GradMA: A Gradient-Memory-based Accelerated Federated Learning with Alleviated Catastrophic Forgetting

Kangyang Luo, Xiang Li, Yunshi Lan, Ming Gao
East China Normal University
Shanghai, China
52205901003@stu.ecnu.edu.cn, {xiangli, yslan, mgao}@dase.ecnu.edu.cn

Abstract

Federated Learning (FL) has emerged as a de facto machine learning area and received rapid increasing research interests from the community. However, catastrophic forgetting caused by data heterogeneity and partial participation poses distinctive challenges for FL, which are detrimental to the performance. To tackle the problems, we propose a new FL approach (namely GradMA), which takes inspiration from continual learning to simultaneously correct the server-side and worker-side update directions as well as take full advantage of server’s rich computing and memory resources. Furthermore, we elaborate a memory reduction strategy to enable GradMA to accommodate FL with a large scale of workers. We then analyze convergence of GradMA theoretically under the smooth non-convex setting and show that its convergence rate achieves a linear speed up w.r.t the increasing number of sampled active workers. At last, our extensive experiments on various image classification tasks show that GradMA achieves significant performance gains in accuracy and communication efficiency compared to SOTA baselines. We provide our code here: https://github.com/lkyddd/GradMA.

1. Introduction

Federated Learning (FL) [18, 26] is a privacy-preserving distributed machine learning scheme in which workers jointly participate in the collaborative training of a centralized model by sharing model information (parameters or updates) rather than their private datasets. In recent years, FL has shown its potential to facilitate real-world applications, which falls broadly into two categories [10]: the cross-silo FL and the cross-device FL. The cross-silo FL corresponds to a relatively small number of reliable workers, usually organizations, such as healthcare facilities [9] and financial institutions [41], etc. In contrast, for the cross-device FL, the number of workers can be very huge and unreliable, such as mobile devices [26], IoT [27] and autonomous driving cars [22], among others. In this paper, we focus on cross-device FL.

The privacy-preserving and communication-efficient properties of the cross-device FL make it promising, but it also confronts practical challenges arising from data heterogeneity (i.e., non-iid data distribution across workers) and partial participation [5, 12, 20, 39]. Specifically, the datasets held by real-world workers are generated locally according to their individual circumstances, resulting in the distribution of data on different workers being not identical. Moreover, owing to the flexibility of worker participation in many scenarios (e.g., IoT and mobile devices), workers can join or leave the FL system at will, thus making the set of active workers random and time-varying across communication rounds. Note that we consider a worker participates or is active at round $t$ (i.e., the index of the communication round) if it is able to complete the computation task and send back model information at the end of round $t$.

The above-mentioned challenges mainly bring catastrophic forgetting (CF) [25, 30, 37] to FL. In a typical FL process, represented by FedAvg [26], a server updates the centralized model by iteratively aggregating the model information from workers that generally is trained over several steps locally before being sent to the server. On the one hand, due to data heterogeneity, the model is updated on private data in local training, which is prone to overfit the current knowledge and forget the previous experience, thus leading to CF [8]. In other words, the updates of the local models are prone to drift and diverge increasingly from the update of the centralized model [12]. This can seriously deteriorate the performance of the centralized model. To ameliorate this issue, a variety of existing efforts regularize the objectives of the local models to align the centralized optimization objective [1, 12, 13, 17, 19]. On the other hand, the server can only aggregate model information from active workers per communication round caused by partial participation. In this case, many existing works directly dis-
In this paper, to alleviate CF caused by data heterogeneity and stragglers, we bring forward a new FL approach, dubbed as GradMA (Gradient-Memory-based Accelerated Federated Learning), which takes inspiration from continual learning (CL) [4,14,24,29,44], two memory-based CL methods, we invoke quadratic programming (QP) and memorize updates to correct the update directions. On the worker side, GradMA harnesses the gradients of the local model in the previous step and the centralized model, and the parameters difference between the local model in the current step and the centralized model as constraints of QP to adaptively correct the gradient of the local model. Furthermore, we maintain a memory state to memorize accumulated update of each worker on the server side. GradMA then explicitly takes the memory state to constrain QP to augment the momentum (i.e., the update direction) of the centralized model. Here, we need the server to allocate memory space to store memory state. However, it may be not feasible in FL scenarios with a large size of workers, which can increase the storage cost and the burden of computing QP largely. Therefore, we carefully craft a memory reduction strategy to alleviate the said limitations. In addition, we theoretically analyze the convergence of GradMA in the smooth non-convex setting.

