

FedADMM: A Robust Federated Deep Learning Framework with Adaptivity to System Heterogeneity

Yonghai Gong^{* 1}, Yichuan Li^{* 2} and Nikolaos M. Freris^{† 1}

¹*School of Computer Science and Technology, University of Science and Technology of China, Hefei, China*

²*Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, IL 61820, USA*

gongyh@mail.ustc.edu.cn, yli129@illinois.edu, nfr@ustc.edu.cn

Abstract—Federated Learning (FL) is an emerging framework for distributed processing of large data volumes by edge devices subject to limited communication bandwidths, heterogeneity in data distributions and computational resources, as well as privacy considerations. In this paper, we introduce a new FL protocol termed FedADMM based on primal-dual optimization. The proposed method leverages dual variables to tackle statistical heterogeneity, and accommodates system heterogeneity by tolerating variable amount of work performed by clients. FedADMM maintains identical communication costs per round as FedAvg/Prox, and generalizes them via the augmented Lagrangian. A convergence proof is established for nonconvex objectives, under no restrictions in terms of data dissimilarity or number of participants per round of the algorithm. We demonstrate the merits through extensive experiments on real datasets, under both IID and non-IID data distributions across clients. FedADMM consistently outperforms all baseline methods in terms of communication efficiency, with the number of rounds needed to reach a prescribed accuracy reduced by up to 87%. The algorithm effectively adapts to heterogeneous data distributions through the use of dual variables, without the need for hyperparameter tuning, and its advantages are more pronounced in large-scale systems.

I. INTRODUCTION

The abundance of data constitutes the fuel for the proliferation of Machine Learning (ML) in all aspects of life. In many applications, data are being generated in large quantities in a distributed fashion: for instance, crowdsourcing, mobile phones, autonomous vehicles, medical centers, distributed sensors in smart grids, to name but a few. The sheer size of the data alongside the urge to protect personal privacy prevent traditional distributed machine learning practices from being directly applied. The well-studied framework of data-center distributed learning [1], [2] often involves maneuvering raw data from machine to machine, which not only violates privacy restrictions but also becomes infeasible when limited networking resources fail to cope with large data sizes. Besides, the rapid increase of the computational power of personal devices such as smartphones surges pushing computation to the edge as opposed to the cloud.

Federated Learning (FL) was proposed in [3], [4] as a new paradigm for tackling large-scale distributed machine learning problems with distinct features compared to the data-center



Fig. 1: Architecture of FL. Active clients ① download the current global model from the server, ② perform local training based on their data, ③ upload an update message to the central server. Server ④ aggregates clients' updates to form a new global model.

setting: (i) massive client population; (ii) limited communication resources; (iii) low device participation rate and unreliable connections; (iv) stringent privacy considerations. A canonical setting of FL is shown in Fig. 1, where a server coordinates with clients to compute a global ML model. The introduction of this framework was carried in [4], wherein the authors proposed FedAvg as the main algorithm. At each round of FedAvg, a small fraction of clients is selected and each selected client downloads the global model. The selected clients then proceed to update the model by running SGD using their local data, and upload the updated model to the server. Finally, a new global model is obtained by the server through averaging the updated models. Despite the fact that FedAvg is arguably the most widely adopted method of choice for FL, it does not address either *system heterogeneity* or *statistical heterogeneity*. Specifically, FedAvg does not explicitly handle the *straggler* problem in a heterogeneous network: the process may stall in the face of unreliable network connections or client unavailability. Moreover, FedAvg may diverge when data distributions across clients violate the IID (independent and identically distributed) assumption [4]–[7].

The authors in [8] proposed FedProx, which augments a proximal term to the local training problem to counter possible divergence resulting from statistical heterogeneity. Additionally, this method accommodates system heterogeneity by allowing variable amount of work at each selected client. Nevertheless, the performance of FedProx is sensitive

^{*} Equal contribution. [†] Corresponding author: Nikolaos M. Freris.

to the selection of the proximal coefficient, whose tuning depends on prior knowledge on statistical distributions and system sizes [8]. SCAFFOLD [9] introduces client and server *control variates*, which effectively act as tracking variables for local and global information. It is reported to outperform FedAvg/Prox through utilizing the similarity of data across clients, but its communication cost per round is doubled due to the extra *control variates*.

Another branch of work has roots in primal-dual optimization methods, such as Generalized Method of Multipliers and Alternating Direction Method of Multipliers (ADMM) [10]–[12] that have enjoyed much success in distributed optimization. Such methods provide a general framework which not only encompasses a wide range of objectives—such as low-rank matrix completion, nonconvex neural network training, and reinforcement learning [13]–[15]—but it can also be tailored for specific application demands, such as event-triggered communication and inexact computing [16]–[20].

In this paper, we propose FedADMM, a primal-dual algorithm that automatically adapts to data heterogeneity while achieving significant convergence speedup compared to the state-of-the-art. Each round of FedADMM involves a small fraction of clients and allows a plethora of computing choices for per user. A distinctive attribute of FedADMM is the use of dual variables to guide the local training process (see Section III-A for detailed discussion), which effectively account for automatic adaptation to data distributions without tuning. This specifically addresses a key challenge encountered in the FL setting known as avoiding *client drift*: local training performed at clients has to be carefully designed according to statistical variations so as to prevent the model from overfitting to a specific selected client’s data. We note that compared to SCAFFOLD, where the acceleration is achieved through storing and communicating the *control variates* that capture local gradient information, FedADMM maintains the exact same communication cost per round as FedAvg/Prox. Using a limited amount of extra storage for the dual variables (which is readily available and inexpensive in practical FL scenarios), FedADMM effectively harnesses the capabilities of edge devices and achieves significant communication and computation savings due to the automated adaptation to statistical variations.

Contributions:

We present a new framework for Federated Learning, termed FedADMM.

- Through the use of local augmented Lagrangian and dual variables, FedADMM addresses both system and statistical heterogeneity.
- By introducing a generalized server update step size, the proposed method achieves a fine balance between convergence speed and robustness against oscillations.
- We establish a convergence rate of $\mathcal{O}(\frac{1}{T}) + \mathcal{O}(1)$ for the proposed algorithm under less restrictive assumptions than the state-of-the-art. Moreover, this rate matches the complexity lower bound for reaching an ε -stationary solution.

- A crucial feature of FedADMM is that no restrictions are imposed on the client activation scheme other than the necessary infinitely often participation. This is achieved by clients solving local augmented Lagrangian subproblems that are strongly convex, which ensures progress is made without diverging under a most general participation scheme.
- We have conducted numerous experiments on distributed neural network training from real datasets in both IID and non-IID settings, which attest that FedADMM achieves high accuracy in fewer rounds compared to the state-of-the-art (an average 72% and a maximum 87% of communication savings over the best performing baseline are reported over several datasets, population sizes, and both statistical and system heterogeneity). Notably, the advantages of FedADMM are most pronounced in large-scale networks with heterogeneous data distributions.

II. RELATED WORK

Existing primal-dual algorithms for FL include FedHybrid [21] and FedPD [22]. FedHybrid is a synchronous method (requiring all clients to update at each round) that tackles system heterogeneity by separating clients into two groups, one performing gradient updates and the other Newton updates. However, this setting does not encapsulate federated deep learning (Newton’s method is not applicable for large neural networks, while also convergence is established under strong convexity assumptions). FedPD employs gradient-based (GD or SGD) local training with variable amount of work at clients. Similar to our proposed method, it invokes dual variables to capture the discrepancy between local models and the global model. Nevertheless, FedPD adopts full client participation, in that (i) *all clients* update their local models and dual variables at each round and (ii) at each round, with fixed probability, either *all clients* communicate with the server to update the global model, or there is no communication. Therefore, the frequency of updating the global model is limited by the probability of communicating. We note that all clients in FedHybrid/PD are continuously engaged in computing, which incurs large computational overhead to devices. Moreover, requiring all clients to communicate at the same time not only imposes strict requirements on clients’ availability and network bandwidth, but also exacerbates the straggler problem (where the server has to wait for the slowest client before proceeding to the next round).

In contrast, our proposed solution (i) applies partial client participation for both computation and communication, (ii) updates the global model at each round, and (iii) leverages tracking in the global model update rule (which accounts for accelerated convergence with smoother paths). The server in FedADMM communicates with a fraction of clients during each round and clients only perform local training when selected. We establish convergence without restrictions on statistical heterogeneity among the networks, or the fraction of participating users. This is in contradistinction to the

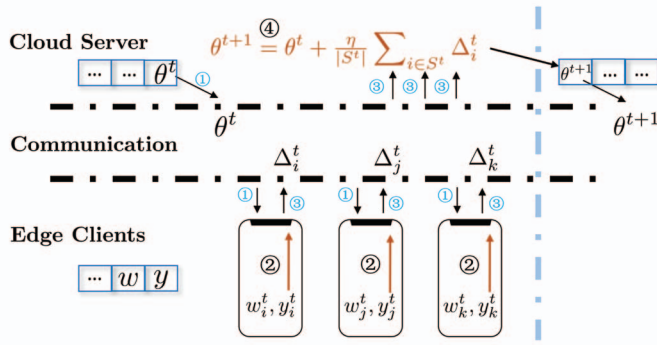


Fig. 2: FedADMM on selected clients during the t -th round. Selected clients ① download the current server model, ② carry local training and update their dual variables (Alg. 1, line 6), ③ upload update messages (Alg. 1, line 8). Server ④ aggregates messages from selected edge clients to update the global model.

analysis of other baseline methods [5], [8], [23] which requires assumptions on data dissimilarity that may be unrealistic in practices when data are individually generated by each client [24]–[26]. Our proposed method has close connections to randomized ADMM [27], which in the FL setting amounts to randomly activating clients and the server. However, it requires that subproblems to be solved exactly, which may be impractical in consideration of heterogeneous hardware conditions and data volumes. On the other hand, FedADMM accounts for inexact solutions from local problems and allows variable amount of training performed at clients (quantified by a tunable parameter). Another related line of work is termed asynchronous ADMM [28]–[30], whose focus is on alleviating the straggler problem mentioned earlier. However, assumptions imposed therein such as the *bounded delay* assumption (i.e., each user needs to be active at least once every some number of rounds) may never be satisfied in FL settings.

