= V(w^\lambda, w_t) + \frac{\lambda}{2M}\sum_{i=1}^{M}\left\|\Delta_i^t\right\|^2 \\ &= D_{\psi_t}(w_t \mid w^\lambda) - \frac{\lambda}{2M}\sum_{i=1}^{M}\left\|\Delta_i^t\right\|^2 \\ &= D_{\phi_t}(\theta_t + \lambda\bar{\Delta}_t \mid \theta_t) - \frac{\lambda}{2M}\sum_{i=1}^{M}\left\|\Delta_i^t\right\|^2 \\ &= \phi_t(\theta_t + \lambda\bar{\Delta}_t) - \phi_t(\theta_t) - \langle\nabla\phi_t(\theta_t), \lambda\bar{\Delta}_t\rangle - \frac{\lambda}{2M}\sum_{i=1}^{M}\left\|\Delta_i^t\right\|^2. \end{aligned}$$

Considering the first-order optimality condition on λ, we have

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}\lambda}\phi_t(\theta_t + \lambda\bar{\Delta}_t) - \langle\nabla\phi_t(\theta_t), \bar{\Delta}_t\rangle - \frac{1}{2M}\sum_{i=1}^{M}\left\|\Delta_i^t\right\|^2 \qquad (4) \\ &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\phi_t(\theta_t + \lambda\bar{\Delta}_t) - \langle w_t, \bar{\Delta}_t\rangle - \frac{1}{2M}\sum_{i=1}^{M}\left\|\Delta_i^t\right\|^2 \\ &= h_t(\lambda) - \frac{1}{2M}\sum_{i=1}^{M}\left\|\Delta_i^t\right\|^2, \\ &= 0, \end{aligned}$$

which implies the optimal $\lambda_* = h_t^{-1}(\frac{1}{2M}\sum_{i=1}^{M}\|\Delta_i^t\|^2) = \eta_g^t$ uniquely exists as in the proof of Theorem 3.2. Note that $\eta_g^t \geq 0$ since expression (4) is increasing in λ and negative at $\lambda = 0$. Thus, $V(w)$ is minimized at $w = \nabla\phi_t(\theta_t + \eta_g^t\bar{\Delta}_t)$. This completes the proof.

D Proof for Theorem 4.3

Let $D_t = D_{\psi_t}$ for simplicity. We have

$$\begin{aligned} D_t(w^* \mid w_{t+1}) - D_t(w^* \mid w_t) &= D_{\phi_t}(\theta_{t+1} \mid \theta^*) - D_{\phi_t}(\theta_t \mid \theta^*) \\ &= \phi_t(\theta_{t+1}) - \phi_t(\theta_t) - \langle \nabla \phi_t(\theta^*), \theta_{t+1} - \theta_t \rangle \\ &= \phi_t(\theta_{t+1}) - \phi_t(\theta_t) - \eta_g^t \langle w^*, \bar{\Delta}_t \rangle \\ &= H_t(\eta_g^t) + \eta_g^t \langle w_t - w^*, \bar{\Delta}_t \rangle, \end{aligned}$$

where $H_t(\eta) = \phi_t(\theta_t + \eta \bar{\Delta}_t) - \phi_t(\theta_t) - \eta \langle w_t, \bar{\Delta}_t \rangle$. For the second equality, we used the duality $w^* = \nabla \phi_t(\theta^*)$. From the mean-value theorem, there exists $\bar{\eta} \in [0, \eta_g^t]$ such that

$$\frac{H_t(\eta_g^t) - H_t(0)}{\eta_g^t - 0} = \frac{H_t(\eta_g^t)}{\eta_g^t} = \frac{\mathrm{d}H_t}{\mathrm{d}\eta}(\bar{\eta}) = h_t(\bar{\eta}) \leq h_t(\eta_g^t).$$

The last inequality follows from the fact that h_t is increasing and $\bar{\eta} \leq \eta_g^t$. From the definition of η_g^t, we have

$$\frac{H_t(\eta_g^t)}{\eta_g^t} \leq h_t(\bar{\eta}_g^t) = \frac{1}{2M} \sum_{i=1}^M \left\| \Delta_i^t \right\|^2.$$

Substituting the above inequality, we have

$$D_t(w^* \mid w_{t+1}) - D_t(w^* \mid w_t) \leq \frac{\eta_g^t}{2} \cdot \underbrace{\frac{1}{M} \sum_{i=1}^M \left\| \Delta_i^t \right\|^2}_{:= R_2} + \eta_g^t \underbrace{\langle w_t - w^*, \bar{\Delta}_t \rangle}_{:= R_1}.$$

In a similar way as in the proof of Theorem 1 in Jhunjhunwala et al. (2023), R_2 can be bounded as

$$\begin{aligned} R_2 &= \frac{1}{M} \sum_{i=1}^M \left\| \Delta_i^t \right\|^2 \\ &= \frac{\eta_l^2}{M} \sum_{i=1}^M \left\| \sum_{k=0}^{\tau-1} \nabla F_i(w_{i,k}^t) \right\|^2 \\ &\leq \frac{\tau \eta_l^2}{M} \sum_{i=1}^M \sum_{k=0}^{\tau-1} \left\| \nabla F_i(w_{i,k}^t) \right\|^2 \\ &\leq \frac{3\tau \eta_l^2 L^2}{M} \sum_{i=1}^M \sum_{k=0}^{\tau-1} \left\| w_{i,k}^t - w_t \right\|^2 + 6\tau^2 \eta_l^2 L (F(w^t) - F(w^*)) + 3\tau^2 \eta_l^2 \sigma_*^2. \end{aligned}$$