To sum up, we highlight our contributions as follows:

- We formulate a novel FL approach GradMA, which aims to simultaneously correct the server-side and worker-side update directions and fully harness the server’s rich computing and memory resources. Meanwhile, we tailor a memory reduction strategy for GradMA to reduce the scale of QP and memory cost.

- For completeness, we analyze the convergence of GradMA theoretically in the smooth non-convex setting. As a result, the convergence result of GradMA achieves the linear speed up as the number of selected active workers increases.

- We conduct extensive experiments on four commonly used image classification datasets (i.e., MNIST, CIFAR-10, CIFAR-100 and Tiny-ImageNet) to show that GradMA is highly competitive compared with other state-of-the-art baselines. Meanwhile, ablation studies demonstrate efficacy and indispensability for core modules and key parameters.

2. Related Work

FL with Data Heterogeneity. FedAvg, the classic distributed learning framework for FL, is first proposed by McMahan et al. [26]. Although FedAvg provides a practical and simple solution for aggregation, it still suffers performance deterioration when the data among workers is non-iid [18]. Shortly thereafter, a panoply of modifications for FedAvg have been proposed to handle said issue. For example, FedProx [19] constrains local updates via adding a proximal term to the local objectives. Scaffold [12] uses control variate to augment the local updates. FedDyn [1] dynamically regularizes the objectives of workers to align global and local objectives. Moon [17] corrects the local training by conducting contrastive learning in model-level. Meanwhile, there exists another line of works to improve the global performance of FL through performing knowledge distillation [13, 23, 42, 45, 46] on the server side or worker side. FedMLB [13] architecturally regularizes the local objectives via online knowledge distillation. However, other approaches incur additional communication overhead [42,46] or pseudo data [23,45]. Going beyond the aforementioned approaches, FL with momentum is an effective way to tackle worker drift problem caused by data heterogeneity and accelerate the convergence. Specifically, on the server side, FedAvgM [7] maintains a momentum buffer, whereas FedADAM [28] and FedAMS [33] both adopt adaptive gradient-descent methods to speed up training. FedCM [38] keeps a state, carrying global information broadcasted by the server, on the worker side to address data heterogeneity issue. DOMO [36] and Mime [11] maintain momentum buffers on both server side and worker side to improve the training performance.

FL with Partial Participation. In addition to data heterogeneity issue, another key hurdle to FL stems from partial participation. The causes for partial participation can be roughly classified into two categories. One is the difference in the computing power and communication speed of different workers. A natural way to cope with this situation is to allow asynchronous updates [2, 34, 40]. The other is the different availability mode, in which workers can abort the training midway (i.e., stragglers) [10]. To do so, many approaches may collect information from only a subset of workers to update the centralized model [1,7,11,12,18,26,39]. However, the server in the mentioned approaches simply ignores and discards the information of the stragglers,
which can lead to other problems such as under-utilization of computation and memory [45], slower convergence [18], and biased/unfair use of workers’ information [10]. Recently, MIFA [5] corrects the gradient bias by exploiting the memorized latest updates of all workers, which avoids excessive delays caused by inactive workers and mitigates CF to some extent.

**Continual Learning.** CL is a training paradigm that focuses on scenarios with a continuously changing class distribution of each task and aims at overcoming CF. Existing works for CL can be roughly divided into three branches: expansion-based methods [21, 44], regularization-based methods [14, 32] and memory-based methods [4, 24, 29]. Note that unlike CL, we focus on alleviating CF in distributed data, not sequential data. There are a handful of recent studies that consider FL with CL. For example, FedWeIT [43] focuses on sequential data. FedCurv [30] trains objectives based on all-reduce protocol. FedReg [37] and FCCL [8] require generated pseudo data and public data, respectively.

3. Preliminaries

This section defines the objective function for FL and introduces QP.