III. ALGORITHM

The objective of FL is cast as a loss minimization problem:

$$\underset{\theta \in \mathbb{R}^d}{\text{minimize}} \quad \sum_{i=1}^m f_i(\theta), \quad (1)$$

where m is the number of clients and $f_i(\theta) = \frac{\alpha_i}{n_i} \sum_{k=1}^{n_i} l_{ik}(\theta)$ captures the weighted local training loss. Common choices of weights include $\alpha_i = \frac{n_i}{n}$, where $n = \sum_{i=1}^m n_i$ (i.e., weighting clients proportionally to their data volumes) or $\alpha_i = 1$ (i.e., equal weights, which serves to avoid overfitting to clients with more data, and is also the choice in our experiments). The goal of FL is to solve (1) over model parameter θ using data distributed across clients. In contrast to data-center distributed machine learning, communication is the bottleneck for FL applications. The large number of clients in federated systems renders infeasible synchronous communication (i.e., all clients communicating with the server at a given round) with large message sizes; this is due to bandwidth limitations and the existence of stragglers. Moreover, the large data volumes involved along with the leakage of personal information from

Algorithm 1 FedADMM

```

1: Input: Total number of rounds  $T$ , server step size  $\eta$ ,
   proximal coefficient  $\rho$ .
2: for  $t = 0, 1, \dots, T-1$  do
3:   Server selects  $S^t \subset [m]$ 
   Clients: // In parallel
4:   for  $i \in S^t$  do
5:     download  $\theta^t$  from the server
6:      $(w_i^{t+1}, y_i^{t+1}) \leftarrow \text{ClientUpdate}(i, \theta^t)$ 
7:     compute update message  $\Delta_i^t$  as in (4)
8:     upload  $\Delta_i^t$  to the server
9:   end for
   Server:
10:   $\theta^{t+1} \leftarrow \theta^t + \frac{\eta}{|S^t|} \sum_{i \in S^t} \Delta_i^t$ 
11: end for

ClientUpdate( $i, \theta$ ): // Store  $w_i$  and  $y_i$ 
12: Input: Local epoch number  $E_i$ , client learning rate  $\eta_i$ .
13: create batches  $\mathcal{B}$ 
14: for  $k = 0, 1, \dots, E_i - 1$  do
15:   for batch  $b \in \mathcal{B}$  do
16:     compute batch gradient  $\nabla f_i(w_i, b)$ 
17:      $w_i \leftarrow w_i - \eta_i (\nabla f_i(w_i, b) + y_i + \rho(w_i - \theta))$ 
18:   end for
19: end for
20:  $y_i \leftarrow y_i + \rho(w_i - \theta)$ 
21: return  $w_i, y_i$ 

```

raw data prevents the server from collecting local data for centralized processing. Thus, a candidate algorithm for FL should meet the following requirements: (i) communication efficiency; (ii) low client participation rate per round; (iii) handling both system and statistical heterogeneity; (iv) no transmission of client raw data. We proceed to present FedADMM whose design meets all these guidelines¹.

A. Proposed Algorithm: FedADMM

We first reformulate (1) into the following consensus setting that is more suitable for our development:

$$\underset{w_i, \theta \in \mathbb{R}^d}{\text{minimize}} \quad \sum_{i=1}^m f_i(w_i), \quad \text{s.t. } w_i = \theta, \quad \forall i \in [m]. \quad (2)$$

Problem (2) is equivalent to (1) in the sense that the optimal solutions coincide. This formulation naturally fits into the FL setting, where w_i can be interpreted as the local model held by client i and θ as the global model held by the server.

At the beginning of the t -th round, a subset of clients is selected (denoted as S^t) and each selected client downloads the server model θ^t . We note that a favorable attribute of FedADMM lies in that client selection (step 3 of Alg. 1) can be carried by any mechanism (e.g., an adaptive rule based on

¹Although not the focus of this paper, we note that standard privacy-preserving methods, such as differential privacy and secure multi-party computation [31]–[33] can be combined with FedADMM.

device operating conditions such as energy and bandwidth) as long as it guarantees non-zero probability of participation. Local training is performed to update the client model as $w_i^{t+1} \approx \underset{w_i}{\operatorname{argmin}} \mathcal{L}_i(w_i, y_i^t, \theta^t)$, where

$$\mathcal{L}_i(w_i, y_i^t, \theta^t) = f_i(w_i) + (y_i^t)^\top (w_i - \theta^t) + \frac{\rho}{2} \|w_i - \theta^t\|^2. \quad (3)$$

We note that $y_i^t \in \mathbb{R}^d$ is the local dual variable held by client i and $\rho > 0$ is the coefficient of the quadratic term. Both the dual variables and the quadratic term serve to strike a balance between updating model parameters using local data while staying consistent with the server model (which serves to incorporate information from all participants). We note that FedProx [8] similarly solves (3) with $y_i^t \equiv 0$. While this provides some level of safeguard against *client drift*, competitive performance of FedProx relies on careful tuning of ρ [8], [9]. The addition of the dual variables helps FedADMM to achieve automatic adaptation to heterogeneous data distributions and significantly alleviates the problem of hyperparameter tuning; we postpone detailed discussions on dual variables to the end of this section. In addition to the local model update (Alg. 1, lines 14-19), the dual variable for each selected client is updated in line 20. We combine the primal and dual variables to what we call *augmented model* ($w_i + \frac{1}{\rho} y_i$), and use $\Delta_i^t \in \mathbb{R}^d$ to denote the update message of client i to the server, that is the difference between successive *augmented models*:

$$\Delta_i^t = \left(w_i^{t+1} + \frac{1}{\rho} y_i^{t+1} \right) - \left(w_i^t + \frac{1}{\rho} y_i^t \right). \quad (4)$$

The server updates the global model after gathering update messages from selected clients as:

$$\theta^{t+1} = \theta^t + \frac{\eta}{|S^t|} \sum_{i \in S^t} \Delta_i^t, \quad (5)$$

where $\eta > 0$ is the server gathering step size. Different choices of η are suitable for different scenarios in terms of scale of the system and statistical variations. We empirically observe that setting $\eta = 1$ gives rise to fast training speed, while setting $\eta = |S^t|/m$ helps to eliminate oscillatory behaviors when significant heterogeneity is detected. In contrast to FedAvg/Prox, where the global model is updated using information only from the current models of clients $i \in S^t$, the server in FedADMM effectively incorporates past information. This is accomplished through the tracking update rule in (5) and serves to provide additional safeguard against oscillations emanating from heterogeneous data and the stochastic nature of the algorithm.

We emphasize that in a practical scenario (3) is not required to be solved exactly, but instead the updated local model w_i^{t+1} can be computed inexactly in the following sense:

$$\|\nabla_{w_i} \mathcal{L}_i(w_i^{t+1}, y_i^t, \theta^t)\|^2 \leq \varepsilon_i, \quad (6)$$

where $\varepsilon_i \geq 0$ prescribes the local accuracy. Note that a variable accuracy level is allowed for different clients (a smaller

ε_i corresponds to a more accurate solution). This is useful since user devices in FL systems feature varying hardware conditions such as computational power, network connectivity, battery level, etc. For the sake of simplicity and comparison with baseline methods, this is done in Algorithm 1 by running E_i epochs of SGD to compute w_i^{t+1} (line 14); the equivalence with (6) emanates from the fact that local subproblems (3) are strongly convex. Nevertheless, other updating schemes are also feasible such as gradient descent and quasi-Newton updates like L-BFGS. Therefore, we accommodate system heterogeneity by letting clients decide to perform different amount of work according to their local environments. Finally, we note that each client in FedADMM holds the primal-dual pair (w_i, y_i) , while it only communicates the augmented model difference Δ_i^t , i.e., the communication cost per round is identical with FedAvg and FedProx.

Dual variables: Contrary to primal only methods (FedAvg/Prox and SCAFFOLD), the augmented Lagrangian \mathcal{L}_i is used for the training of local model parameters w_i . Invoking the augmented Lagrangian is effective for two reasons: (i) by a proper selection of the quadratic proximal coefficient ρ , it is possible to make \mathcal{L}_i in (3) strongly convex with respect to w_i (this is true even for non-convex f_i); therefore, theoretical guarantees for reaching a prescribed local accuracy level ε_i in (6) are readily obtainable. (ii) the dual variable y_i in (3), in addition to penalizing the discrepancy between the local model and the global model, further provides a quantitative measure of the benefit brought by letting local models be different from the server model during the initial stage of training. To illustrate, note that the stationary points for each subproblem, i.e., the solutions to $\nabla f_i(\theta) = 0$, tend to be different from one another, especially in FL settings with heterogeneous data distributions across clients. The stationary condition of (1), $\sum_{i=1}^m \nabla f_i(\theta^*) = 0$, indicates that the cost reduction by following any given $-\nabla f_i(\theta^*)$ may be partly or entirely cancelled by the increase of other $f_j(\theta^*)$, $j \neq i$. This is encoded within the KKT conditions for problem (2): $\nabla f_i(w_i^*) + y_i^* = 0$ for all $i \in [m]$ and $\sum_{i=1}^m y_i^* = 0$. Therefore, $y_i \in \mathbb{R}^d$ can be interpreted as a signed “price vector” (with entries being positive or negative) which not only quantifies the cost of w_i^{t+1} being different from θ^t , but also provides a *direction* of the adjustments needed for agreement. This effectively achieves an automatic adaptation to statistical variations in that it safeguards against *client drift*, i.e., w_i converging to client optima.

B. Connections with Existing Work

We show how FedADMM generalizes FedAvg and FedProx. Recall the augmented Lagrangian defined in (3). By setting $y_i \equiv 0$ (i.e., also omitting line 20 in Alg. 1), we recover the local training problem of FedProx. If additionally ρ is set to 0, one recovers the local training problem of FedAvg. The main motivation of these two terms is to tackle a fundamental challenge reminiscent of FL, i.e., the balance between (i) extensive local training (aiming for faster convergence speed of the entire FL process, thus a reduction of the to-

tal number of communication rounds), and (ii) deterring client models from overfitting to their local data. In addition to the quadratic proximal term introduced in FedProx, FedADMM further employs dual variables, whose merits in guiding the update of local models w_i were discussed in the prequel. We note that SCAFFOLD similarly stores additional variables at clients and the server to counter statistical variations. However, the variable introduced therein can not be combined into a single message (as is accomplished with our *augmented model* and update messages in (4)), whence SCAFFOLD *doubles* the size of the uploading message to the server. On the other hand, FedADMM only stores an additional vector at clients but maintains the exact same upload size. To this end, we believe FedADMM is an effective scheme for harnessing local device capabilities to alleviate communication costs, which constitutes a primordial motivation for FL. Extensive experiments illustrate a substantial reduction of communication in all test cases (72% on average) over the best performing baseline.