Here, we used Lemma 5 in Jhunjhunwala et al. (2023) for the last inequality. For R_1, we have

$$\begin{aligned} R_1 &= \frac{1}{M} \sum_{i=1}^M \langle w_t - w^*, \Delta_i^t \rangle \\ &= \frac{\eta_l}{M} \sum_{i=1}^M \sum_{k=0}^{\tau-1} \langle w_t - w^*, \nabla F_i(w_{i,k}^t) \rangle. \end{aligned}$$

As shown in the proof of Theorem 1 in Jhunjhunwala et al. (2023), the right-hand side can be bounded as

$$R_1 \geq \eta_l \tau (F(w_t) - F(w^*)) - \frac{\eta_l L}{2M} \sum_{i=1}^M \sum_{k=0}^{\tau-1} \left\| w_{i,k}^t - w_t \right\|^2.$$

Combining the above two inequalities, we have

$$\begin{aligned} D_t(w^* \mid w_{t+1}) - D_t(w^* \mid w_t) &\le 2\eta_g^t\eta_l\tau(1-3\eta_l\tau L)(F(w_t)-F(w^*)) + 3\eta_g^t\eta_l^2\tau^2\sigma_*^2 \\ &\quad + (3\eta_g^t\eta_l^2\tau + \eta_g^t\eta_l L)\frac{1}{M}\sum_{i=1}^{M}\sum_{k=0}^{\tau-1}\left\|w_{i,k}^t - w_t\right\|^2 \\ &\le -\frac{\eta_g^t\eta_l\tau}{3}(F(w_t)-F^*) + \eta_g^t\cdot O(\eta_l^3\tau^2\sigma_*^2 + \eta_l^2\tau^2(\tau-1)L\sigma_*^2). \end{aligned}$$

Averaging over all rounds, we have

$$\begin{aligned} \frac{\sum_{t=0}^{T-1}\eta_g^t(F(w_t)-F(w^*))}{\sum_{t=0}^{T-1}\eta_g^t} &\le 3\cdot\frac{\sum_{t=0}^{T-1}D_t(w^*\mid w_t) - D_t(w^*\mid w_{t+1})}{\sum_t \eta_g^t\eta_l\tau} + O(\eta_l\tau\sigma_*^2 + \eta_l^2\tau(\tau-1)L\sigma_*^2) \\ &\le O\left(\frac{D_0(w^*\mid w_0) + \sum_{t=1}^{T-1}(D_t(w^*\mid w_t) - D_{t-1}(w^*\mid w_t))}{\sum_t \eta_g^t\eta_l\tau}\right) \\ &\quad + O(\eta_l\tau\sigma_*^2 + \eta_l^2\tau(\tau-1)L\sigma_*^2). \end{aligned}$$

Since F is convex, we have

$$F(\bar{w}_T) - F(w^*) \le \frac{\sum_{t=0}^{T-1}\eta_g^t(F(w_t)-F(w^*))}{\sum_{t=0}^{T-1}\eta_g^t}.$$

This completes the proof.

E Proof for Corollary 4.4

The numerator in the first term can be bounded as

$$\begin{aligned} \sum_{t=0}^{T-1} D_t(w^*\mid w_t) - D_t(w^*\mid w_{t+1}) &\le \sum_{t=0}^{T-1}(D_t(w^*\mid w_t) - D_{t-1}(w^*\mid w_t)) \\ &= \sum_{t=0}^{T-1}\frac{1}{2}(w_t-w^*)^\top(G_t-G_{t-1})(w_t-w^*) \\ &\le \sum_{t=0}^{T-1}\frac{1}{2}\|w_t-w^*\|_\infty^2\|g_t-g_{t-1}\|_1 \\ &= \sum_{t=0}^{T-1}\frac{D^2}{2}(\|g_t\|_1 - \|g_{t-1}\|_1) \\ &\le \frac{D^2}{2}\|g_{T-1}\|_1 \\ &= \frac{D^2}{2}\operatorname{tr}(G_{T-1}) \end{aligned}$$

where we define $D_{-1}(w_0 \mid w^*) = 0$ and $g_t = \operatorname{diag}(G_t)$. The second equality follows from the monotonicity of $[g_t]_k$. Substituting the above inequality, we obtain

$$\begin{aligned} F\left(\frac{\sum_{t=0}^{T-1}\eta_g^t w_t}{\sum_{t=0}^{T-1}\eta_g^t}\right) - F(w^*) &\le \frac{\sum_{t=0}^{T-1}\eta_g^t(F(w_t)-F(w^*))}{\sum_{t=0}^{T-1}\eta_g^t} \\ &= O\left(\frac{D^2\operatorname{tr}(G_{T-1})}{\sum_{t=0}^{T-1}\eta_g^t\eta_l\tau}\right) + O(\eta_l\tau\sigma_*^2) + O(\eta_l^2\tau(\tau-1)L\sigma_*^2) \end{aligned}$$

This completes the proof.

F Detailed Analysis on Benefit of Adaptivity

From the assumption, we have

$$\begin{aligned}[G_t]_{k,k} &= \sqrt{\sum_{t=0}^{t}[\bar{\Delta}_t]_k^2} \\ &= \Theta\left(k^{-\beta}\sqrt{\sum_{s=0}^{t} a_s^2}\right).\end{aligned}$$