In practice, FL is designed to minimize the empirical risk over data distributed across multiple workers without compromising local data. The following optimization problem is often considered:

$$\min_{x \in \mathbb{R}^d} f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x) = \frac{1}{n_i} \sum_{r=1}^{n_i} F_i(x; \xi_r^{(i)})$$, (1)

where $N$ is the number of workers. Moreover, the local objective $f_i : \mathbb{R}^d \to \mathbb{R}$ measures the local empirical risk over data distribution $D_i$, i.e., $\xi_r^{(i)} \sim D_i$, with $n_i$ samples available at $i$-th worker. Note that $D_i$ can be different among workers. In this work, we consider the typical centralized setup where $N$ workers are connected to one central server.

Next, we introduce QP, which is a fundamental optimization problem with well-established solutions and can be widely seen in the machine learning community, to correct the server-side and work-side update directions. In this paper, we can model our goal via QP, which is posed in the following primal form:

$$\min_{\bar{p}} \frac{1}{2} \|p - \bar{p}\|^2 \quad \text{s.t.} \ (\bar{p}, M[i]) \geq 0, \forall i \in [C]$$, (2)

where $p \in \mathbb{R}^d$ and $M \in \mathbb{R}^{d \times C}$ ($C \in \mathbb{N}$). One can see that the goal of (2) is to seek a vector $\bar{p}$ that is positively correlated with $M[i] \in \mathbb{R}^d, \forall i \in [C]$ while being close to $p$. By discarding the constant term $p^\top p$, we rewrite (2) as:

$$\min_{\bar{p}} \frac{1}{2} \bar{p}^\top \bar{p} - p^\top \bar{p} \quad \text{s.t.} \ M^\top \bar{p} \geq 0 \in \mathbb{R}^C$$. (3)

**Algorithm 1** GradMA: A Gradient-Memory-based Accelerated Federated Learning

1. **Input:** learning rates ($\eta_1, \eta_2$), the number of all workers $N$, the number of sampled active workers per communication round $S$, control parameters ($\beta_1, \beta_2$), synchronization interval $I$ and memory size $m (S \leq m \leq \min\{d, N\})$.
2. **Initial state** $x_0^{(i)} = x_0 \in \mathbb{R}^d$ ($\forall i \in [N]$), $\bar{m}_0 = 0$.
3. **Initial counter** $\{c(i) = 0, \forall i \in [N]\}$.
4. **Initial memory state** $D = \{\}$. $buf = \{\}$, $new\_buf = \{\}$.
5. **for** $t = 0, 1, \ldots, T - 1$ **do**
6. **On server:**
7. Server samples a subset $S_t$ with $S$ active workers and transmits $x_t$ to $S_t$.
8. **counter**, $D$, $buf$, $new\_buf$ ← mem_red $(m, S_t, \{ \text{counter, } D, buf, new\_buf \})$.
9. **On workers:**
10. **for** $i \in S_t$ **parallel do**
11. $x_{t+1} = \text{Worker}_\text{Update}(x_t^{(i)}, x_t, \eta_1, I)$,
12. sends $d_{t+1}^{(i)} = x_t - x_t^{(i)}$ to server.
13. **end for**
14. **end for**
15. **On server:**
16. $D, x_{t+1}, \bar{m}_{t+1} = \text{Server}_\text{Update}([d_{t+1}^{(i)}, i \in S_t], \bar{m}_t, D, \eta_2, \beta_1, \beta_2, buf, new\_buf)$.
17. Sends $x_{t+1}$ to sampled active workers in the next round.
18. $new\_buf = \{\}$.
19. **end for**
20. **Output:** $x_T$

However, this is a QP problem on $d$ variables, which are updates of the model. Generally, $d$ can be enormous, resulting in $d$ being much larger than $C$. We thus solve the dual formulation of the above QP problem:

$$\min_{z} \frac{1}{2} z^\top M^\top M z + p^\top M z \quad \text{s.t.} \ z \succeq 0 \in \mathbb{R}^C$$. (4)

Once we solve for the optimal dual variable $z^*$, we can recover the optimal primal solution as $\bar{p} = M z^* + p$.