IV. ANALYSIS

In this section, we provide the convergence analysis for FedADMM. We adopt the following standard assumptions.

Assumption 1. Each local loss function $f_i(\cdot)$ is L -Lipschitz smooth, i.e., $\forall w, w' \in \mathbb{R}^d$, the following inequality holds:

$$\|\nabla f_i(w) - \nabla f_i(w')\| \leq L\|w - w'\|, \quad i \in [m].$$

Assumption 2. The objective of problem (1) is lower bounded, i.e., there exists $f^* \in \mathbb{R}$ such that $\sum_{i=1}^m f_i(w) \geq f^*, \forall w \in \mathbb{R}^d$.

The above assumptions are minimally restrictive, since loss functions are typically non-negative and the structure of the deep neural networks considered in the experiments ascertains Lipschitz smoothness. We further denote the aggregated Lagrangian as $\mathcal{L} = \sum_{i=1}^m \mathcal{L}_i$ and we define a non-negative function $V(\{w_i^t\}, \{y_i^t\}, \theta^t)$ to measure the optimality gap:

$$V^t := \|\nabla_{\theta} \mathcal{L}^t\|^2 + \sum_{i=1}^m \left(\|\nabla_{w_i} \mathcal{L}^t\|^2 + \|w_i^t - \theta^t\|^2 \right). \quad (7)$$

Since each term of V^t is non-negative, $V^t = 0$ if and only if $\|\nabla_{\theta} \mathcal{L}^t\| = \|\nabla_{w_i} \mathcal{L}_i^t\| = \|w_i^t - \theta^t\| = 0, \forall i$. It can be further verified that these terms are 0 if and only if a stationary solution of (2) is reached.

Theorem 1. Let assumptions 1 and 2 hold. Assume each client has a probability of being selected at each round that is lower bounded by a positive constant $p_{\min} > 0$. Let $\eta = |S^t|/m$, and select $\rho > (1 + \sqrt{5})L$, then the following holds:

$$\frac{1}{mT} \sum_{t=0}^{T-1} \mathbb{E}[V^t] \leq \frac{1}{mT} \frac{c_2}{c_1} (\mathcal{L}^0 - f^* + \frac{m}{2L} \varepsilon_{\max}) + c_3 \varepsilon_{\max}, \quad (8)$$

where $\varepsilon_{\max} = \max_i \varepsilon_i$, $c_1 = p_{\min} \left(\frac{\rho-2L}{2} - \frac{2L^2}{\rho} \right)$, $c_2 = 3(L^2 + \rho^2) + 2(1 + \frac{2L^2}{\rho^2})$, and $c_3 = \left(3 + \frac{16}{\rho^2} + \frac{c_2}{c_1} \cdot \frac{\rho+16L}{2L\rho} \right)$.

Remark 1: Note that the loss functions considered in our experiments admit $L = \mathcal{O}(1)$, i.e., the Lipschitz constant does not scale with the client population (this is because the input

TABLE I: Number of communication rounds required to reach an ε -stationary solution. Constants B, G measure data dissimilarity and boundedness of gradients in (9) and (10) respectively, m is the total number of clients, and S denotes the number of active clients.

Method	Number of communication rounds
FedAvg [4], [9]	$\mathcal{O}(1/\varepsilon^2 \cdot (m - S)/mS + G/\varepsilon^{3/2} + B^2/\varepsilon)$
FedProx ¹ [8]	$\mathcal{O}(B^2/\varepsilon)$
SCAFFOLD [9]	$\mathcal{O}(1/\varepsilon^2 + 1/\varepsilon \cdot (m/S)^{\frac{2}{3}})$
FedPD ² [22]	$\mathcal{O}(1/\varepsilon)$
FedADMM	$\mathcal{O}(1/\varepsilon \cdot (m/S))$

¹FedProx requires $S > B^2$ to ensure convergence.

²FedPD requires all clients to communicate at the same time.

data are uniformly bounded, e.g., images). To reach an ε -exact solution in such cases, **Theorem 1** suggests a complexity of $\mathcal{O}(1/\varepsilon p_{\min})$ by setting $\rho \in \mathcal{O}(L) = \mathcal{O}(1)$. We compare the convergence property of FedADMM with other baselines in Table I. Note that only FedProx and FedPD among baseline methods enjoy the same favorable dependency on solution accuracy as FedADMM, i.e., $\mathcal{O}(1/\varepsilon)$. However, they both impose additional restrictive conditions to achieve this. In particular, FedProx requires selecting $S > B^2$, where B is a measure of data dissimilarity across users as in (9). This imposes constraints on minimum number of active clients. On the contrary, our analysis for FedADMM establishes convergence under arbitrary client participation and does not require (9) (i.e., $B = \infty$ is also acceptable). Besides, FedPD requires all clients to communicate simultaneously (i.e., it does not allow partial client participation), which we believe is unrealistic for large networks where stragglers and unavailability of machines is ubiquitous. On the other hand, SCAFFOLD and FedADMM both allow arbitrarily large values of G (gradient norm bound as in (10)) and B , but SCAFFOLD doubles the communication cost per client compared to FedADMM.

Remark 2: For simplicity, we stated **Theorem 1** assuming users are active with a non-zero probability at each round. In fact, the weaker condition is needed that each client is active infinitely often, which is necessary for the correctness of any iterative distributed algorithm (otherwise some clients are ignored). This in turn allows for a general activation scheme and thus incurs no lower bound on minimum number of active clients per round. The reason that this is possible is the introduction of dual variables and the quadratic penalty term which alleviate the effect of possible unbalanced client activation. Further insight can be gained by inspecting the inequality (31), where we derive the expected decrement of the augmented Lagrangian. This suggests that any activation scheme can be supported by our analysis: using time dependent participating probability p_i^t for each client, our convergence proof carries as long as $\sum_{t=1}^{\infty} p_i^t = \infty$ (infinitely often participation) by invoking the Borel-Cantelli Lemma II.

We emphasize that our choice of ρ can be made without any a priori knowledge on system sizes, number of participating clients, and statistical variation across clients. The dynamic nature of FL makes such selection rule of the hyperparameter useful and practical. This is corroborated by our experiments

in Section V which shows that FedADMM performs well with constant ρ in settings with different system sizes and data distributions.

In addition to the assumptions 1-2, the following conditions have been imposed when analyzing existing state-of-the-art methods. Our analysis *does not require any of these conditions*. (*Data dissimilarity*) The local client loss functions satisfy:

$$\sum_{i=1}^m \|\nabla f_i(w)\|^2 \leq \left\| \sum_{i=1}^m \nabla f_i(w) \right\|^2 B^2, \quad (9)$$

for some constant B and all $w \in \{w \mid \|\sum_{i=1}^m \nabla f_i(w)\|^2 > \varepsilon\}$. Note that since the accuracy ε can be chosen arbitrarily small, this condition is restrictive as it would effectively force $B \rightarrow \infty$ as $\varepsilon \rightarrow 0$, unless data distributions are similar across clients. In other words, in scenarios where accurate solutions are pursued and when data distributions are heterogeneous, the analysis in [8] dictates that all clients need to participate in each round (which is highly undesirable in FL) when such an assumption is imposed.

(*Bounded Gradient*) There exists a constant G that uniformly bounds the client loss function gradient, i.e.,

$$\|\nabla f_i(w_i^t)\| \leq G, \quad \forall w_i^t \in \mathbb{R}^d. \quad (10)$$

The above assumption was imposed in [5], [23] to analyze FedAvg for bounding iterate paths during local training. However, it may be unrealistic to impose a uniform bound in the face of large populations and data heterogeneity.

V. EXPERIMENTS

We conducted experiments on three datasets, namely MNIST [34], Fashion MNIST (FMNIST) [35], and CIFAR-10 [36]. We compare our proposed method against FedSGD, FedAvg [4], FedProx [8], and SCAFFOLD [9] using two popular CNN models [4], [34] with two convolutional layers each and number of parameters as detailed in Table II. We do not compare against FedHybrid [21] and FedPD [22], since they are both involve communication rounds that require participation of all users, which we deem unrealistic for large-scale FL scenarios. In brief, our experiments unravel three main findings: (i) FedADMM achieves high accuracy in both IID and non-IID settings within *substantially fewer rounds*, which translates to large communication savings; (ii) FedADMM features *robust convergence* against both statistical and system heterogeneity, and requires no hyperparameter tuning across the IID and non-IID settings. This is in contrast to FedProx whose proximal coefficient needs to be carefully chosen according to data dissimilarity, which is difficult to assess in a practical scenario; (iii) FedADMM outperforms all baseline methods in *all test scenarios*, with the advantages

being most pronounced in large-scale systems with heterogeneous data distributions. This is as opposed to other methods, where an improved performance is restricted to either the IID or the non-IID setting.

A. Experimental Setup

We study ten-class image classification on three real datasets with details specified below. Two CNN models are used for our experiments, both with a convolutional module (two 5×5 convolutional layers, each followed by 2×2 max pooling layers), and a fully connected layer module. The input of the two models is a flattened image with dimension 784 and 3,072, respectively, while the output for both is a class label from 0 to 9. Table II summarizes the datasets, model sizes, and target accuracies used in our experiments. We denote E as the local epoch number, B as the local batch size, and C as the fraction of clients that participate in the training process during each round. All experiments are implemented in PyTorch on a system with 2 Intel® Xeon® Silver 4210 CPUs and 2 NVIDIA® Tesla® V100S GPUs. Clients are selected uniformly at random during each global round, and the number of participating clients is set to 10% ($C = 0.1$) of total clients in all cases, while SGD was chosen as the local solver in all cases. This was done for the sake of comparison with the baselines, but our framework allows for arbitrary client participation and local solver. The system heterogeneity (i.e., variable computational capabilities across clients) is captured by letting each client select the local epoch number uniformly between 1 and E in FedADMM as well as in FedProx. The number of local epochs for FedAvg and SCAFFOLD are fixed to be E ; this was done in order to compare against baselines in their principal description. We adopt random initialization for the global model in all algorithms, zero initialization for dual variables in FedADMM, and zero initialization for *control variates* in SCAFFOLD (as recommended). Recall that the use of *control variates* doubles the upload cost of SCAFFOLD compared to other baselines. We set the server gathering step size η to be 1 unless otherwise specified and present all results averaged over five runs. The code is available at: github.com/YonghaiGong/FedADMM.