Thus, $\mathrm{tr}(G_{T-1}) = \Theta(\sqrt{\sum_{t=0}^{T-1} a_t^2})$ since $\beta > 1$. On the other hand, the numerator can be bounded as

$$\begin{aligned}\sum_t \eta_g^{(t)} &= \sum_{t=0}^{T-1} \frac{\frac{1}{M}\sum_{i=1}^{M}\left\|\Delta_i^t\right\|^2}{\left\|\bar{\Delta}^t\right\|_{G_t^{-1}}^2} \\ &\geq \sum_{t=0}^{T-1} \frac{\left\|\bar{\Delta}^t\right\|^2}{\left\|\bar{\Delta}^t\right\|_{G_t^{-1}}^2} \\ &= \sum_{t=0}^{T-1} \frac{\Theta\left(\sum_{k=1}^{d} a_t^2 k^{-2\beta}\right)}{\Theta\left(\frac{a_t^2}{\sqrt{\sum_{s=0}^{t} a_s^2}}\cdot\sum_{k=1}^{d} k^{-\beta}\right)} \\ &= \Theta\left(\sum_{t=0}^{T-1}\sqrt{\sum_{s=0}^{t} a_s^2}\right),\end{aligned}$$

since $[G_t]_{k,k} = \sqrt{\sum_{s=0}^{t}[\bar{\Delta}_s]_k^2} = \Theta(\sqrt{\sum_{s=0}^{t} a_s^2 k^{-2\beta}})$. Substituting the above two inequalities, we have

$$T_1 = O\left(\frac{D^2\sqrt{\sum_{t=0}^{T-1} a_t^2}}{\eta_l\tau\sum_{t=0}^{T-1}\sqrt{\sum_{s=0}^{t} a_s^2}}\right).$$

If a_t is decreasing, we have

$$\begin{aligned}\frac{D^2}{\eta_l\tau\sum_{t=0}^{T-1}\frac{\sqrt{\sum_{s=0}^{t} a_s^2}}{\sqrt{\sum_{s=0}^{T-1} a_s^2}}} &\leq \frac{D^2}{\eta_l\tau\sum_{t=0}^{T-1}\sqrt{\frac{t}{T}}} \\ &= O\left(\frac{D^2}{\eta_l\tau T}\right).\end{aligned}$$

For the first inequality, we used $\frac{\sum_{s=0}^{t} a_s^2}{\sum_{s=0}^{T-1} a_s^2} \geq t/T$ since we assume $a_s \leq a_{s-1}$. That is, we have

$$\begin{aligned}\frac{\sum_{s=0}^{T-1} a_s^2}{\sum_{s=0}^{t} a_s^2} &= 1 + \frac{\sum_{s=t+1}^{T-1} a_s^2}{\sum_{s=0}^{t} a_s^2} \\ &\leq 1 + \frac{(T-t-1)\cdot a_t^2}{(t+1)\cdot a_t^2} \\ &= \frac{T}{t+1}\end{aligned}$$

and thus $\frac{\sum_{s=0}^{t} a_s^2}{\sum_{s=0}^{T-1} a_s^2} \geq \frac{t}{T}$.

G Convergence Analysis for Non-convex Objectives

In this section, we provide a convergence analysis for non-convex objectives. For that purpose, we introduce the following assumptions.

Assumption G.1. Local loss function F_i is L-smooth i.e., for any $w, w' \in \mathbb{R}^d$,

$$F(w') \leq F(w) + \langle \nabla F(w), w' - w \rangle + \frac{L}{2}\|w' - w\|^2.$$

Also, there exists a constant $\sigma_g \geq 0$ such that for any $w \in \mathbb{R}^d$,

$$\frac{1}{M}\sum_{i=1}^{M}\|\nabla F_i(w) - \nabla F(w)\|^2 \leq \sigma_g^2.$$

Assumption G.2. For any $0 \leq t \leq T-1$, there exist constant $0 \leq \alpha_t \leq \beta_t$ such that for any $w, w' \in \mathbb{R}^d$,

$$\frac{1}{2}\alpha_t\|w - w'\|^2 \leq D_{\psi_t}(w \mid w') \leq \frac{1}{2}\beta_t\|w - w'\|^2$$

and $\beta_t/\alpha_t \leq \kappa$ for some constant $\kappa \geq 1$ independent of t.

The first assumption is standard in the analysis of non-convex optimization (Jhunjhunwala et al., 2023; Reddi et al., 2021) and the second assumption is satisfied by FedDuAdagrad case as shown later. Under the above assumptions, we have the following convergence result.

Theorem G.3. *Suppose that Assumptions G.1 and G.2 hold and clients use full-batch SGD and participate in every round. Then, if $\eta_l \leq \frac{1}{6\tau L\kappa}$ generated by FedDuA satisfies*

$$\min_{t=0,\ldots,T-1}\left\|\nabla F(w^t)\right\|^2 \leq O\left(\frac{F(w_0) - F^*}{\eta_l\tau\sum_{t=0}^{T-1}\tilde{\eta}_g^t}\right) + O(\kappa\eta_l^2\tau^2L^2\sigma_g^2) + O(\kappa\eta_l\tau L\sigma_g^2).$$

where $\tilde{\eta}_g^t = \eta_g^t/\beta_t$.

See Appendix G.1 for the proof. As a corollary, we have the following convergence result for FedDuAdagrad.

Corollary G.4. *Suppose that Assumptions G.1 and G.2 hold and clients use full-batch SGD and participate in every round. Additionally, assume that $\|\nabla F_i(w)\| \leq G$ for any $w \in \mathbb{R}^d$. Then, if $\eta_l \leq \frac{1}{6\tau L}$ generated by FedDuAdagrad satisfies*

$$\min_{t=0,\ldots,T-1}\left\|\nabla F(w^t)\right\|^2 \leq O\left(\frac{F(w_0) - F^*}{\eta_l\tau\sum_{i=0}^{T-1}\left(\frac{1}{M}\sum_{i=1}^{M}\|\Delta_i^t\|/\left\|\bar{\Delta}_t\right\|^2\right)}\right) + O(\eta_l^2\tau^2L^2\sigma_g^2) + O(\eta_l\tau L\sigma_g^2).$$

where $\epsilon = (\eta_l\tau G\sqrt{T})/L$.