4. Proposed Approach: GradMA

We now present the proposed FL approach GradMA, see Alg. 1 for complete pseudo-code. Note that the communication cost of GradMA is the same as that of FedAvg. Next, we detail core modules of GradMA, which include the memory reduction strategy (i.e., mem_red()), Worker_Update() and Server_Update() on lines 9, 12 and 16 of Alg. 1, respectively.
Algorithm 2 Worker Update($x', x, \eta, I$)

1: Sets $x_{t+1}' = x'$, $x_0' = x$.
2: for $\tau = 0, 1, \ldots, I - 1$ do
3: \hspace{1em} $g_{\tau}' = \nabla f_\tau(x_{\tau+1}')$
4: \hspace{1em} $G_{\tau}' = [\nabla f_\tau(x_{\tau+1}'), \nabla f_\tau(x_{\tau} - x_t)]$
5: \hspace{1em} $g_{\tau} = QP(g_{\tau}', G_{\tau}')$
6: \hspace{1em} $x_{\tau+1}' = x_{\tau+1}' - \eta g_{\tau}'$
7: end for
8: Output: $x_I'$.

4.1. Correcting gradient for the worker side

Throughout the local update, we leverage QP to perform correcting gradient directions, see Alg. 2. Here, the input of QP (marked as $QP$ for distinction) is $p \leftarrow g_{\tau}'$ and $M \leftarrow G_{\tau}' \in \mathbb{R}^{t \times 3}$ (line 5 of Alg. 2), and its output is the following vector $g_{\tau}$, which is positively correlated with $\nabla f_\tau(x_{\tau+1}'), \nabla f_\tau(x_{\tau} - x_t)$ and $x_{\tau+1}' - x_t$ while ensuring the minimum $\|g_{\tau}' - g_{\tau}\|_2$.

$$g_{\tau} = G_{\tau}' z_{\tau} + g_{\tau}'$$

where $z_{\tau} = [z_{\tau,1}, z_{\tau,2}, z_{\tau,3}]^T$ and $z_{\tau,3} \geq 0 \in \mathbb{R}^3$. Essentially, the output of QP, $G_{\tau}'$, is a conical combination and serves as an update direction for local training. Particularly, when $z_{\tau,1} = 0$, $z_{\tau,2} = 0$ and $z_{\tau,3} > 0$, Eq. (5) is equivalent to the local update of FedProx [19]. The difference is that the control parameter $\mu$ in FedProx is a hyper-parameter, while $z_{\tau,3}$ is determined adaptively by QP. Specifically, when $g_{\tau}'$ is positively correlated with $x_{\tau+1}' - x_t$, i.e., $\langle g_{\tau}', x_{\tau+1}' - x_t \rangle \geq 0$, $z_{\tau,3}$ is approximately equal to 0; otherwise, $z_{\tau,3}$ is greater than 0. In other words, $x_{\tau+1}' - x_t$ acts as a hard constraint only when $g_{\tau}'$ is negatively correlated with $x_{\tau+1}' - x_t$, which makes $g_{\tau}'$ focus more on local information. Moreover, the calculation mechanisms for $z_{\tau,1}$ and $z_{\tau,2}$ are the same as that for $z_{\tau,3}$. When $z_{\tau,1} > 0$ and $z_{\tau,2} > 0$, it indicates that the update direction $g_{\tau}'$ takes into account the previous step and global information about the model, which is inspired by CL [4, 24]. Intuitively, Eq. (5) adaptively taps previous and global knowledge, thus effectively mitigating CF caused by data heterogeneity.

4.2. Correcting update direction for the server side

Now, we describe the proposed update process of the centralized model on the server side, see Alg. 3 for details. For ease of presentation, we define the number of local updates of the server that the server can store as the memory size $m$. To elaborate, we assume that there is enough memory space on the server such that $m = N$. In this way, at communication round $t$, the update process can be streamlined, which takes the form:

$$d_{t+1} = \frac{1}{S} \sum_{i \in S_t} d_{t+1}^{i}, m_{t+1} = \beta_1 \tilde{m}_t + d_{t+1},$$

$$D[i] \left\{ \begin{array}{l}
\beta_2 D[i] + d_{t+1}^{i}, i \in S_t, \\
\beta_2 D[i], i \notin S_t
\end{array} \right.$$

$$\tilde{m}_{t+1} = QP(g(m_{t+1}^\prime, D), x_{t+1} = x_t - \eta g \tilde{m}_{t+1}.$$}