Data Distribution. For the IID setting, data are evenly distributed to clients. In contrast, to account for data heterogeneity (non-IID setting), we first arrange the training data by label and then distribute them evenly into shards: each client is assigned two shards uniformly at random. We note that this is a rather extreme representative of data heterogeneity, since clients tend to overfit to the specific classes of assigned data.

Hyperparameters. The learning rates of the local SGD solver for all algorithms are selected from the candidate set $\{0.01, 0.1, 0.2, 0.5\}$ for best performance. Recall that both FedADMM and FedProx make use of a quadratic penalty term and require the setting of hyperparameter ρ . It has been observed in [8], [9], and is further supported in our experiments, that careful tuning of ρ is required to achieve the best performance for FedProx (see discussion on **Proximal Parameter** ρ in Section V-B). For fairness, we fixed $\rho = 0.1$

TABLE II: Experimental Setup.

Model	# of para	Dataset	Target Accuracy
CNN 1	1,663,370	MNIST	97%
		FMNIST	80%
CNN 2	1,105,098	CIFAR-10	45%

TABLE III: Comparison of number of communication rounds along with the speedup relative to FedSGD to reach a target accuracy (100+ indicates the target accuracy was not reached in 100 rounds). For MNIST dataset with 100 clients, $E = 5$, $B = 200$ was used for both IID and non-IID data distributions. In all 1,000 client settings, we set $E = 20$, $B = 10$ for the non-IID case, and $E = 20$, $B = \infty$ (full-batch) for the IID case. Reduction of number of communication rounds using FedADMM is computed over the *best baseline method* in each experiment.

	MNIST (100 clients)		MNIST (1,000 clients)		FMNIST (1,000 clients)		CIFAR-10 (1,000 clients)	
	IID	Non-IID	IID	Non-IID	IID	Non-IID	IID	Non-IID
FedSGD	297	250	201	269	390	530	186	202
FedADMM	10(29.7\times)	33(7.6\times)	8(25.1\times)	13(20.7\times)	3(130.0\times)	7(75.7\times)	7(26.6\times)	9(22.4\times)
FedAvg	19(15.6 \times)	77(3.2 \times)	61(3.3 \times)	73(3.7 \times)	10(39.0 \times)	33(16.1 \times)	24(7.8 \times)	50(4.0 \times)
FedProx	29(10.2 \times)	100+	78(2.6 \times)	100+	14(27.9 \times)	61(8.7 \times)	32(5.8 \times)	68(3.0 \times)
SCAFFOLD	27(11.0 \times)	76(3.3 \times)	61(3.3 \times)	84(3.2 \times)	12(32.5 \times)	40(13.3 \times)	37(5.0 \times)	100+
Reduction	47.4 %	56.6%	86.9%	82.2%	70.0%	78.8%	70.8%	82.0%

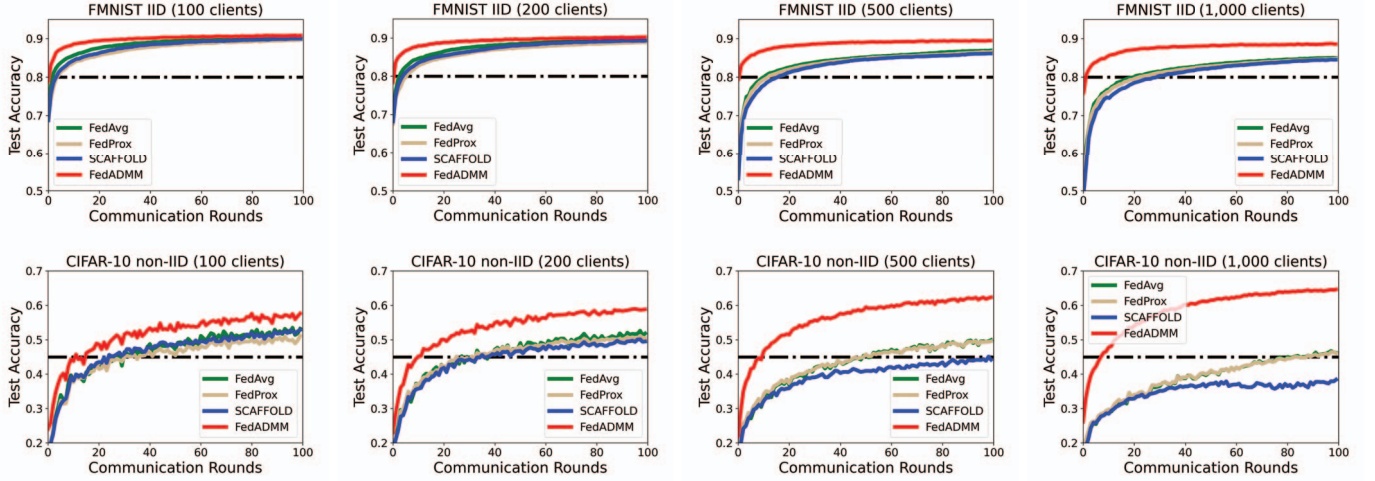


Fig. 3: Convergence paths in systems with variable number of users (prescribed accuracies of 80% for FMNIST and 45% for CIFAR-10 are plotted as black dashed lines). We test how performance varies with fixed hyperparameters (tuned for best performance in the 100-client setting) for different system scales. We conclude that the performance gap of FedADMM increases with the client population.

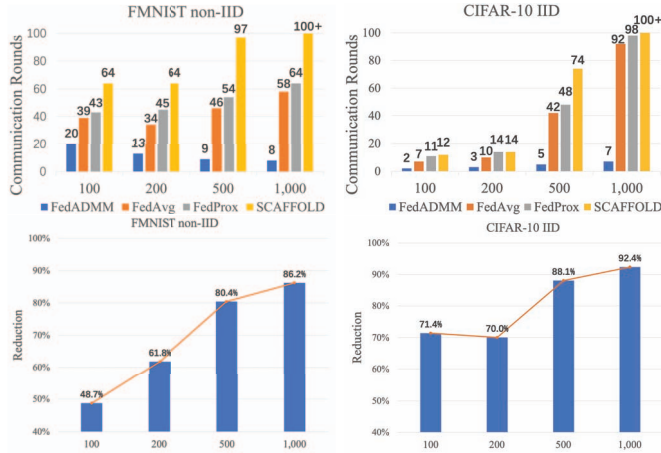


Fig. 4: Number of rounds to reach a prescribed accuracy for variable client population along with the reduction of FedADMM over the best baseline method.

for FedProx except for Table III, where ρ is selected from the candidate set $\{0.001, 0.01, 0.1, 1\}$ to test for its best performance. In contrast, FedADMM outperforms all baselines with fixed $\rho = 0.01$ (empirically fixed) in all experiments. This is in full alignment with our theoretical analysis (see **Theorem**

1 and **Remark 1**) that suggests that ρ can be set as constant, independent of system size and data distributions.

B. Experimental Results

We first demonstrate the communication efficiency of FedADMM by comparing the number of rounds needed to reach a prescribed target accuracy. The results accompanied by a detailed description are summarized in Table III. We note that FedADMM consistently outperforms all baseline methods in all cases tested. In specific, the communication savings achieved by FedADMM compared to the best baseline method in each case averages 72%. Recall also that FedADMM has 50% less training computation than FedAvg and SCAFFOLD in all instances, since those were tested without system heterogeneity considerations, unlike FedProx and FedADMM.

Increasing System Scale. We compare the performance of FedADMM in different system scales using the FMNIST and CIFAR-10 datasets with both IID and non-IID data distributions. We first tuned all algorithms for their best performance in the 100-client setting, and then scale up system sizes with the hyperparameters fixed. The results are illustrated in Fig. 3. Fig. 4 serves to complement Fig. 3: it considers the same datasets but for the reversed setting (IID becomes non-IID and vice versa). We keep the same fraction ($C = 0.1$) of

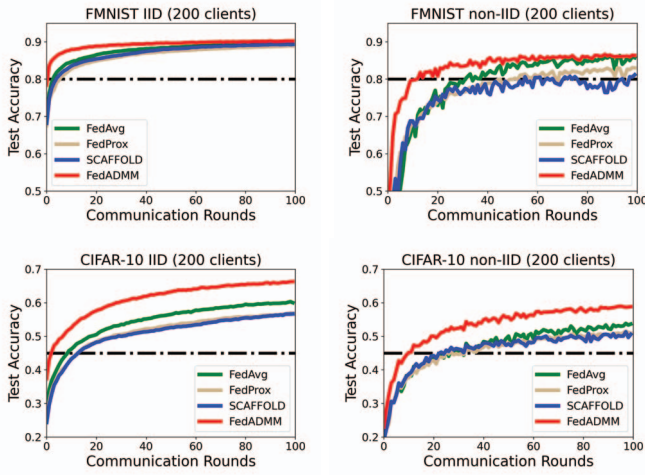


Fig. 5: Comparison of adaptability to heterogeneous data. Target accuracy is 80% for FMNIST and 45% for CIFAR-10. For both cases, $m = 200$, $E = 10$, $B = 50$. In FedADMM, we fix the local learning rates to 0.1 and $\rho = 0.01$, while other algorithms are tuned for best performance. In all scenarios, FedADMM consistently outperforms all baseline methods even with no hyperparameter tuning.

participating clients at each round, which translates to the same amount of data being used per round. Note, however, that an increase of the client population comes with the introduction of additional dual variables for FedADMM. In other words, the same amount of data is processed with more guidance, which serves to yield a larger improvement of FedADMM over baseline methods at larger scales. This favorable attribute is intensified in the non-IID setting where FedADMM is shown to achieve an effective incorporation of local information even faster. We conclude that FedADMM has formidable scalability, as also suggested by **Theorem 1**. In summary, the performance boost at increased scale (especially in the non-IID setting) is attributed to: (i) a higher aggregate number of information exchanges with the server, (ii) a higher total computational power in the network (albeit both at no additional cost per user), and (iii) additional dual variables, which provide extra guidance to the learning process.