See Appendix G.2 for the proof. This matches the convergence rate of FedExP in Jhunjhunwala et al. (2023). In addition, the convergence rate of FedAvg is $O\left(\frac{F(w_0)-F^*}{\eta_l\tau T}\right) + O(\eta_l^2\tau^2L^2\sigma_g^2)$ (Reddi et al., 2021). Since $\sum_{i=0}^{T-1}\frac{1}{M}\sum_{i=1}^{M}\|\Delta_i^t\|/\left\|\bar{\Delta}_t\right\|^2 \geq T$, we can expect the first term in the convergence rate of FedDuAdagrad is smaller than that of FedAvg. On the other hand, FedDuAdagrad has an additional term $O(\eta_l\tau L\sigma_g^2)$ at a price of adaptivity. Note that it can be controlled by selecting a client learning rate η_l appropriately.

G.1 Proof for Theorem G.3

Proof. From the Lipschitz smoothness of F, we have

$$\begin{aligned}
F(w_{t+1}) - F(w_t) &\leq \langle \nabla F(w^t), w^{t+1} - w^t \rangle + \frac{L}{2}\left\|w^{t+1} - w^t\right\|^2 \\
&\leq \langle \nabla F(w^t), w^{t+1} - w^t \rangle + \frac{L}{\alpha_t}D_{\psi_t}(w^t \mid w^{t+1}).
\end{aligned}$$

For the first term, from the mean-value theorem, there exists $\bar{\eta} \in [0, \eta_g^t]$ such that

$$\begin{aligned} w^{t+1} - w^t &= \nabla\phi_t(\theta_t + \eta_g^t \bar{\Delta}_t) - \nabla\phi_t(\theta_t) \\ &= \eta_g^t \nabla^2\phi_t(\theta_t + \bar{\eta}\bar{\Delta}_t)\bar{\Delta}_t \\ &= \eta_g^t \nabla^2\psi_t(\xi_t)^{-1}\bar{\Delta}_t \\ &= \eta_g^t \eta_l \tau \nabla^2\psi_t(\xi_t)^{-1}\bar{h}_t, \end{aligned}$$

where $\xi_t = \theta_t + \bar{\eta}\bar{\Delta}_t$, $h_i^t := \Delta_i^t/(\eta_l\tau)$ and $\bar{h}_t := \frac{1}{M}\sum_{i=1}^M h_i^t$. For the second term, as in the proof of Theorem 4.3, we have

$$\begin{aligned} D_{\psi_t}(w^t \mid w^{t+1}) &= \phi_t(\theta_t + \eta_g^t\bar{\Delta}_t) - \phi_t(\theta_t) - \langle \nabla\phi_t(\theta_t), \eta_g^t\bar{\Delta}_t\rangle \\ &= H_t(\eta_g^t) \le \eta_g^t \cdot h_t(\eta_g^t) \\ &= \eta_g^t \cdot \frac{1}{2M}\sum_{i=1}^M \left\|\Delta_i^t\right\|^2 \\ &= \eta_g^t(\eta_l\tau)^2 \cdot \frac{1}{2M}\sum_{i=1}^M \left\|h_i^t\right\|^2. \end{aligned}$$

Thus, we have

$$F(w_{t+1}) - F(w_t) \le -\eta_g^t\eta_l\tau \underbrace{\langle \nabla F(w^t), \nabla^2\psi_t(\xi_t)^{-1}\bar{h}_t\rangle}_{R_1} + \frac{\eta_g^t L(\eta_l\tau)^2}{2\alpha_t} \underbrace{\frac{1}{M}\sum_{i=1}^M \left\|h_i^t\right\|^2}_{R_2}$$

Bounding R_1 We have

$$\begin{aligned} R_1 &= \langle \nabla F(w^t), \nabla^2\psi_t(\xi_t)^{-1}\bar{h}_t\rangle \\ &= \frac{1}{2}\left\|\nabla F(w^t)\right\|^2_{\psi_t(\xi_t)^{-1}} + \frac{1}{2}\left\|\bar{h}_t\right\|^2_{\psi_t(\xi_t)^{-1}} - \frac{1}{2}\left\|\bar{h}_t - \nabla F(w^t)\right\|^2_{\psi_t(\xi_t)^{-1}} \\ &\ge \frac{1}{2\beta_t}\left\|\nabla F(w^t)\right\|^2 - \frac{1}{2M\alpha_t}\sum_{i=1}^M \left\|h_i^t - \nabla F(w^t)\right\|^2, \end{aligned}$$

where the inequality follows from $\beta_t I \succeq \nabla^2\psi_t(\xi_t) \succeq \alpha_t I$.

Bounding R_2 As in the proof of Theorem 2 in Jhunjhunwala et al. (2023), we have

$$\begin{aligned} R_2 &= \frac{1}{M}\sum_{i=1}^M \left\|h_i^t\right\|^2 \\ &= \frac{3}{M}\sum_{i=1}^M \left\|h_i^t - \nabla F_i(w^t)\right\|^2 + 3\sigma_g^2 + 3\left\|\nabla F(w^t)\right\|^2 \\ &= \frac{3}{M}\sum_{i=1}^M \left\|h_i^t - \nabla F_i(w^t)\right\|^2 + 3\sigma_g^2 + 3\left\|\nabla F(w^t)\right\|^2 \end{aligned}$$