As shown in Eq. (7), we propose that the server allocates memory space to maintain a memory state $D$, which is updated in a momentum-like manner to memorize the accumulated updates of all workers. Each worker only uploads update $d_{t+1}^{i}$ $(i \in [N])$ to the server, and as such the risk of data leakage is greatly reduced. By memorizing accumulated updates of inactive workers, GradMA avoids waiting for any straggler when facing heterogeneous workers with different availability, so as to effectively overcome the adverse effects caused by partial participation. This is different from the recently proposed MIFA [5] (see Alg. 6 in Appendix 1), which stores the latest updates of all workers to perform averaging. However, such a straightforward and naive implementation of integration implicitly increases statistical heterogeneity in situations where different workers have varying data distributions, which can induce bias.

Therefore, the core idea of this paper is how to leverage memorized information to overcome the above challenge effectively. To tackle the challenge, we apply QP (marked as $QP$ for distinction) to seek an update direction $m_{t+1}$ that is positively correlated with buffers $D[i] \in \mathbb{R}^d, \forall i \in [N]$ while being close to $\tilde{m}_{t+1}$. Concretely, the input of $QP$ is $p \leftarrow m_{t+1}$ and $M \leftarrow D \in \mathbb{R}^{d \times N}$, and its output is $m_{t+1}$ (see Eq. (8)), which takes the form $m_{t+1} = D z_{t+1}^\prime + m_{t+1}$, where $z_{t+1}^\prime = [z_{t+1,1}, \ldots, z_{t+1,N}]^T \geq 0 \in \mathbb{R}^N$ is determined adaptively by $QP$. Inherently, $QP$ takes advantage of the accumulated updates of all workers stored on $D$ to correct the update direction $m_{t+1}$ and circumvents the centralized model from forgetting stragglers’ knowledge, thereby alleviating CF induced by partial participation. In particular, one can easily observe that $m_{t+1} = m_{t+1}$ holds if $m = 0$ (that is, $D = 0$). The update process of Alg. 3 is then consistent with that of FedAvgM [7] on the server side. Consequently, Alg. 3 can be considered as an extension of FedAvgM in terms of augmenting updates through allocating memory.

4.3. A Practical Memory Reduction Strategy

It is well known that in realistic FL scenarios, on the one hand, the number of workers may be large; the size of the model may be huge on the other hand, leading to large-scale QP as well as high memory demanding for server to store $D$, which is infeasible and unnecessary in practice.
Therefore, we propose a memory reduction strategy to alleviate this deficiency, which ensures that the size of $D$ does not exceed a pre-given $m$ and $S \leq m \leq \min\{d, N\}$, see Alg. 4 for details. The design ethos of the memory reduction strategy is to keep as much useful information as possible in a given $m$. Specifically, at communication round $t$, the server samples $S$ active workers and performs that $c(i) \leftarrow c(i) + 1$ ($i \in S$) (lines 3 and 15 of Alg. 4). When the memory used is less than the given one, $c(i) \notin \text{buf}$ of sampled active workers enter the buffers $\text{buf}$ and $\text{newbuf}$ in turn (lines 16-17 of Alg. 4). Once the memory used is equal to the given one, $c(i’)$ with the smallest value in $\text{oldbuf}$ is discarded and set $c(i’) = 0$. Also, $D[i’]$ is discarded from $D$ (lines 12-13 of Alg. 4).

5. Convergence Results for GradMA

We now present a convergence analysis of GradMA in the smooth non-convex setting. And the following assumptions are considered.

**Assumption 1** (Global function below bounds). Set $f^* = \inf_{x \in \mathbb{R}^d} f(x)$ and $f^* > -\infty$.

**Assumption 2** ($L$-smooth). $\forall i \in [N]$, the local functions $f_i$ are differentiable, and there exist constant $L > 0$ such that for any $x, y \in \mathbb{R}^d$, $\|\nabla f_i(x) - \nabla f_i(y)\| \leq L \|x - y\|$. 

**Assumption 3** (Bounded data heterogeneity). The degree of heterogeneity of the data distribution across workers can be quantified as $\|\nabla f_i(x) - \nabla f_i(x)^{\ast}\| \leq \rho^2$, for any $i \in [N]$ and some constant $\rho \geq 0$.

