EFL: Elastic Federated Learning on Non-IID Data

Zichen Ma
The Chinese University of Hong Kong, Shenzhen
JD AI Research
China
zichenma1@link.cuhk.edu.cn

Yu Lu
The Chinese University of Hong Kong, Shenzhen
JD AI Research
China
yulu1@link.cuhk.edu.cn

Wenye Li
The Chinese University of Hong Kong, Shenzhen
Shenzhen Research Institute of Big Data
China
wyli@cuhk.edu.cn

Shuguang Cui
The Chinese University of Hong Kong, Shenzhen
Shenzhen Research Institute of Big Data
China
shuguangcui@cuhk.edu.cn

Abstract

Federated learning involves training machine learning models over devices or data silos, such as edge processors or data warehouses, while keeping the data local. However, training in heterogeneous and potentially massive networks introduces bias into the system, originating from the non-IID data and the low participation rate. In this paper, we propose Elastic Federated Learning (EFL), an unbiased federated training framework capable of tackling the heterogeneity[1] in the system. EFL extends lifelong learning to realistic federated settings, makes the most informative parameters less volatile during training, and utilizes the incomplete local updates. It is also an efficient and effective algorithm that compresses both upstream and downstream communications with a convergence guarantee. We empirically demonstrate the efficacy of our framework on a variety of non-IID datasets and show the competitive performance of the algorithm on robustness and efficiency.

1 Introduction

Federated learning (FL) has been an attractive distributed machine learning paradigm where participants jointly learn a global model without data sharing (McMahan et al., 2017a). It embodies the principles of focused collection and data minimization and can mitigate many of the systemic privacy risks and costs resulting from traditional, centralized machine learning (Kairouz et al., 2019). While there are plenty of works on federated optimization, bias in the system remains a crucial challenge. The origins of bias are from (i) the statistical heterogeneity that data are not independent and identically distributed (IID) across clients; (ii) the low participation rate due to limited computing and communication resources (systemic heterogeneity), e.g., network condition, battery, processors, etc.

Existing FL methods empower participants to accomplish several local updates, and the server will abandon struggling clients, which attempt to alleviate the communication burden. The popular algorithm, FedAvg (McMahan et al., 2017a), first allows clients to perform a small number of epochs of local stochastic gradient descent (SGD), then successfully completed clients communicate their model updates back to the server, and stragglers will be abandoned.

While there are many variants of FedAvg, and they have shown empirical success in the non-IID settings, these algorithms do not fully address bias in the system. The solutions are sub-optimal as they either employ a small shared global subset of data (Zhao et al., 2018) or a more significant number of models with increased communication costs (Karimireddy et al., 2020b; Li et al., 2018; 2019). Moreover, to the best of our knowledge, previous models do not consider the low participation rate, which may restrict the potential availability of training datasets and weaken the system's applicability.

This paper develops Elastic Federated Learning (EFL), an unbiased algorithm that addresses the issue of bias mitigation in FL, which is robust to non-IID data (statistical heterogeneity), and different client behaviors (systemic heterogeneity). **Contributions** of the paper are as follows: First, EFL is robust to the non-IID data setting. It incorporates an elastic term into the local objective to improve the algorithm's stability and makes the most informative parameters

[1] Details of the heterogeneity are summarized in Appendix A.1

identified by the Fisher information matrix less volatile. Theoretically, we provide the convergence guarantees for the algorithm.

Second, EFL still converges even when the system has a low participation rate, i.e., many clients may be inactive or return incomplete updates. It utilizes partial information by scaling the corresponding aggregation coefficient. We show that the low participation rate will not impact the convergence, but its tolerance diminishes as the training continues.

Finally, the proposed EFL is a communication-efficient algorithm that compresses both upstream and downstream communications. We provide the convergence analysis of the compressed algorithm and extensive empirical results on different datasets. The algorithm requires both fewer gradient evaluations and communicated bits to converge.