Data Heterogeneity. To demonstrate the adaptability of FedADMM to heterogeneous data distributions, we fix parameters of FedADMM while tuning other methods for their best performance. Fig. 5 reveals the robust performance of FedADMM to statistical variations (with no hyperparameter tuning). This is a positive attribute for FL applications where data distributions are unpredictable and may be hard to characterize due to strict restrictions on clients' data privacy. As discussed in Section III, the dual variables in FedADMM serve as an automatic adaptation mechanism to data heterogeneity, by quantitatively recording the discrepancy between the local model and the global model.

Server step size. We explore the effect of adjusting the server gathering step size η in a system with 100 clients. In the left figure of the first row of Fig. 6, we observe that when data distributions are IID across clients, all choices of η result in

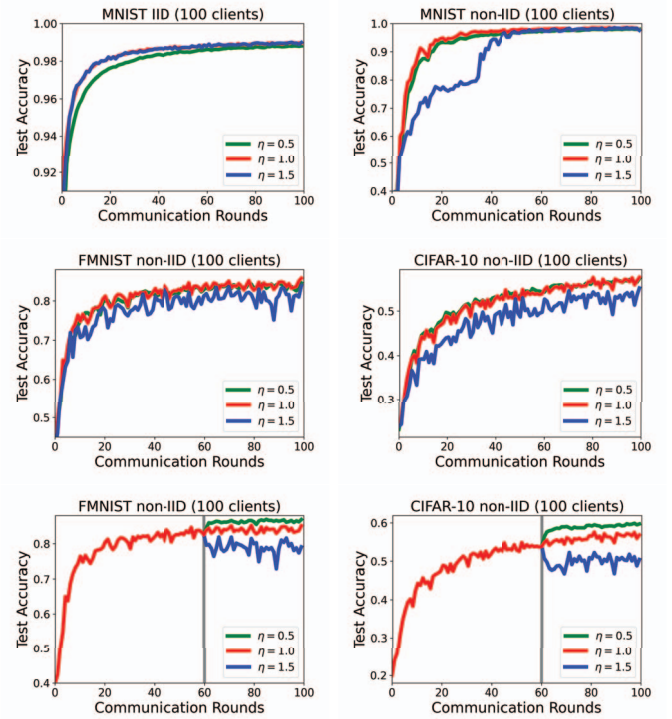


Fig. 6: Performance comparison using different server step sizes for FedADMM. The vertical line indicates the point of when the step size is adjusted. The nominal step size $\eta = 1.0$ yields consistent performance, while a decrease of the step size at later stages of the process can further improve the algorithm.

similar performance, with the smallest $\eta = 0.5$ being slightly slower at initial stages. This is due to the fact that when data are evenly distributed among clients, more drastic updates are allowed. When data distributions are non-IID, care must be taken in incorporating local information to the global model. This is reflected by the stalling in the case of setting $\eta = 1.5$. The nominal value $\eta = 1.0$ showcases consistent performance in all cases. We additionally tested the effect of adjusting η at later stages of the process (60 rounds): a decrease of the step size serves to incorporate past information in a finer fashion, thus improving the test accuracy.

Increasing Local Work. We investigate how the local epoch number E (capturing the amount of local training) affects the performance of FedADMM. Table IV and Fig. 7 indicate that a more accurate global model can be obtained by increasing the amount of local training. This is in accord with our theoretical analysis (**Theorem 1**): the fact that the local subproblems are

TABLE IV: Number of rounds required for FedADMM to reach given accuracy (97% for MNIST and 45% for CIFAR-10) with variable amount of local training.

		$E = 1$	$E = 5$	$E = 10$
MNIST	IID	27	10	6
	non-IID	56	33	32
CIFAR-10	IID	24	12	10
	non-IID	30	14	11

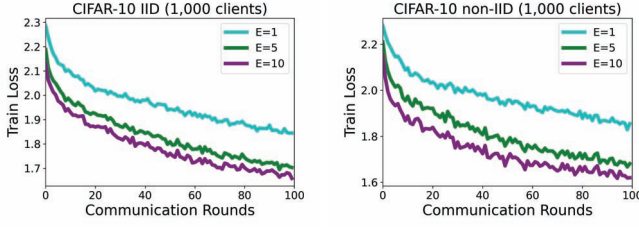


Fig. 7: Performance boost with increased local epoch number E . Convergence of FedADMM is always ensured even with fixed learning rate and increasing amount of local training.

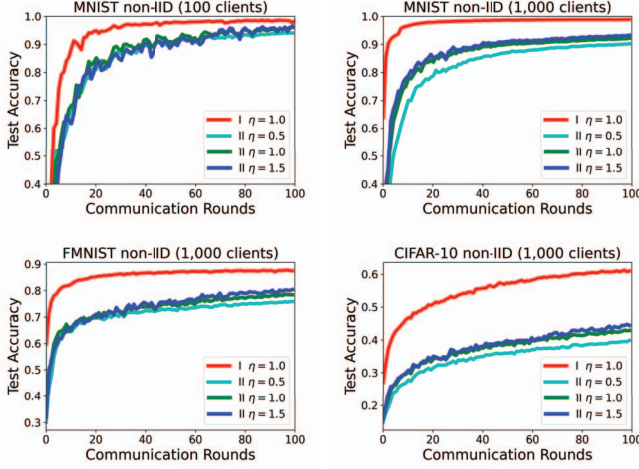


Fig. 8: Different choices on local training initialization for FedADMM: I stands for initializing with the local model w_i^t , and II stands for initializing with the global model θ^t (over different server step sizes). Warm starting local training with the local model w_i^t yields superior results in all cases.

cast as strongly convex guarantees that the larger the local workload, the smaller the ε_i in (6), and thus the better the convergence.

Local Initialization. We further study the performance of FedADMM by investigating the impact of different initialization for the local training subproblems at selected clients. As shown in Fig. 8, using client model parameters w_i^t as a warm start for

TABLE V: Number of communication rounds to reach a prescribed accuracy (80% for FMNIST, 97% for MNIST). We use fixed $\rho = 0.01$ for FedADMM while varying ρ for FedProx. Note that different values of ρ give drastically different results for FedProx, while FedADMM is consistently superior with constant ρ .

Dataset	ρ	200 clients		500 clients	
		IID	non-IID	IID	non-IID
MNIST	FedADMM (0.01)	3	29	5	14
	FedProx (0.01)	100+	100+	47	82
	FedProx (0.1)	25	83	59	100+
	FedProx (1)	81	100+	100+	100+
FMNIST	FedADMM (0.01)	2	13	2	9
	FedProx (0.01)	5	34	11	45
	FedProx (0.1)	7	45	13	54
	FedProx (1)	43	100+	44	100+

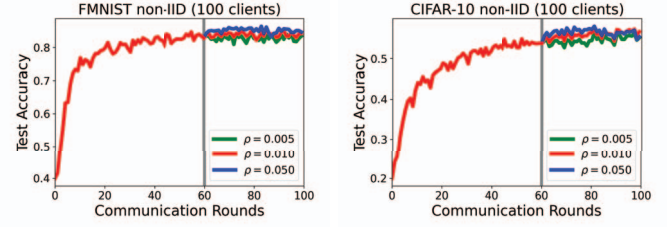


Fig. 9: Performance comparison with different value of proximal hyperparameter ρ in FedADMM. The vertical line indicates the point when ρ is changed.

the local SGD demonstrates significant advantages compared to using the global model θ^t . This provides additional motivation for clients to store their local model w_i as compared to using the global model θ as the starting point for local training.

Proximal Parameter ρ . To counter *client drift* in non-IID data distributions, FedADMM and FedProx both use a quadratic proximal term for the local training problem. The proximal coefficient ρ in FedProx has to be carefully tuned to achieve competitive performance across different settings [8], [9]. Such tuning is dependent on system sizes and data distributions, as we also demonstrate in Table V. Note that the best ρ value (0.01) for FedProx in the FMNIST dataset gives the worst performance for MNIST in the 200-client setting. Moreover, the performance of FedProx with respect to ρ is not monotone, which makes tuning even more challenging. On the other hand, FedADMM dominates all tested instances of FedProx with fixed $\rho = 0.01$. This is also supported by our theoretical analysis (**Theorem 1** and **Remark 1** supports a constant ρ). Additional insights can be gained by a simple dynamic adaptation of ρ for FedADMM in Fig. 9. A smaller value (0.01) at initial stages of training allows efficient incorporation of local data when the global model is not informed, while an increase of ρ at later stages reduces discrepancies between client models and the global model.

Imbalanced Data Volumes. We explore the more realistic scenario that clients hold different data volumes. In this setting,

TABLE VI: Statistics of imbalanced datasets.

Dataset	Clients	Samples	Mean	Stdev
FMNIST	200	60,000	300	171.03
CIFAR-10	200	50,000	250	142.52

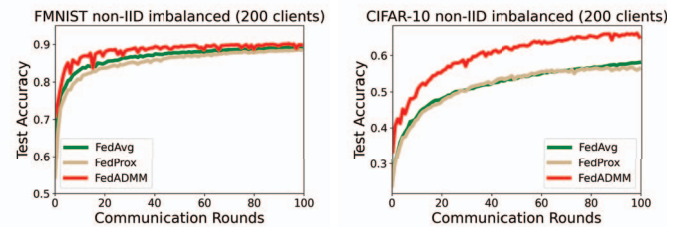


Fig. 10: FedADMM achieves the best performance over baseline algorithms for imbalanced datasets ($E = 10$, $B = 50$).

we first sort the training data points by labels, then divide the training data into 10,000 shards, each with 5 data points for CIFAR-10 and 6 data points for FMNIST, respectively. We divide 200 clients evenly into 100 groups. Each member of the group is assigned with a number of shards that equals the group index, except for the last group that collects the remaining data. The setting is summarized in Table VI. Fig. 10 shows that FedADMM achieves higher accuracy than other baselines, especially in the CIFAR-10 dataset.

VI. CONCLUSION

FedADMM is a new federated learning method that handles both statistical and system heterogeneity, without inducing extra communication costs per round. By storing an extra dual variable at each client, a significant speedup over the state-of-the-art is achieved which translates to substantial communication savings. Moreover, FedADMM features automatic adaptation to statistical variations within the system without the need for hyperparameter tuning. Such adaptation safeguards against client drift that causes FedAvg to diverge. We have established an optimal convergence rate (which matches the problem complexity lower bound), with no assumptions on data dissimilarity or gradient boundedness. The analytical results are corroborated by extensive experiments that reveal that FedADMM is well-suited for FL applications with large system sizes and heterogeneous data distributions.