**Assumption 4** (Bounded optimal solution error for $QP_1$). Given $g^{(i)} = \nabla f_i(x)$ (see Alg. 2), then there exists $\varepsilon_i > 0$ such that $\|g^{(i)} - g^{(i)}\| \leq \varepsilon_i$.

**Assumption 5** (Bounded optimal solution error for $QP_2$). Given $\beta_2 \in [0, 1]$ and $m$ (see Alg. 3), then there exists $\varepsilon_g > 0$ such that $\|m - \tilde{m}\|^2 \leq \varepsilon_g^2/1-\beta_2$.

Assumptions 1 and 2 are commonly used in the analysis of distribution learning [12, 19, 35]. Assumption 3 quantifies inter-worker variances, i.e., data heterogeneity [12, 19]. Assumptions 4 and 5 are necessary for the theoretical analysis of GradMA, which constrain the upper bound on the optimal solution errors of $QP_1$ and $QP_2$, respectively. Intuitively, the assumptions hold if the local updates for all worker make sense [3]. Note that the upper bound for Assumption 5 follows an intuitive observation: more accumulated updates of workers (i.e., the larger $\beta_2$) can provide more accumulated update information for the centralized model. Next, we state our convergence results for GradMA.

**Theorem 1** Assume Assumptions 1-5 exist. Let $\eta_t \leq \frac{1}{1000 + LT}, \eta_t \leq \frac{(1 - \beta_1)^2S(N - 1)}{10L(N - 1) + 4N(S - 1)}, \eta_t \leq \frac{320T^2\eta_t^2L^2}{(1 - \beta_1)^2S(N - 1)} \leq 1$. For all $t \in [0, \cdots, T - 1]$, the following relationship generated by Alg. 1 holds:

$$
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla f(x_t)\|^2\right] \leq \frac{8(1 - \beta_1)(f(x_0) - f^*)}{T\eta_0\eta_T} + C_1\varepsilon_1^2 + C_2\varepsilon_2^2 + C_3\rho^2,
$$

where the expectation $\mathbb{E}$ is w.r.t the sampled active workers per communication round, and $C_1 = 8 + 320T^2\eta_t^2L^2 + \frac{64\eta_0\eta_2\eta_t(1 + 40\eta_t^2L^2)^2}{(1 - \beta_1)^2S(N - 1)} \cdot \frac{N - S}{S(N - 1)} + \frac{20\eta_2L}{(1 - \beta_1)^2(1 - \beta_2)\eta_0} + \frac{8}{(1 - \beta_2)^2\eta_0^2} \cdot C_3 = C_1 - 8.$
A detailed proof of Theorem 1 is presented in the Appendix 3.

**Corollary 1** Assume Assumptions 1-5 exist. We set $\eta_l = \frac{1}{\tau \epsilon_l}$, $\eta_g = \frac{1}{2\epsilon_g}$, $\epsilon_l = \frac{1}{\tau}$ and $\epsilon_g = \frac{1}{\tau}$. For $T \geq \max \left\{ 160, \frac{(\beta_5(S(N-1)+4N(S-1))^2}{P(I-\beta_5)^T(S(N-1))^2}, \frac{(b+1280)^{0.5}}{4} \right\}$, where $b = \frac{128(N-S)}{(1-\beta_5)^T(S(N-1))^2}$ in Theorem 1, we have:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla f(x_t)\|^2] = O \left( \frac{f_{0.5}}{S^{0.5}T^{0.5}} + \frac{1}{T} \right).$$

An immediate observation from Corollary 1 is that GradMA can achieve the linear speed up as the number of sampled active workers $S$ increases. This convergence rate matches the well-known best result in FL approaches in literature [39] under the smooth non-convex setting.

### 6. Empirical Study

In this section, we empirically investigate GradMA on four datasets (MNIST [16], CIFAR-10, CIFAR-100 [15] and Tiny-Imagenet) commonly used for image classification tasks.