2 Related Work

Federated Optimization Recently we have witnessed significant progress in developing novel methods that address different challenges in FL (Kairouz et al., 2019; Li et al., 2020a). Concretely, there have been various works on different aspects of FL, including preserving users' privacy (McMahan et al., 2017b; Agarwal et al., 2018; Zhu et al., 2020) and lowering communication costs (Reisizadeh et al., 2020; Dai et al., 2019; Basu et al., 2019; Li et al., 2020b). Some works develop algorithms for the homogeneous setting, where the data samples of all users are sampled from the same probability distribution (Stich, 2018; Wang & Joshi, 2018; Zhou & Cong, 2017; Lin et al., 2018). More related to our paper, several works have been proposed to handle the heterogeneity in FL (Zhao et al., 2018; Haddadpour & Mahdavi, 2019; Wang et al., 2020; Zhu et al., 2021; Li et al., 2021), including regularizing model weight updates (Li et al., 2020a), allowing personalized models for clients (Fallah et al., 2020; T Dinh et al., 2020), or introducing new model aggregation schemes (Yurochkin et al., 2019; Mansour et al., 2020). Still, the solutions are not optimal as they violate privacy requirements or increase the communication burden.

EFL differs from these approaches by simultaneously mitigating non-IID data issues with the elastic term and alleviating the low participation rate problem, which originates from systemic heterogeneity, while stabilizing the training. To our best knowledge, we present the first work investigating clients with different behaviors in heterogeneous settings, which appeals to practical applications. Furthermore, our approach can also address communication costs concerns, where communications between clients and the server are compressed and preserve the algorithm's efficiency.

Lifelong Learning/Continuous Learning The problem is defined as learning separate tasks sequentially using a single model without forgetting the previously learned tasks. Several popular approaches have been proposed in this context, such as data distillation (Parisi et al., 2018), model expansion (Rusu et al., 2016; Draelos et al., 2017), and memory consolidation (Soltoggio, 2015; Shin et al., 2017). A particularly successful one is EWC (Kirkpatrick et al., 2017), a method to aid the sequential learning of tasks.

To draw an analogy between FL and the problem of Continuous Learning, we consider the problem of learning a model on each client in the non-IID setting as a separate learning problem. It is not surprising that similar approaches are being applied to solve FL and Continuous Learning problems. One such example is data distillation, in which representative data samples are shared between tasks (Hou et al., 2018; Zhao et al., 2018). However, FL is frequently used to achieve data privacy. Sending data directly from one client to another (or the server) would break privacy. Therefore, we seek other information to be shared between the tasks.

The answer to what kind of information to use may be found in Kirkpatrick et al. (2017). The authors present a new algorithm for Continuous Learning, Elastic Weight Consolidation (EWC), aiming to prevent catastrophic forgetting when moving from learning task A to learning task B. The idea is to identify the coordinates in the network parameters that are the most informative for task A, while task B is learned by penalizing the learner for changing these parameters. The basic assumption is that deep neural networks are over-parameterized enough to have good chances of finding an optimal solution to task B in the neighborhood of the previously learned task A's solution.

Communication-efficient Distributed Learning A wide variety of methods have been proposed to reduce the amount of communication in distributed machine learning. The substantial existing research focuses on: (i) communication delay reduces the communication frequency by performing local optimization (Konečnỳ et al., 2016; McMahan et al., 2017a); (ii) sparsification reduces updates' entropy by restricting changes to only a small subset of parameters (Aji & Heafield, 2017; Tsuzuku et al., 2018); (iii) dense quantization reduces the entropy of the weight updates by restricting all updates to a reduced set of values (Alistarh et al., 2017; Bernstein et al., 2018).

Out of all the above-listed methods, only *FedAvg* and *signSGD* compress upstream and downstream communications. All other methods are of limited utility in FL settings, as they leave communications from the server to clients uncompressed.