Combining the above two inequalities, we have

$$\begin{aligned}
F(w_{t+1}) - F(w_t) &\leq \frac{\eta_g^t \eta_l}{\alpha_t}\tau\Bigg(-\frac{\alpha_t}{2\beta_t}\left\|\nabla F(w^t)\right\|^2 + \frac{1}{2M}\sum_{i=1}^{M}\left\|h_i^t - \nabla F(w^t)\right\|^2 \\
&\quad + \frac{\eta_l \tau L}{2}\left(3\sigma_g^2 + 3\left\|\nabla F(w^t)\right\|^2 + \frac{3}{M}\sum_{i=1}^{M}\left\|h_i^t - \nabla F_i(w^t)\right\|^2\right)\Bigg) \\
&\leq \frac{\eta_g^t \eta_l}{\alpha_t}\left(-\frac{1}{4\kappa}\left\|\nabla F(w^t)\right\|^2 + \frac{1}{M\kappa}\sum_{i=1}^{M}\left\|h_i^t - \nabla F(w^t)\right\|^2 - 3\eta_l\tau L_G\sigma_g^2\right) \\
&\leq \frac{\eta_g^t \eta_l}{\alpha_t}\left(-\frac{1}{8\kappa}\left\|\nabla F(w^t)\right\|^2 + 5\eta_l^2 L^2 \tau(\tau-1)\sigma_g^2 - 3\eta_l\tau L_G\sigma_g^2\right)
\end{aligned}$$

The second inequality follows from $\eta_l \leq \frac{1}{6\tau L_G \kappa}$ and $\beta_t/\alpha_t \leq \kappa$, and the third inequality follows from Lemma 7 in Jhunjhunwala et al. (2023). Rearranging the above inequality, we have

$$\left\|\nabla F(w^t)\right\|^2 \leq O\left(\kappa \cdot \frac{F(w_t) - F(w_{t+1})}{\eta_g^t \eta_l/\alpha_t}\right) + O(\kappa\eta_l^2\tau^2 L^2\sigma_g^2) + O(\kappa\eta_l\tau L\sigma_g^2).$$

Averaging over $t = 0, \ldots, T-1$, we have

$$\min_{t=0,\ldots,T-1}\left\|\nabla F(w^t)\right\|^2 \leq \frac{1}{T}\sum_{t=0}^{T-1}\left\|\nabla F(w^t)\right\|^2 \leq O\left(\frac{F(w_0) - F^*}{\eta_l \sum_{t=0}^{T-1}\tilde{\eta}_g^t}\right) + O(\kappa\eta_l^2\tau^2 L^2\sigma_g^2) + O(\kappa\eta_l\tau L\sigma_g^2).$$

This completes the proof. □

G.2 Proof for Corollary G.4

Proof. First, we show that Assumption G.2 holds for FedDuAdagrad with $\alpha_t = \epsilon$ and $\beta_t = 2\epsilon$. From the definition of ψ_t, we have

$$\begin{aligned}
D_{\psi_t}(w \mid w') &= \frac{1}{2}(w - w')^\top G_t (w - w') \\
&= \frac{1}{2}\sum_{k=1}^{d}[G_t]_{k,k}(w_k - w'_k)^2.
\end{aligned}$$

Since $[\bar{\Delta}_t]_k = \frac{1}{M}\sum_{i=1}^{M}[\Delta_i^t]_k$ and $\|\Delta_i^t\| \leq \eta_l\tau G$, we have

$$\begin{aligned}
[G_t]_{k,k} &= \epsilon + \sqrt{\sum_{s=0}^{t}[\bar{\Delta}_s]_k^2} \\
&\leq \sqrt{\sum_{s=0}^{t}(\eta_l\tau G)^2} \\
&= \epsilon + \sqrt{t+1}\cdot\eta_l\tau G \leq 2\epsilon.
\end{aligned}$$

Thus, we have

$$\epsilon I \preceq G_t \preceq 2\epsilon I,$$

which implies that Assumption G.2 holds.

Therefore, from Theorem G.3, we have

$$\min_{t=0,\ldots,T-1}\left\|\nabla F(w^t)\right\|^2 \leq O\left(\frac{F(w_0) - F^*}{\eta_l \sum_{t=0}^{T-1}\tilde{\eta}_g^t}\right) + O(\eta_l^2\tau^2 L^2\sigma_g^2) + O(\eta_l\tau L\sigma_g^2).$$

It remains to bound $\sum_{t=0}^{T-1} \tilde{\eta}_g^t$. From the definition of $\tilde{\eta}_g^t$, we have

$$\begin{aligned}\sum_{t=0}^{T-1} \tilde{\eta}_g^t &= \sum_{t=0}^{T-1} \frac{\eta_g^t}{\beta_t} \\ &= \sum_{t=0}^{T-1} \frac{\frac{1}{M}\sum_{i=1}^M \|\Delta_i^t\|^2}{\left\|\bar{\Delta}^t\right\|_{G_t^{-1}}^2} \cdot \frac{1}{\beta_t} \\ &= \sum_{t=0}^{T-1} \frac{\frac{1}{M}\sum_{i=1}^M \|\Delta_i^t\|^2}{\left\|\bar{\Delta}_t\right\|^2} \cdot \frac{\left\|\bar{\Delta}^t\right\|^2}{\left\|\bar{\Delta}^t\right\|_{G_t^{-1}}^2} \cdot \frac{1}{\beta_t}.\end{aligned}$$

Since $\beta_t = 2\epsilon$ and $\left\|\bar{\Delta}^t\right\|_{G_t^{-1}}^2 \leq \frac{1}{\epsilon}\left\|\bar{\Delta}^t\right\|^2$, we have

$$\begin{aligned}\sum_{t=0}^{T-1} \tilde{\eta}_g^t &\geq \sum_{t=0}^{T-1} \frac{\frac{1}{M}\sum_{i=1}^M \|\Delta_i^t\|^2}{\left\|\bar{\Delta}_t\right\|^2} \cdot \epsilon \cdot \frac{1}{2\epsilon} \\ &\geq \sum_{t=0}^{T-1} \frac{\frac{1}{M}\sum_{i=1}^M \|\Delta_i^t\|^2}{2\left\|\bar{\Delta}_t\right\|^2}.\end{aligned}$$

Substituting the above inequality, we obtain the result. □

H Detailed Experimental Setup

H.1 Compute Resources and Time

Our experiments were conducted on Intel(R) Xeon(R) Silver 4316 CPU @ 2.30GHz and 8 NVIDIA A100-SXM4-80GB GPUs. The training process (500 rounds) takes about 3 hours for each method and dataset.