#### 6.1. Experimental Setup

To gauge the effectiveness of Worker_Update() and Server_Update(), we perform ablation study of GradMA. For this purpose, we design Alg. 7 (marked as GradMA-W) and Alg. 8 (marked as GradMA-S), as specified in Appendix 1. Meanwhile, we compare other baselines, including FedAvg [26], FedProx [19], MOON [17], FedMLB [13], FedDyn [1], MimeLite [11], MIFA [5] and slow-momentum variants of FedAvg, FedProx, MIFA, MOON and FedMLB (i.e., FedAvgM [7], FedProxM, MIFAM, MOONM and FedMLBM), in terms of test accuracy and communication efficiency in different FL scenarios. For fairness, we divide the baselines into three groups based on FedAvg’s improvements on the worker side, server side, or both. See Table 1 and Table 2 for details. Furthermore, on top of GradMA-S, we empirically study the effect of the control parameters ($\beta_1, \beta_2$) and verify the effectiveness of men_red() by setting varying memory sizes $m$.

All our experiments are performed on a centralized network with 100 workers. And we fix synchronization interval $I = 5$. To explore the performances of the approaches, we set up multiple different scenarios w.r.t. the number of sampled active workers $S$ per communication round and data heterogeneity. Specifically, we set $S \in \{5, 10, 50\}$. Moreover, we use Diraclet process $D_{\omega}(\omega)$ [1,46] to strictly partition the training set of each dataset across 100 workers. We set $\omega \in \{0.01, 0.1, 1.0\}$. A visualization of the data partitions for the four datasets at varying $\omega$ values can be found in Fig. 8 in Appendix 2. For MNIST, a neural network (NN) with three linear hidden layers is implemented for each worker. For CIFAR-10 (CIFAR-100, Tiny-Imagenet), each worker runs a Lenet-5 [16] (VGG-11 [31], Resnet20 [6]) architecture. Due to the space limitation, we relegate detailed hyper-parameters tuning and full experimental results to Appendix 2.

#### 6.2. Performance Analysis

**Effects of data heterogeneity.** From Table 1, one can see that the performances of all approaches degrade severely with decreasing $\omega$ on MNIST, CIFAR-10 and Tiny-Imagenet, with GradMA being the only approach that is robust while surpasses other baselines with an overwhelming margin against most scenarios. In particular, the higher data heterogeneity, the more superior performance for GradMA. Also, as shown in Table 2 and Fig. 3, GradMA requires much less communication rounds to reach a given test accuracy compared to baselines against most scenarios. These results validate our idea in the sense that the advantage of GradMA comes from the effective adaptive utilization of workers’ information on both the worker side and server side, which alleviates negative impacts caused by the discrepancy of data distributions among workers.

**Impacts of stragglers.** We explore the impacts of different $S$ on MNIST, CIFAR-100 and Tiny-Imagenet. A higher $S$ means more active workers upload updates per communication round. From Table 1, we can clearly see that the performance of all approaches improves uniformly with increasing $S$ on CIFAR-100 and Tiny-Imagenet, where GradMA consistently dominates other baselines in terms of test accuracy. Meanwhile, Fig. 3 shows that the learning efficiency of GradMA consistently outperforms other baselines (see Appendix 2 for more results). However, for MNIST, the test accuracy for most of the approaches does not intuitively improve with increasing $S$. We conjecture that for simple classification tasks and models, the more active workers participating in training, the more prone the centralized model is to overfitting.

**Comments on GradMA-W and GradMA-S.** We now
Table 1. Top test accuracy (%) overview given different FL scenarios.

<table>
<thead>
<tr>
<th>Alg</th>
<th>MINST-NN, S = 10</th>
<th>CIFAR-10, S = 10</th>
<th>CIFAR-10+VGG-11, S = 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>FedAvg</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FedProx</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FedMLB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOON</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scaffold</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FedMA-W</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FedMA-S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MimeLite</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GradMA-W</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GradMA-S</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Communication rounds to reach given test accuracy ac under different FL scenarios. Note that since Scaffold and MimeLite have twice as much communication load per communication round as the other approaches, we use 2× to show the distinction.

<table>
<thead>
<tr>
<th>Alg</th>
<th>MNIST, S = 10</th>
<th>CIFAR-10, S = 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>FedAvg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FedProx</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FedMLB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOON</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scaffold</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FedMA-W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FedMA-S</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MimeLite</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GradMA-W</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GradMA-S</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2. Test accuracy curves selected of GradMA-S as well as baselines over CIFAR-10 and Tiny-Imagenet.