Algorithm 1 EFL. N clients are indexed by k, p^k is the probability that the k-th client is selected, η_τ is the learning rate, E is the maximum number of time steps each round has, and $0 \leq s^k_\tau \leq E$ is the number of local updates the k-th client completes in the τ-th round.

```
Server executes:
  initialize $\omega_0$; $R^G_0, R^k_0 \leftarrow 0$.
  for round $\tau$ = 1,2, ... do
      $S_\tau \leftarrow$ (selecting a subset of $N$ clients)
      for client $k \in S_t$ in parallel do
          $\Delta\omega^k_{\tau E}, u^k_\tau, v^k_\tau \leftarrow$ ClientUpdate($k$, $ST(\Delta\omega^G_{(\tau-1)E})$, $\sum_k u^k_{\tau-1}, \sum_k v^k_{\tau-1}$)
          $\Delta\omega^G_{\tau E} = R^G_{(\tau-1)E} + \sum_k p^k_\tau \Delta\omega^k_{\tau E}$
          $R^G_{\tau E} = \Delta\omega^G_{\tau E} - ST(\Delta\omega^G_{\tau E})$
          $\omega^G_{(\tau+1)E} = \omega^G_{\tau E} + \Delta\omega^G_{\tau E}$
      end for
  end for
ClientUpdate($k$, $ST(\Delta\omega^G_{(\tau-1)E})$, $\sum_k u^k_{\tau-1}, \sum_k v^k_{\tau-1}$):
  $\omega^k_{\tau E} = \omega^k_{(\tau-1)E} + ST(\Delta\omega^G_{(\tau-1)E})$
  for $j = 0, ..., s^k_\tau - 1$ do
      $\omega^k_{\tau E+j+1} = \omega^k_{\tau E+j} - \eta_\tau g^k_{\tau E+j}$
  end for
  $\Delta\omega^k_{\tau E} = R^k_{(\tau-1)E} + \omega^k_{\tau E+s^k_\tau} - \omega^k_{\tau E}$
  $R^k_{\tau E} = \Delta\omega^k_{\tau E} - ST(\Delta\omega^k_{\tau E})$
  $u^k_\tau = \text{diag}(I_{\tau,k})$
  $v^k_\tau = \text{diag}(I_{\tau,k})\omega^k_{\tau E+s^k_\tau}$
  return $ST(\Delta\omega^k_{\tau E}), u^k_\tau, v^k_\tau$ to the server
```

3 Elastic Federated Learning

In this section, we first formulate the EFL method, an unbiased algorithm that alleviates the heterogeneity in federated machine learning. Then, we establish the convergence bound for the algorithm when it trains on the non-IID data at the low participation rate. Finally, we discuss the impacts of irregular clients on the algorithm's convergence.

3.1 Problem Formulation

EFL is designed to mitigate the heterogeneity in the system, wherein the problem is originated from the non-IID data across clients and the low participation rate. In particular, the aim is to minimize:

$$\min_\omega F(\omega) = \sum_{k=1}^{N} p^k \widetilde{F}_k(\omega), \tag{1}$$

where N is the total number of clients, n_k is the number of available samples the k-th client owns, $n = \sum_{k=1}^{N} n_k$, and $p^k = \frac{n_k}{n}$ denotes the probability of selecting the k-th client. Here ω represents the model's parameters, and $\widetilde{F}_k(\omega)$ is the local objective of the k-th client.

Assuming there are at most T rounds. For the τ-th round, the clients are connected via a central aggregating server and seek to optimize the following objective locally:

$$\widetilde{F}_{\tau,k}(\omega) = f_k(\omega) + \frac{\lambda}{2}\sum_{i=1}^{N}(\omega - \omega^i_{\tau-1})^T \text{diag}(I_{\tau-1,i})(\omega - \omega^i_{\tau-1}), \tag{2}$$

where $f_k(\omega)$ is the local empirical risk over all available samples at the k-th client. $\omega^i_{\tau-1}$ is the model parameters of the i-th client in the $(\tau-1)$-th round. $I_{\tau-1,i} = I(\omega^i_{\tau-1})$ is the Fisher information matrix, which is the negative expected Hessian of the log-likelihood function. $\text{diag}(I_{\tau-1,i})$ is the matrix that preserves values of the diagonal of the Fisher information matrix, which aims to penalize parts of the parameters that are too volatile in a round.