H.2 Dataset and Model

We summarize the datasets and models used in our main experiments in Table 3. We also provide the architecture of LSTM and CNN in Table 4 and Table 5, respectively. For ViT experiments, we use the architecture in omiita (2024).

Table 3: Datasets and models used in our experiments

Dataset	Task	Model	# of Classes	License
Synthetic dataset	Regression	Linear	N/A	N/A
CIFAR-10	Image classification	ResNet-18	10	MIT License
CIFAR-100	Image classification	ResNet-18	100	MIT License
FEMNIST	Image classification	CNN	62	BSD 2-Clause
Shakespeare	Next character prediction	LSTM	79	BSD 2-Clause

Synthetic dataset Here, we briefly describe the synthetic dataset used in our experiments. For client $i = 1, \ldots, 20$, we generate 30 samples $\{(x_j, y_j)\}$ ($x_j \in \mathbb{R}^{1000}$) by sampling $x_j \sim \mathcal{N}(0, \Sigma)$ and $y_j = \langle w_{i,j}, x_j \rangle$. Here, Σ is a diagonal matrix with its k-th diagonal element $\Sigma_{k,k} = k^{-1.1}$, and $w_{i,j} \sim \mathcal{N}(w_i, 1), w_i \sim \mathcal{N}(0, 0.1)$.

H.3 Hyperparameters

For a fair comparison, we tune the hyperparameters for each method using grid search. We run algorithms for 500 rounds for the synthetic dataset and 50 rounds for the real-world datasets, and employ the hyperparameters which yield the best validation accuracy averaged over the last 5 rounds. We summarize the best hyperparameters in Table 6. Other hyperparameters are kept the same across all methods. Following Jhunjhunwala et al. (2023), we use weight decay of 10^{-4}, learning rate decay of 0.998, and gradient clipping to stabilize the training for image classification tasks.

Table 4: Architecture of LSTM

Layer	Output Shape	# of Params
Input	[80]	0
Embedding	[80, 256]	20,224
LSTM	[80, 256]	1,052,672
Dropout	[80, 256]	0
Dense	[256, 79]	20,303

Table 5: Architecture of CNN

Layer	Output Shape	# of Params	Kernel Size
Input	[1, 28, 28]	0	
Conv2d	[32, 26, 26]	320	(3, 3)
Conv2d	[64, 24, 24]	18,496	(3, 3)
Dropout	[64, 24, 24]	0	
Dense	[128]	1,179,776	
Dropout	[128]	0	
Dense	[62]	7,998	

Table 6: Best hyperparameters ($\log_{10}$ scale)

	FedAvg		FedExp		FedAvgM		FedExpM	
dataset	η_l	η_g	η_l	ϵ_g	η_l	η_g	η_l	ϵ_g
CIFAR100	-2	0	-2	-4	-2	0	-2	-3
CIFAR10	-2	0	-2	-4	-2	-1	-2	-4
FEMNIST	-1	0	-1	-2	-2	0	-1	-3
shakespeare	0	0	0	-2	0	0	0	-4

	FedAdagrad		FedDuAdagrad		FedAdam		FedDuAdam	
dataset	η_l	η_g	η_l	ϵ_g	η_l	η_g	η_l	ϵ_g
CIFAR100	-2	-4	-2	-1	-2	-4	-2	-1
CIFAR10	-2	-4	-2	-1	-2	-4	-2	-2
FEMNIST	-1	-2	-1	-1	-2	-2	-1	-1
shakespeare	0	-2	0	0	0	-2	0	0