VII. FULL PROOF

In this section, we present the full proof of our theoretical results. Our proof of convergence is inspired by [27] but differs in the following aspects allusive to our specific design: (i) the server updates at each round; (ii) the local training problems are solved inexactly with the level of inexactness captured by the local parameter ε_i as in (6); (iii) an additional step size η for server aggregation is included. For ease of exposition, we fix $\eta = |S^t|/m$ in the analysis and stack variables as follows: $w^t = [(w_1^t)^\top, \dots, (w_m^t)^\top]^\top$, $y^t = [(y_1^t)^\top, \dots, (y_m^t)^\top]^\top \in \mathbb{R}^{md}$. We define the *virtual* dual update as follows:

$$\|\nabla_{w_i} \mathcal{L}_i(\hat{w}_i^{t+1}, y_i^t, \theta^t)\|^2 \leq \varepsilon_i, \quad (11)$$

$$\hat{y}_i^{t+1} = y_i^t + \rho(\hat{w}_i^{t+1} - \theta^t). \quad (12)$$

In other words, the virtual update for client i is the value for (w_i^{t+1}, y_i^{t+1}) if it participates in the current round, i.e., if $i \in S^t$, $w_i^{t+1} = \hat{w}_i^{t+1}$ and $y_i^{t+1} = \hat{y}_i^{t+1}$. As a consequence, we have:

$$\mathbb{E}^t[w_i^{t+1} - w_i^t] = p_i(\hat{w}_i^{t+1} - w_i^t),$$

where $\mathbb{E}^t[\cdot]$ denotes the conditional expectation at step t , and p_i denotes the probability of client i being active. We proceed to bound $\|\hat{y}_i^{t+1} - y_i^t\|$ in the following lemma.

Lemma 1: Recall assumption 1 and the local accuracy ε_i in (6). The consecutive difference between dual variables can be bounded as follows:

$$\|\hat{y}_i^{t+1} - y_i^t\|^2 \leq 8\varepsilon_i + 2L^2\|\hat{w}_i^{t+1} - w_i^t\|^2, \quad (13)$$

$$\|y_i^{t+1} - y_i^t\|^2 \leq 8\varepsilon_i + 2L^2\|w_i^{t+1} - w_i^t\|^2. \quad (14)$$

Proof: We first prove (13) by defining the local error term as

$$\begin{aligned} e_i^{t+1} &= \nabla_{w_i} \mathcal{L}_i(\hat{w}_i^{t+1}, y_i^t, \theta^t) \\ &= \nabla f_i(\hat{w}_i^{t+1}) + y_i^t + \rho(\hat{w}_i^{t+1} - \theta^t) \\ &= \nabla f_i(\hat{w}_i^{t+1}) + \hat{y}_i^{t+1}, \end{aligned} \quad (15)$$

where the last equality holds from the definition of virtual update rule (12). Note that $\|e_i^{t+1}\| = \|\nabla_{w_i} \mathcal{L}_i(\hat{w}_i^{t+1}, y_i^t, \theta^t)\| \leq \sqrt{\varepsilon_i}$. By rearranging and taking the difference,

$$\begin{aligned} \|\hat{y}_i^{t+1} - y_i^t\| &= \|e_i^{t+1} - e_i^t + \nabla f_i(w_i^t) - \nabla f_i(\hat{w}_i^{t+1})\| \\ &\leq \|e_i^{t+1}\| + \|e_i^t\| + \|\nabla f_i(w_i^t) - \nabla f_i(\hat{w}_i^{t+1})\| \\ &\leq 2\sqrt{\varepsilon_i} + L\|\hat{w}_i^{t+1} - w_i^t\|. \end{aligned}$$

Using the inequality $\|\sum_{i=1}^m a_i\|^2 \leq m \sum_{i=1}^m \|a_i\|^2$, we obtain:

$$\|\hat{y}_i^{t+1} - y_i^t\|^2 \leq 8\varepsilon_i + 2L^2\|\hat{w}_i^{t+1} - w_i^t\|^2,$$

which is the desired. We note that (13) holds for all $i \in [m]$ and proceed to show (14) as follows. For $i \in S^t$, $y_i^{t+1} = \hat{y}_i^{t+1}$, $w_i^{t+1} = \hat{w}_i^{t+1}$, and therefore (14) reduces to (13). For $i \notin S^t$, $w_i^{t+1} = w_i^t$, $y_i^{t+1} = y_i^t$, and therefore (14) trivially holds. ■

We denote the aggregated Lagrangian as $\mathcal{L}(w^t, y^t, \theta^t) = \sum_{i=1}^m \mathcal{L}_i(w_i^t, y_i^t, \theta^t)$ and proceed to bound its change after one round.

Lemma 2: The aggregated Lagrangian $\mathcal{L}^t \equiv \mathcal{L}(w^t, y^t, \theta^t)$ iterates satisfy:

$$\begin{aligned} \mathcal{L}^{t+1} - \mathcal{L}^t &\leq \sum_{i \in S^t} \left(\left(\frac{2L-\rho}{2} + \frac{2L^2}{\rho} \right) \|w_i^{t+1} - w_i^t\|^2 \right. \\ &\quad \left. + \frac{\varepsilon_i}{2L} + \frac{8\varepsilon_i}{\rho} \right) - \frac{m\rho}{2} \|\theta^{t+1} - \theta^t\|^2. \end{aligned} \quad (16)$$

Proof: We decompose the difference into three parts:

$$\begin{aligned} \mathcal{L}^{t+1} - \mathcal{L}^t &= \mathcal{L}(w^{t+1}, y^{t+1}, \theta^{t+1}) - \mathcal{L}(w^{t+1}, y^{t+1}, \theta^t) \\ &\quad + \mathcal{L}(w^{t+1}, y^{t+1}, \theta^t) - \mathcal{L}(w^{t+1}, y^t, \theta^t) \\ &\quad + \mathcal{L}(w^{t+1}, y^t, \theta^t) - \mathcal{L}(w^t, y^t, \theta^t). \end{aligned} \quad (17)$$

The three difference terms in (17) correspond to the θ^{t+1} -update, y^{t+1} -update, and w^{t+1} -update, respectively. We proceed to bound each term separately. Note that $\mathcal{L}(\cdot)$ is strongly convex with respect to θ with parameter $m\rho$. Therefore,

$$\begin{aligned} \mathcal{L}(w^{t+1}, y^{t+1}, \theta^t) &\geq \mathcal{L}^{t+1} + \langle \nabla_{\theta} \mathcal{L}^{t+1}, \theta^t - \theta^{t+1} \rangle \\ &\quad + \frac{m\rho}{2} \|\theta^{t+1} - \theta^t\|^2. \end{aligned} \quad (18)$$

By definition,

$$\nabla_{\theta} \mathcal{L}^{t+1} = \rho \left(m\theta^{t+1} - \sum_{i=1}^m (w_i^{t+1} + \frac{1}{\rho} y_i^{t+1}) \right). \quad (19)$$

When the server step size is chosen as $\eta = |S^t|/m$, we have

$$\begin{aligned} \theta^{t+1} &= \theta^t + \frac{1}{m} \sum_{i \in S^t} \left(w_i^{t+1} + \frac{1}{\rho} y_i^{t+1} - (w_i^t + \frac{1}{\rho} y_i^t) \right) \\ &= \theta^t + \frac{1}{m} \sum_{i=1}^m (u_i^{t+1} - u_i^t), \end{aligned}$$

where we have defined the *augmented model* $u_i^t = w_i^t + \frac{1}{\rho}y_i^t$ and used the fact that for $i \notin S^t$, $u_i^{t+1} - u_i^t = 0$. After telescoping, we obtain

$$\theta^{t+1} = \theta^0 + \frac{1}{m} \sum_{i=1}^m (u_i^{t+1} - u_i^0).$$

From our initialization, $w_i^0 = \theta^0$ and $y_i^0 = 0$, it follows that $\frac{1}{m} \sum_{i=1}^m u_i^0 = \theta^0$. Therefore, $\theta^{t+1} = \frac{1}{m} \sum_{i=1}^m u_i^{t+1}$. Using this fact along with (19), we obtain

$$\nabla_{\theta} \mathcal{L}^{t+1} = \rho \sum_{i=1}^m \left(u_i^{t+1} - (w_i^{t+1} + \frac{1}{\rho}y_i^{t+1}) \right) = 0, \quad (20)$$

where we used the definition $u_i^{t+1} = w_i^{t+1} + \frac{1}{\rho}y_i^{t+1}$. Therefore, we can rewrite (18) as

$$\mathcal{L}^{t+1} - \mathcal{L}(w^{t+1}, y^{t+1}, \theta^t) \leq -\frac{m\rho}{2} \|\theta^{t+1} - \theta^t\|^2, \quad (21)$$

which is the bound for the first difference term in (17). By definition, the second difference term in (17) can be expressed as:

$$\begin{aligned} & \mathcal{L}(w^{t+1}, y^{t+1}, \theta^t) - \mathcal{L}(w^{t+1}, y^t, \theta^t) \\ &= \sum_{i \in S^t} \langle y_i^{t+1} - y_i^t, w_i^{t+1} - \theta^t \rangle \\ &= \sum_{i \in S^t} \frac{1}{\rho} \|y_i^{t+1} - y_i^t\|^2, \end{aligned}$$

where the last equality follows from (12) and the fact that for $i \in S^t$, $y_i^{t+1} = \hat{y}_i^{t+1}$, $w_i^{t+1} = \hat{w}_i^{t+1}$. Using Lemma 1, we obtain

$$\begin{aligned} & \mathcal{L}(w^{t+1}, y^{t+1}, \theta^t) - \mathcal{L}(w^{t+1}, y^t, \theta^t) \\ & \leq \sum_{i \in S^t} \left(\frac{8\varepsilon_i}{\rho} + \frac{2L^2}{\rho} \|w_i^{t+1} - w_i^t\|^2 \right), \end{aligned} \quad (22)$$

which is the bound for the second difference term in (17). For the third difference term, we first note that $-\nabla f_i(\cdot)$ is Lipschitz as well. Therefore, the following holds:

$$\begin{aligned} -f_i(w_i^t) & \leq -f_i(w_i^{t+1}) + \langle -\nabla f_i(w_i^{t+1}), w_i^t - w_i^{t+1} \rangle \\ & \quad + \frac{L}{2} \|w_i^{t+1} - w_i^t\|^2, \end{aligned} \quad (23)$$