Algorithm 2 Compression Method ST. q is the sparsity, tensor $T \in \mathbb{R}^n$, and $\hat{T} \in \{-\mu, 0, \mu\}^n$

```
**ST($T$):**
    $k = \max(nq, 1)$; $e = top_k(|T|)$
    $mask = (|T| \geq e) \in \{0, 1\}^n$; $T^{mask} = mask \times T$
    $\mu = \frac{1}{k} \sum_{i=1}^{n} |T_i^{mask}|$
    $\hat{T} = \mu \times sign(T^{mask})$
    return $\hat{T}$
```

We propose adding the elastic term (the second term of Equation (2)) to the local subproblem to restrict the most informative parameter changes. It alleviates bias originating from the non-IID data and stabilizes the training. We can further rearrange Equation (2) as

$$\widetilde{F}_{\tau,k}(\omega) = f_k(\omega) + \frac{\lambda}{2}\omega^T \sum_{i=1}^{N} \text{diag}(I_{\tau-1,i})\omega - \lambda\omega^T \sum_{i=1}^{N} \text{diag}(I_{\tau-1,i})\omega_{\tau-1}^{i} + Z, \tag{3}$$

where Z is a constant. Let $u_{\tau-1}^k = \text{diag}(I_{\tau-1,k})$, and $v_{\tau-1}^k = \text{diag}(I_{\tau-1,k})\omega_{\tau-1}^k$.

Suppose ω^* is the minimizer of the global objective F, and denote by $\widetilde{F}_k^*$ the optimal value of $\widetilde{F}_k$. Let the degree to which data at the k-th client is distributed differently than that at other clients as $D_k = \widetilde{F}_k(\omega^*) - \widetilde{F}_k^*$, where $D = \sum_{k=1}^{N} p^k D_k$. We consider discrete time steps $t = 0, 1, ...$ Model weights are aggregated and synchronized when t is a multiple of E, i.e., each round consists of E time steps. In the τ-th round, EFL, presented in Algorithm 1, executes the following steps:

First, the server broadcasts the compressed latest global weight updates $ST(\Delta\omega_{(\tau-1)E}^G)$, $\sum_k u_{\tau-1}^k$, and $\sum_k v_{\tau-1}^k$ to participants. Each client then updates its local weight by $\omega_{\tau E}^k = \omega_{(\tau-1)E}^k + \Delta\omega_{(\tau-1)E}^G$.

Second, each client runs SGD on its local objective $\widetilde{F}_k$ for $j = 0, ..., s_\tau^k - 1$:

$$\omega_{\tau E+j+1}^k = \omega_{\tau E+j}^k - \eta_\tau g_{\tau E+j}^k, \tag{4}$$

where η_τ is a learning rate that decays with τ, $0 \leq s_\tau^k \leq E$ is the number of local updates the client completes in the τ-th round, $g_t^k = \nabla\widetilde{F}_k(\omega_t^k, \xi_t^k)$ is the stochastic gradient of the k-th client, and ξ_t^k is a mini-batch sampled from client k's local data. $\overline{g}_t^k = \nabla\widetilde{F}_k(\omega_t^k)$ is the full batch gradient at client k, and $\overline{g}_t^k = \mathbb{E}_{\xi_t^k}[g_t^k]$, $\Delta\omega_{\tau E}^k = R_{(\tau-1)E}^k + \omega_{\tau E+s_\tau^k}^k - \omega_{\tau E}^k$, where each client computes the residual as

$$R_{\tau E}^k = \Delta\omega_{\tau E}^k - ST(\Delta\omega_{\tau E}^k). \tag{5}$$

$ST(\cdot)$ is the compression method presented in Algorithm 2. The client sends the compressed local updates $ST(\Delta\omega_{\tau E}^k)$, u_τ^k, and v_τ^k back to the coordinator.

Finally, the server aggregates the next global weight by

$$\begin{aligned}
\omega_{(\tau+1)E}^G &= \omega_{\tau E}^G + \Delta\omega_{\tau E}^G \\
&= \omega_{\tau E}^G + R_{(\tau-1)E}^G + \sum_{k=1}^{N} p_\tau^k \Delta\omega_{\tau E}^k \\
&= \omega_{\tau E}^G + R_{(\tau-1)E}^G - \sum_{k=1}^{N} p_\tau^k \sum_{j=0}^{s_\tau^k} \eta_\tau g_{\tau E+j}^k,
\end{aligned} \tag{6}$$

where $R_{\tau E}^G = \Delta\omega_{\tau E}^G - ST(\Delta\omega_{\tau E}^G)$.

As mentioned in Section 1, clients' low participation rate in an FL system is prevalent in reality. EFL mainly focuses on two situations that lead to the low participation rate, which are not yet well discussed previously: (i) incomplete clients that can only submit partially complete updates; (ii) inactive clients that cannot respond to the server.