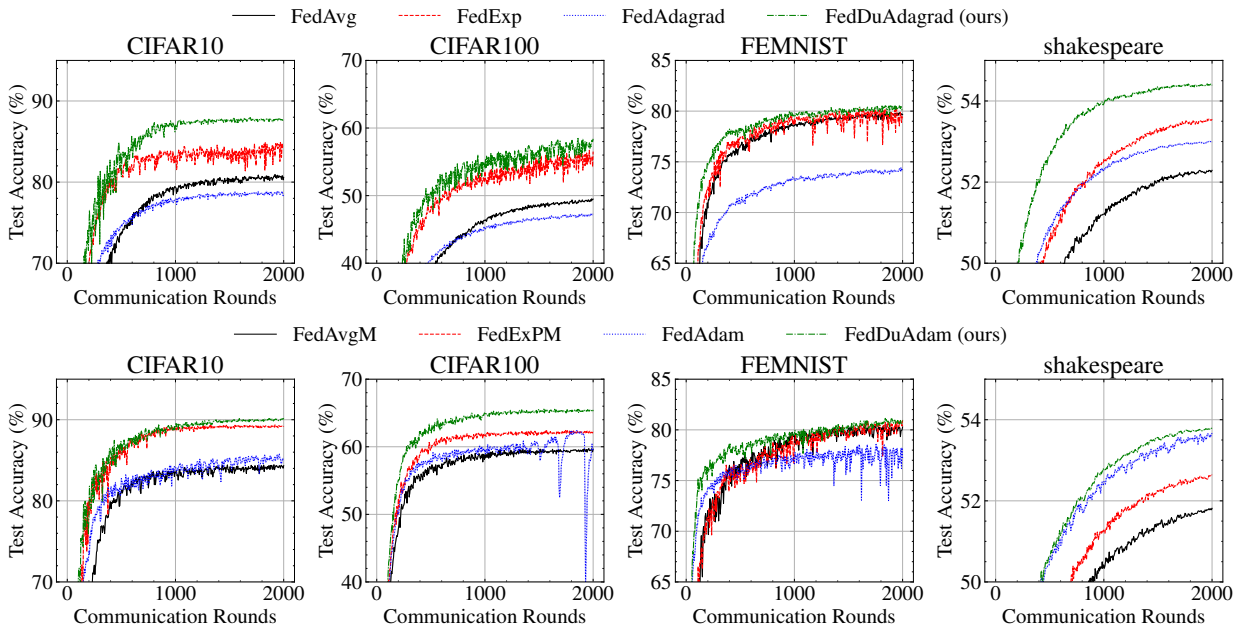

Figure 5: Long term behavior of each algorithm.

Synthetic dataset For the synthetic dataset, We tune η_l over $\{10^{-3}, 10^{-5/2}, 10^{-2}, 10^{-3/2}, 10^{-1}\}$, and η_g over $\{10^{-1}, 10^{-1/2}, 10^{0}, 10^{1/2}, 10^{1}\}$ for FedAvg(M) and $\{10^{-2}, 10^{-3/2}, 10^{-1}, 10^{-1/2}, 10^{0}\}$ for FedOPT. We fix $\epsilon = 0, \epsilon_g = 0$.

Image classification datasets For the image classification tasks, we tune η_l over $\{10^{-2}, 10^{-3/2}, 10^{-1}, 10^{-1/2}, 10^{0}\}$. The grid of η_g is $\{10^{-1}, 10^{-1/2}, 10^{0}, 10^{1/2}, 10^{1}\}$ for FedAvg(M) and SCAFFOLD, and $\{10^{-4}, 10^{-7/2}, 10^{-3}, 10^{-5/2}, 10^{-2}\}$ for FedOPT. The grid of ϵ_g is $\{10^{-3}, 10^{-5/2}, 10^{-2}, 10^{-3/2}, 10^{-1}\}$ for FedDuA, and $\{10^{-4}, 10^{-7/2}, 10^{-3}, 10^{-5/2}, 10^{-2}\}$ for FedExP(M). We use $\{10^{-1}, 10^{-1/2}, 10^{0}, 10^{1/2}, 10^{1}\}$ for FedDuA with Mime since the best η_l is relatively large. We fix $\epsilon = 10^{-9}$ for adaptive methods if not specified.

NLP dataset For the NLP task, we tune η_l over $\{10^{-2}, 10^{-3/2}, 10^{-1}, 10^{-1/2}, 10^{0}\}$. The grid of η_g is $\{10^{-1}, 10^{-1/2}, 10^{0}, 10^{1/2}, 10^{1}\}$ for FedAvg(M) and SCAFFOLD, and $\{10^{-3}, 10^{-5/2}, 10^{-2}, 10^{-3/2}, 10^{-1}\}$ for FedOPT. The grid of ϵ_g is $\{10^{-3}, 10^{-5/2}, 10^{-2}, 10^{-3/2}, 10^{-1}\}$ for FedDuA, and $\{10^{-1}, 10^{-1/2}, 10^{0}, 10^{1/2}, 10^{1}\}$ for FedExP(M). We fix $\epsilon = 10^{-9}$ for adaptive methods if not specified.

I Additional Experimental Results

I.1 Long-term Behavior

To compare the long-term behavior of our proposed method and baselines, we ran the experiments for 2000 rounds, which is sufficiently long to observe the convergence behavior. We show the results in Fig. 5. We see that FedDuA consistently outperforms other methods in terms of convergence speed and final accuracy.

I.2 Comparison with FedProx

In this section, we provide a comparison with FedProx (Li et al., 2020). This is an algorithm that modifies the local objective and thus can be combined with the FedDuA framework. For the FedProx-type local training procedure, we tune the additional hyperparameter μ with the grid $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ following the original paper. As shown in Fig. 6, FedProx-type local training does not improve performance in our setup.

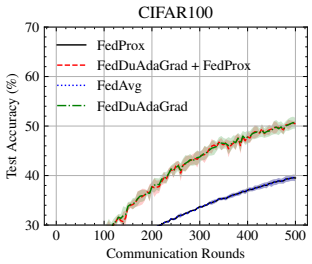

Figure 6: Comparison with FedProx.

I.3 Training Loss and Global Learning Rate

Here, we provide the curve of training loss and global learning rate for FedDuA and baselines. As shown in Fig. 8, FedDuA converges faster than other methods in terms of training loss.