Figure 3. Test accuracy curves selected of GradMA as well as baselines over CIFAR-10 and CIFAR-100.

discuss the empirical performances of GradMA-W and GradMA-S and observe that GradMA-S beats GradMA-W by a significant margin in different FL scenarios, and even slightly outperforms GradMA in a few cases (see Table 1 and Table 2). To put it differently, GradMA leads GradMA-S in most FL scenarios, suggesting that the combination of Worker_Update() and Server_Update() can have a positive effect and thus improve performance. Meanwhile, GradMA-W trumps baselines in most cases, which suggests that Worker_Update() can mitigate the issue of CF and thus augment the centralized model. In addition, we can draw an

empirical conclusion that correcting the update direction of the centralized model on the server can greatly boost accuracy compared to correcting that of the local model for each worker. Selected learning curves shown in Fig. 1 and Fig. 4 verify the above statements.

Next, we further explore effects of (\beta_1, \beta_2) and m on the performance of GradMA-S on MNIST and CIFAR-10.

**Varying control parameters (\beta_1, \beta_2)**. In order to explore effects of (\beta_1, \beta_2) in more detail, we set \beta_1, \beta_2 \in \{0.0,0.1,0.3,0.5,0.7,0.9\}. And we fix m = 100 and S = 10. Notice that some similarities exist between GradMA-S
and MIFA (MIFAM) when $\beta_1 = 0.0$ and $\beta_2 = 0.0$ ($\beta_1 = 0.0$ and $\beta_2 > 0.0$), i.e., they both memorize the latest updates of stragglers at the server side. From Table 1 and Fig. 5 (refer to Appendix .2 for more results), we can see that GradMA-S with $\beta_2 = 0.0$ considerably beats MIFA and MIFAM regardless of value of $\beta_1$. Furthermore, we observe that GradMA-S with $\beta_2 > 0.0$ outperforms GradMA-S with $\beta_2 = 0.0$ in most cases, and the best test accuracy is located in the region of $\beta_2 > 0.0$. This indicates that the accumulated updates of stragglers can provide more effective update information for the centralized model to refine the performance of GradMA-S compared to the latest updates of stragglers.

Varying memory sizes $m$. In a real-world FL scenario, the memory space on the server side determines the value of the tunable parameter $m$ for GradMA-S. Here, we fix $S = 10$ and set $m \in \{0, 10, 20, 40, 60, 80, 100\}$ to carefully look into the performance of GradMA-S with varying $m$.

Notably, GradMA-S with $m = 0$ is equivalent to FedAvgM. From Fig. 6, FedAvgM performs comparably to GradMA-S with $m > 0$ for MNIST under moderate data heterogeneity setting (i.e., $\omega = 0.1$). In contrast, the performance of FedAvgM sharply degrades and is seriously worse than that of GradMA-S with $m > 0$ under high data heterogeneity setting (i.e., $\omega = 0.01$). Meanwhile, for CIFAR-10, GradMA-S with $m > 0$ consistently surpasses FedAvgM, even under mild data heterogeneity setting (i.e., $\omega = 1.0$). Besides, we can see that the performance of GradMA-S does not intuitively and monotonically improve with increasing $m$. This indicates that the quality of the memory reduction strategy is an essential ingredient affecting the performance of GradMA-S for a given $m$. Therefore, how to tailor a more effective memory reduction strategy is one of our future works. From Fig. 7, the learning curves selected also echo the said statements (see Appendix .2 for more results).

7. Conclusions

In this paper, we propose a novel FL approach GradMA, which corrects the update directions of the server and workers simultaneously. Specifically, on the worker side, GradMA utilizes the gradients of the local model in the previous step and the centralized model, and the parameters difference between the local model in the current round and the centralized model as constraints of QP to adaptively correct the update direction of the local model. On the server side, GradMA takes the memorized accumulated gradients of all workers as constraints of QP to augment the update direction of the centralized model. Meanwhile, we provide the convergence analysis theoretically of GradMA in the smooth non-convex setting. Also, we conduct extensive experiments to verify the superiority of GradMA.

8. Acknowledgments

This work has been supported by the National Natural Science Foundation of China under Grant No.U1911203, and the National Natural Science Foundation of China under Grant No.61977025.
References


[22] Yijing Li, Xiaofeng Tao, Xuefei Zhang, Junjie Liu, and Jin Xu. Privacy-preserved federated learning for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 1