Client k is inactive in the τ-th round if $s_\tau^k = 0$, i.e., it does not perform the local training, and k is incomplete if $0 < s_\tau^k < E$. s_τ^k is a random variable that can follow an arbitrary distribution. It can generally be time-varying, i.e.,

Table 1: The number of required communication rounds to reach ϵ-accuracy. SC refers to strongly convex, NC is non-convex, and δ in MIME bounds Hessian dissimilarity. EFL preserves the optimal statistical rates (first term in SCAFFOLD) while improving the optimization.

Algorithm	Bounded gradient	Convexity	# Communication rounds
SCAFFOLD	✓	μ-SC	$\frac{G^2}{\mu S\epsilon} + \frac{G}{\mu\sqrt{\epsilon}} + \frac{L}{\mu}$
MIME	✓	μ-SC	$\frac{G^2}{\mu S\epsilon} + \frac{\delta}{\mu}$
VRL-SGD	×	NC	$\frac{N\sigma^2}{S\epsilon^2} + \frac{N}{\epsilon}$
FedAMP	✓	NC	$\frac{G^2}{LS\epsilon} + \frac{G}{\epsilon^{\frac{3}{2}}} + \frac{L^2}{\sqrt{\epsilon}}$
EFL	✓	μ-SC	$\frac{G^2}{\mu S\epsilon} + \frac{L}{\mu\sqrt{\epsilon}}$

it may follow different distributions at different time steps. EFL also allows the aggregation coefficient p_τ^k to vary with τ, and in the following subsection, we explore various schemes of choosing p_τ^k and their impacts on the model convergence.

EFL incorporates the sparsification and quantization to compress the upstream (from clients to the server) and the downstream (from the server to clients) communications. It is not economical to communicate the elements at full precision as the standard top-k sparsification (Aji & Heafield, 2017) method does. As a result, EFL quantizes the top-k components of the sparsified updates to the mean population magnitude, leaving the updates with a ternary tensor containing $\{-\mu, 0, \mu\}$, The details are summarized in Algorithm 2.

3.2 Convergence Analysis

Five assumptions are made to help analyze the convergence behaviors of the EFL algorithm.

Assumption 1 *(L-smoothness) $\widetilde{F}_1, ..., \widetilde{F}_N$ are L-smooth, and F is also L-smooth.*

Assumption 2 *(Strong convexity) $\widetilde{F}_1, ..., \widetilde{F}_N$ are μ-strongly convex, and F is also μ-strongly convex.*

Assumption 3 *(Bounded variance) The variance of the stochastic gradients is bounded by $\mathbb{E}_\xi||g_t^k - \overline{g}_t^k||^2 \leq \sigma_k^2$*

Assumption 4 *(Bounded gradient) The expected squared norm of the stochastic gradients at each client is bounded by $\mathbb{E}_\xi||g_t^k||^2 \leq G^2$.*

Assumption 5 *(Bounded aggregation coefficient) The aggregation coefficient has an upper bound, which is given by $p_\tau^k \leq \theta p^k$.*

Assuming $\mathbb{E}[p_\tau^k], \mathbb{E}[p_\tau^k s_\tau^k], \mathbb{E}[(p_\tau^k)^2 s_\tau^k]$, and $\mathbb{E}[\sum_{k=1}^N p_\tau^k - 2 + \sum_{k=1}^N p_\tau^k s_\tau^k]$ exist for all rounds τ and clients k, and $\mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k] \neq 0$. The convergence bound can be derived as

Theorem 1 *Under Assumptions 1 to 5, for learning rate $\eta_\tau = \frac{16E}{\mu((\tau+1)E+\gamma)\mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k]}$, the EFL satisfies*

$$\mathbb{E}||\omega_{\tau E}^G - \omega^*||^2 \leq \frac{C_\tau}{(\tau E + \gamma)^2} + \frac{H_\tau J}{\tau E + \gamma}, \tag{7}$$