and $\forall i \in S^t$:

$$\begin{aligned} & \mathcal{L}_i(w_i^{t+1}, y_i^t, \theta^t) - \mathcal{L}_i(w_i^t, y_i^t, \theta^t) \\ &= f_i(w_i^{t+1}) - f_i(w_i^t) + \langle y_i^t, w_i^{t+1} - w_i^t \rangle \\ & \quad + \frac{\rho}{2} (\|w_i^{t+1} - \theta^t\|^2 - \|w_i^t - \theta^t\|^2) \\ & \leq \langle \nabla f_i(w_i^{t+1}), w_i^{t+1} - w_i^t \rangle + \frac{L}{2} \|w_i^{t+1} - w_i^t\|^2 \\ & \quad + \langle y_i^t, w_i^{t+1} - w_i^t \rangle + \frac{\rho}{2} \langle w_i^{t+1} + w_i^t - 2\theta^t, w_i^{t+1} - w_i^t \rangle \\ & \stackrel{(i)}{=} \langle \nabla f_i(w_i^{t+1}) + y_i^t + \frac{\rho}{2} \cdot 2(w_i^{t+1} - \theta^t), w_i^{t+1} - w_i^t \rangle \\ & \quad + \frac{L}{2} \|w_i^{t+1} - w_i^t\|^2 + \frac{\rho}{2} \langle w_i^t - w_i^{t+1}, w_i^{t+1} - w_i^t \rangle \\ & \stackrel{(iii)}{\leq} \frac{1}{2L} \|\nabla f_i(w_i^{t+1}) + y_i^{t+1}\|^2 + \frac{2L-\rho}{2} \|w_i^{t+1} - w_i^t\|^2 \\ & \stackrel{(iv)}{\leq} \frac{\varepsilon_i}{2L} + \frac{2L-\rho}{2} \|w_i^{t+1} - w_i^t\|^2. \end{aligned} \quad (24)$$

In (24): (i) follows from (23) and the identity $\|a\|^2 - \|b\|^2 = (a+b)^\top(a-b)$; (ii) follows from splitting $w_i^{t+1} + w_i^t - 2\theta^t = 2(w_i^{t+1} - \theta^t) + w_i^t - w_i^{t+1}$; (iii) follows from the dual update: $y_i^{t+1} = y_i^t + \rho(w_i^{t+1} - \theta^t)$ and the identity $a^\top b \leq \frac{1}{2L}\|a\|^2 + \frac{L}{2}\|b\|^2$; (iv) follows from (15) and (11). Therefore, the third difference term in (17) can be bounded as:

$$\mathcal{L}(w^{t+1}, y^t, \theta^t) - \mathcal{L}^t \leq \sum_{i \in S^t} \frac{2L-\rho}{2} \|w_i^{t+1} - w_i^t\|^2 + \sum_{i \in S^t} \frac{\varepsilon_i}{2L}. \quad (25)$$

After combining (21), (22) and (25), we obtain the desired. ■

We proceed to establish a lower bound for the augmented Lagrangian in the following lemma.

Lemma 3. Recall the lower bound $\sum_{i=1}^m f_i(w) \geq f^*$ in assumption 2. For $t \geq 0$, the augmented Lagrangian is lower bounded as $\mathcal{L}^{t+1} \geq f^* - \frac{1}{2L} \sum_{i=1}^m \varepsilon_i$ by selecting $\rho \geq 2L$.

Proof: By definition,

$$\begin{aligned} \mathcal{L}^{t+1} &= \sum_{i=1}^m \left(f_i(w_i^{t+1}) + \langle y_i^{t+1}, w_i^{t+1} - \theta^{t+1} \rangle \right. \\ & \quad \left. + \frac{\rho}{2} \|w_i^{t+1} - \theta^{t+1}\|^2 \right) \\ & \stackrel{(i)}{=} \sum_{i=1}^m \left(f_i(w_i^{t+1}) + \langle \nabla f_i(w_i^{t+1}), \theta^{t+1} - w_i^{t+1} \rangle \right. \\ & \quad \left. + \frac{\rho}{2} \|w_i^{t+1} - \theta^{t+1}\|^2 + \langle e_i^{t+1}, w_i^{t+1} - \theta^{t+1} \rangle \right) \\ & \stackrel{(ii)}{\geq} \sum_{i=1}^m \left(f_i(\theta^{t+1}) + \frac{\rho-2L}{2} \|w_i^{t+1} - \theta^{t+1}\|^2 - \frac{1}{2L} \varepsilon_i \right) \\ & \geq \sum_{i=1}^m f_i(\theta^{t+1}) - \frac{1}{2L} \sum_{i=1}^m \varepsilon_i \geq f^* - \frac{1}{2L} \sum_{i=1}^m \varepsilon_i. \end{aligned} \quad (26)$$

To show equality (i) holds, it suffices to show: $\forall i, y_i^{t+1} = e_i^{t+1} - \nabla f_i(w_i^{t+1})$. This holds for $i \in S^t$ by (15) and the fact that $\hat{w}_i^{t+1} = w_i^{t+1}$, $\hat{y}_i^{t+1} = y_i^{t+1}$. For $i \notin S^t$, $w_i^{t+1} = w_i^t = w_i^{t_i+1}$, where we denote t_i as the most recent time step when client i is selected before step t . Therefore, for $i \notin S^t$, $y_i^{t+1} = y_i^t = y_i^{t_i+1} = e_i^{t_i+1} - \nabla f_i(w_i^{t_i+1}) = e_i^{t+1} - \nabla f_i(w_i^{t+1})$. The inequality (ii) follows from a consequence of assumption 1, the identity $a^\top b \geq -\frac{L}{2}\|a\|^2 - \frac{1}{2L}\|b\|^2$, and $\|e_i^{t+1}\| \leq \varepsilon_i$. ■

Recall the non-negative function $V^t \equiv V(w^t, y^t, \theta^t)$ defined in (7):

$$V^t = \|\nabla_{\theta} \mathcal{L}^t\|^2 + \sum_{i=1}^m (\|\nabla_{w_i} \mathcal{L}^t\|^2 + \|w_i^t - \theta^t\|^2).$$

We establish the convergence of the FedADMM in the following.

Theorem 1. Let assumptions 1 and 2 hold. Assume each client has a probability of participating at a given round that is lower bounded by a positive constant $p_{\min} > 0$. Consider $\eta = |S^t|/m$, and select $\rho > (1 + \sqrt{5})L$, then the following holds:

$$\frac{1}{mT} \sum_{t=0}^{T-1} \mathbb{E}[V^t] \leq \frac{1}{mT} \frac{c_2}{c_1} (\mathcal{L}^0 - f^* + \frac{m}{2L} \varepsilon_{\max}) + c_3 \varepsilon_{\max}, \quad (8)$$

where $\varepsilon_{\max} = \max_i \varepsilon_i$, $c_1 = p_{\min} \left(\frac{\rho-2L}{2} - \frac{2L^2}{\rho} \right)$, $c_2 = 3(L^2 + \rho^2) + 2(1 + \frac{2L^2}{\rho^2})$, and $c_3 = 3 + \frac{16}{\rho^2} + \frac{c_2}{c_1} \cdot \frac{\rho+16L}{2L\rho}$.

Proof: By definition, the following holds:

$$\|\nabla_{w_i} \mathcal{L}_i^t\|^2 = \|\nabla f_i(w_i^t) + y_i^t + \rho(w_i^t - \theta^t)\|^2.$$

Recall (15) where

$$e_i^{t+1} = \nabla f_i(\hat{w}_i^{t+1}) + y_i^t + \rho(\hat{w}_i^{t+1} - \theta^t).$$

Therefore, we obtain

$$\begin{aligned} \|\nabla_{w_i} \mathcal{L}_i^t\| &= \|\nabla f_i(w_i^t) + y_i^t + \rho(w_i^t - \theta^t)\| \\ &= \left\| \nabla f_i(w_i^t) + y_i^t + \rho(w_i^t - \theta^t) \right. \\ &\quad \left. - \left(\nabla f_i(\hat{w}_i^{t+1}) + y_i^t + \rho(\hat{w}_i^{t+1} - \theta^t) \right) + e_i^{t+1} \right\| \\ &= \|\nabla f_i(w_i^t) - \nabla f_i(\hat{w}_i^{t+1}) + \rho(w_i^t - \hat{w}_i^{t+1}) + e_i^{t+1}\| \\ &\leq \|\nabla f_i(w_i^t) - \nabla f_i(\hat{w}_i^{t+1})\| + \rho\|w_i^t - \hat{w}_i^{t+1}\| + \|e_i^{t+1}\| \end{aligned}$$

Using assumption 1, and the accuracy for inexact primal updates, we obtain

$$\|\nabla_{w_i} \mathcal{L}_i^t\|^2 \leq 3 \left((L^2 + \rho^2) \|\hat{w}_i^{t+1} - w_i^t\|^2 + \varepsilon_i \right).$$

Therefore, it holds that

$$\sum_{i=1}^m \|\nabla_{w_i} \mathcal{L}_i^t\|^2 \leq 3 \sum_{i=1}^m \left((L^2 + \rho^2) \|\hat{w}_i^{t+1} - w_i^t\|^2 + \varepsilon_i \right). \quad (27)$$

Moreover,

$$\begin{aligned} \sum_{i=1}^m \|w_i^t - \theta^t\|^2 &= \sum_{i=1}^m \|w_i^t - \hat{w}_i^{t+1} + \hat{w}_i^{t+1} - \theta^t\|^2 \\ &\leq \sum_{i=1}^m 2 \left(\|\hat{w}_i^{t+1} - w_i^t\|^2 + \|\hat{w}_i^{t+1} - \theta^t\|^2 \right). \end{aligned}$$

From the virtual update for \hat{y}_i^{t+1} in (12),

$$\|\hat{w}_i^{t+1} - \theta^t\|^2 = \frac{1}{\rho^2} \|\hat{y}_i^{t+1} - y_i^t\|^2. \quad (28)$$