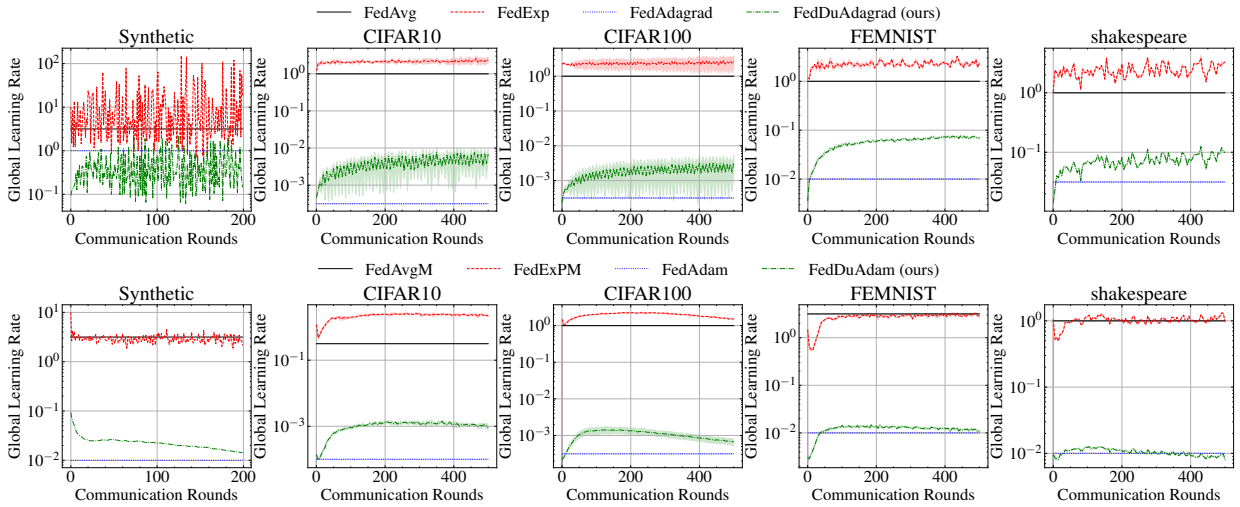

Figure 7: Comparison of global learning rates η_g for different methods.

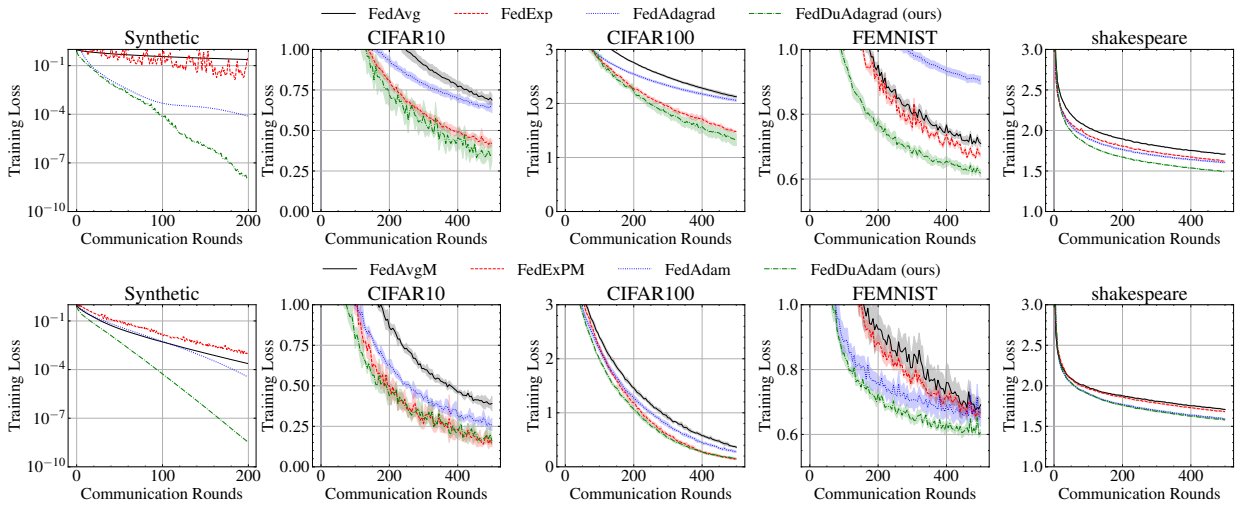

Figure 8: Comparison of training loss for different methods.

I.4 Extreme Non-IID and Lower Client Participation Scenarios

To see the effect of client heterogeneity in real-world scenarios, we conduct experiments with different values of α and different numbers of participating clients M. As shown in Table 7, FedDuA consistently outperforms other methods in terms of final accuracy.

I.5 Additional Results for ViT

We provide additional results for ViT experiments with CIFAR10 in Table 8. We see that FedDuA outperforms other methods by a large margin, which demonstrates the effectiveness of our method in training Transformer-based models in federated learning.

Table 7: Comparison of methods on CIFAR10 and CIFAR100.

Dataset	DuAdam	DuAdagrad	Exp	ExpM
CIFAR10 ($\alpha = 0.1$)	**80.0 ± 1.2**	69.0 ± 1.5	73.7 ± 1.0	78.0 ± 1.8
CIFAR100 ($\alpha = 0.1$)	**58.4 ± 0.9**	45.6 ± 0.4	44.6 ± 1.0	53.5 ± 0.7
CIFAR10 ($M = 10$)	**83.4 ± 1.2**	79.3 ± 1.0	79.5 ± 1.1	**83.4 ± 1.0**
CIFAR100 ($M = 10$)	**61.0 ± 0.8**	49.2 ± 0.5	48.5 ± 0.7	56.2 ± 0.7

Dataset	Adam	Adagrad	AvgM	Avg
CIFAR10 ($\alpha = 0.1$)	75.9 ± 1.8	67.1 ± 1.2	72.7 ± 0.7	63.4 ± 1.0
CIFAR100 ($\alpha = 0.1$)	54.1 ± 0.6	44.0 ± 0.3	52.2 ± 0.8	38.9 ± 0.6
CIFAR10 ($M = 10$)	79.1 ± 2.1	70.4 ± 0.8	79.1 ± 0.9	72.1 ± 0.7
CIFAR100 ($M = 10$)	52.2 ± 1.0	37.6 ± 0.3	55.7 ± 0.9	39.3 ± 0.6

Table 8: Comparison on CIFAR10 (ViT).

Dataset	DuAdam	DuAdagrad	Exp	ExpM
CIFAR10 (ViT)	**76.4 ± 1.3**	68.0 ± 0.4	62.9 ± 1.6	57.9 ± 5.2

Dataset	Adam	Adagrad	AvgM	Avg
CIFAR10 (ViT)	73.0 ± 0.7	68.4 ± 0.4	69.6 ± 0.6	59.3 ± 1.3