Therefore, we obtain:

$$\begin{aligned} \sum_{i=1}^m \|w_i^t - \theta^t\|^2 &\leq \sum_{i=1}^m 2 \left(\|\hat{w}_i^{t+1} - w_i^t\|^2 + \frac{1}{\rho^2} \|\hat{y}_i^{t+1} - y_i^t\|^2 \right) \\ &\leq 2 \sum_{i=1}^m \left(\left(1 + \frac{2L^2}{\rho^2} \right) \|\hat{w}_i^{t+1} - w_i^t\|^2 + \frac{8\varepsilon_i}{\rho^2} \right), \quad (29) \end{aligned}$$

where the last inequality follows from Lemma 1. From (20) we obtain:

$$\begin{aligned} V^t &= \underbrace{\|\nabla_{\theta} \mathcal{L}^t\|^2}_{=0} + \sum_{i=1}^m \left(\|\nabla_{w_i} \mathcal{L}_i^t\|^2 + \|w_i^t - \theta^t\|^2 \right) \\ &\stackrel{(27)-(29)}{\leq} \sum_{i=1}^m \left(3(L^2 + \rho^2) + 2(1 + \frac{2L^2}{\rho^2}) \right) \|\hat{w}_i^{t+1} - w_i^t\|^2 \\ &\quad + \sum_{i=1}^m \left(3 + \frac{16}{\rho^2} \right) \varepsilon_i \\ &\leq \sum_{i=1}^m c_2 \|\hat{w}_i^{t+1} - w_i^t\|^2 + \left(3m + \frac{16m}{\rho^2} \right) \varepsilon_{\max}, \quad (30) \end{aligned}$$

where the last inequality follows from the definitions: $\varepsilon_{\max} = \max_i \varepsilon_i$, $c_2 = 3(L^2 + \rho^2) + 2(1 + \frac{2L^2}{\rho^2})$. From Lemma 2, we have

$\mathcal{L}^{t+1} - \mathcal{L}^t \leq \sum_{i \in S^t} \left(\left(\frac{2L-\rho}{2} + \frac{2L^2}{\rho} \right) \|w_i^{t+1} - w_i^t\|^2 + \frac{\varepsilon_i}{2L} + \frac{8\varepsilon_i}{\rho} \right) - \frac{m\rho}{2} \|\theta^{t+1} - \theta^t\|^2$. After taking conditional expectation at step t , by using linearity of expectation and rearranging, we obtain:

$$\begin{aligned} \mathbb{E}^t [\mathcal{L}^t - \mathcal{L}^{t+1}] &\geq p_{\min} \sum_{i=1}^m \left(\frac{\rho-2L}{2} - \frac{2L^2}{\rho} \right) \|\hat{w}_i^{t+1} - w_i^t\|^2 \\ &\quad - m \left(\frac{\varepsilon_{\max}}{2L} + \frac{8\varepsilon_{\max}}{\rho} \right) + \frac{m\rho}{2} \|\bar{\theta}^{t+1} - \theta^t\|^2, \quad (31) \end{aligned}$$

where we denote $\mathbb{E}^t [\theta^{t+1}] = \bar{\theta}^{t+1}$ and used Jensen's inequality: $\mathbb{E}^t [\|\theta^{t+1} - \theta^t\|^2] \geq \|\mathbb{E}^t [\theta^{t+1}] - \theta^t\|^2 = \|\bar{\theta}^{t+1} - \theta^t\|^2$.

Defining $c_1 = p_{\min} \left(\frac{\rho-2L}{2} - \frac{2L^2}{\rho} \right)$, we rewrite (31) as

$$\mathbb{E}^t [\mathcal{L}^t - \mathcal{L}^{t+1}] + m \left(\frac{\varepsilon_{\max}}{2L} + \frac{8\varepsilon_{\max}}{\rho} \right) \geq c_1 \sum_{i=1}^m \|\hat{w}_i^{t+1} - w_i^t\|^2. \quad (32)$$

Using (30) and (32) leads to:

$$V^t \leq \frac{c_2}{c_1} \mathbb{E}^t [\mathcal{L}^t - \mathcal{L}^{t+1}] + mc_3 \varepsilon_{\max}, \quad (33)$$

where $c_3 = 3 + \frac{16}{\rho^2} + \frac{c_2}{c_1} \left(\frac{1}{2L} + \frac{8}{\rho} \right)$. Taking expectation on both sides gives

$$\mathbb{E}[V^t] \leq \frac{c_2}{c_1} \mathbb{E}[\mathcal{L}^t - \mathcal{L}^{t+1}] + mc_3 \varepsilon_{\max}.$$

After telescoping the above and dividing by mT , we obtain:

$$\frac{1}{mT} \sum_{t=0}^{T-1} \mathbb{E}[V^t] \leq \frac{1}{mT} \frac{c_2}{c_1} \mathbb{E}[\mathcal{L}^0 - \mathcal{L}^{T-1}] + c_3 \varepsilon_{\max}.$$

From Lemma 3, it holds that $\mathbb{E}[\mathcal{L}^0 - \mathcal{L}^{T-1}] \leq \mathcal{L}^0 - f^* + \frac{m}{2L} \varepsilon_{\max}$. Therefore, the following holds:

$$\frac{1}{mT} \sum_{t=0}^{T-1} \mathbb{E}[V^t] \leq \frac{1}{mT} \frac{c_2}{c_1} (\mathcal{L}^0 - f^* + \frac{m}{2L} \varepsilon_{\max}) + c_3 \varepsilon_{\max},$$

which is the desired. \blacksquare

REFERENCES

- [1] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *COMPSTAT*, pp. 177–186, 2010.
- [2] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. a. Ranzato, A. Senior, P. Tucker, K. Yang, Q. Le, and A. Ng, "Large scale distributed deep networks," in *NeurIPS*, pp. 1223–1231, 2012.
- [3] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv:1610.02527*, 2016.
- [4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *AISTATS*, pp. 1273–1282, 2017.
- [5] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-iid data," in *ICLR*, 2020.
- [6] F. Haddadpour and M. Mahdavi, "On the convergence of local descent methods in federated learning," *arXiv:1910.14425*, 2019.
- [7] A. Khaled, K. Mishchenko, and P. Richtárik, "Tighter theory for local SGD on identical and heterogeneous data," in *AISTATS*, pp. 4519–4529, 2020.
- [8] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Machine Learning and Systems*, pp. 429–450, 2020.
- [9] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *ICML*, pp. 5132–5143, 2020.
- [10] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, 1996.
- [11] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976.
- [12] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, p. 1–122, 2011.
- [13] K. Li, M. Sundin, C. R. Rojas, S. Chatterjee, and M. Jansson, "Alternating strategies with internal ADMM for low-rank matrix reconstruction," *Signal Processing*, vol. 121, pp. 153–159, 2016.
- [14] G. Taylor, R. Burmeister, Z. Xu, B. Singh, A. Patel, and T. Goldstein, "Training neural networks without gradients: A scalable ADMM approach," in *ICML*, pp. 2722–2731, 2016.
- [15] S. El Bsati, H. B. Ammar, and M. E. Taylor, "Scalable multitask policy gradient reinforcement learning," in *AAAI*, pp. 1847–1853, 2017.
- [16] Y. Liu, W. Xu, G. Wu, Z. Tian, and Q. Ling, "Communication-censored ADMM for decentralized consensus optimization," *IEEE Transactions on Signal Processing*, vol. 67, no. 10, pp. 2565–2579, 2019.
- [17] T.-H. Chang, M. Hong, and X. Wang, "Multi-agent distributed optimization via inexact consensus ADMM," *IEEE Transactions on Signal Processing*, vol. 63, no. 2, pp. 482–497, 2015.
- [18] V. Smith, S. Forte, C. Ma, M. Takáč, M. I. Jordan, and M. Jaggi, "CoCoA: A general framework for communication-efficient distributed optimization," *Journal of Machine Learning Research*, vol. 18, no. 230, pp. 1–49, 2018.
- [19] Y. Li, N. M. Freris, P. Voulgaris, and D. Stipanović, "DN-ADMM: Distributed Newton ADMM for multi-agent optimization," in *IEEE Conference on Decision and Control*, pp. 3343–3348, 2021.
- [20] Y. Li, Y. Gong, N. M. Freris, P. Voulgaris, and D. Stipanović, "BFGS-ADMM for large-scale distributed optimization," in *IEEE Conference on Decision and Control*, pp. 1689–1694, 2021.
- [21] X. Niu and E. Wei, "FedHybrid: A hybrid primal-dual algorithm framework for federated optimization," *arXiv:2106.01279*, 2021.
- [22] X. Zhang, M. Hong, S. Dhople, W. Yin, and Y. Liu, "FedPD: A federated learning framework with adaptivity to non-iid data," *IEEE Transactions on Signal Processing*, vol. 69, pp. 6055–6070, 2021.
- [23] S. U. Stich, "Local SGD converges fast and communicates little," in *ICLR*, 2019.
- [24] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-Edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning," *IEEE Network*, vol. 33, no. 5, pp. 156–165, 2019.
- [25] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, B. McMahan, T. Van Overveldt, D. Petrou, D. Ramage, and J. Roselander, "Towards federated learning at scale: System design," in *Machine Learning and Systems*, vol. 1, pp. 374–388, 2019.
- [26] N. H. Tran, W. Bao, A. Zomaya, M. N. H. Nguyen, and C. S. Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *IEEE INFOCOM*, pp. 1387–1395, 2019.
- [27] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM Journal on Optimization*, vol. 26, no. 1, pp. 337–364, 2016.
- [28] R. Zhang and J. Kwok, "Asynchronous distributed ADMM for consensus optimization," in *ICML*, pp. 1701–1709, 2014.
- [29] T.-H. Chang, M. Hong, W.-C. Liao, and X. Wang, "Asynchronous distributed ADMM for large-scale optimization—part i: Algorithm and convergence analysis," *IEEE Transactions on Signal Processing*, vol. 64, no. 12, pp. 3118–3130, 2016.
- [30] Y. Zheng, Y. Song, D. J. Hill, and Y. Zhang, "Multiagent system based microgrid energy management via asynchronous consensus ADMM," *IEEE Transactions on Energy Conversion*, vol. 33, no. 2, pp. 886–888, 2018.
- [31] C. Dwork, "Differential privacy: A survey of results," in *Theory and Applications of Models of Computation*, pp. 1–19, Springer Berlin Heidelberg, 2008.
- [32] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.
- [33] A. Triastcyn and B. Faltings, "Federated learning with Bayesian differential privacy," in *IEEE Big Data*, pp. 2587–2596, 2019.
- [34] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [35] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms," *arXiv:1708.07747*, 2017.
- [36] A. Krizhevsky, "Learning multiple layers of features from tiny images," tech. rep., University of Toronto, 2009.