where $\gamma = \max\{\frac{4E^2\theta}{\min_\tau \mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k]}, \frac{32E(1+\theta)L}{\mu \min_\tau \mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k]}\}$, $H_\tau = \sum_{t=0}^{\tau-1} \mathbb{E}[r_t]$, $r_t \in \{0, 1\}$ indicates the ratio $\frac{\mathbb{E}[p_\tau^k s_\tau^k]}{p^k}$ has the same value for all k, $C_\tau = \max\{\gamma^2 \mathbb{E}||\omega_0^G - \omega^*||^2, (\frac{16E}{\mu})^2 \sum_{t=0}^{\tau-1} \frac{\mathbb{E}[B_t]}{(\mathbb{E}[\sum_{k=1}^N p_t^k s_t^k])^2}\}$, $B_t = \sum_{k=1}^N (p_t^k)^2 s_t^k \sigma_k^2 + 2(2+\theta)L\sum_{k=1}^N p_t^k s_t^k D_k + (2 + \frac{\mu}{2(1+\theta)L})E(E-1)G^2(\sum_{k=1}^N p_t^k s_t^k + \theta(\sum_{k=1}^N p_t^k - 2) + \sum_{k=1}^N p_t^k s_t^K) + 2EG^2 \sum_{k=1}^N \frac{(p_t^k)^2}{p^k} s_t^k$, $J = \max_\tau\{\frac{32E\sum_{k=1}^N \mathbb{E}[p_\tau^k s_\tau^k]}{D_k/\mu\mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k]}\}$.

Based on Theorem 1, $C_\tau = O(\tau)$, which means $\omega_{\tau E}^G$ will finally converge to a global optimum as $\tau \to \infty$ if H_τ increases sub-linearly with τ. Table 1 summarizes the required number of communication rounds of SCAFFOLD (Karimireddy et al., 2020b), MIME (Karimireddy et al., 2020a), VRL-SGD (Liang et al., 2019), FedAMP (Huang et al., 2021), and EFL. The proposed EFL algorithm achieves a tighter bound than methods that assume μ-strongly convex. The proof of Theorem 1 is summarized in Appendix A.2.

4 Impacts of Irregular Clients

Non-IID data are prevalent in FL, where the data distribution on the i-th client P_i is different from P_j on the j-th client, i.e., data are highly skewed, extremely imbalanced, and vary across clients (McMahan et al., 2017a; Zhao et al., 2018). From a statistical perspective, it leads to distribution shifts, which raises the difficulties of model convergence (Hsieh et al., 2020). In reality, non-IID data issues naturally arise in recommender systems and personalized advertisement placement. For example, different mobile phone users who read news articles may be interested in different news categories like politics, sports, or fashion; advertisement platforms might need to send different types of ads to different groups of customers.

Clients with limited computation resources (CPU, RAM), low power, or poor network conditions may fail to communicate with the server, i.e., be inactive or return incomplete updates. These undesirable behaviors introduce the bias into the system and are proved to degrade the system's performance since they magnify the discrepancies between irregular models and the global model (Sahu et al., 2018; Chen et al., 2018).

EFL can absorb incomplete and inactive clients into the federated training, unlike previous algorithms, which abandon these clients. This section investigates the impacts of clients' different behaviors, including being inactive, incomplete, new client arrival, and client departure.

4.1 Inactive Client

If inactive clients exist, the convergence rate changes to $O(\frac{\sum_{t=0}^{\tau-1} y_t J}{\tau E+E^2})$. y_t indicates if there are inactive clients in the t-th round or not. Furthermore, the term converges to zero if $y_t \in O(\tau)$, which means that a mild degree of the inactive client will not discourage the convergence. A client can frequently become inactive due to the limited resources in reality. Permanently removing the client, in this case, may improve the model performance. Specially, we will remove the client if the system without this client leads to a smaller training loss when it terminates at the deadline T.

Suppose a client a is inactive with the probability $0 < y^a < 1$ in each round, and let $f_0(\tau)$ be the convergence bound if we keep the client, $f_1(\tau)$ be the bound if it is abandoned at τ_0. For f_0 with sufficiently many steps, the first term in Equation (7) shrinks to zero, and the second term converges to $y^a J$. Thus, $f_0 \approx y^a J$, and we can obtain that $f_1(\tau) = \frac{\widetilde{C}_\tau}{(E(\tau-\tau_0)+\widetilde{\gamma})^2}$ for some $\widetilde{C}_\tau$ and $\widetilde{\gamma}$, thus we have

Corollary 1 *An inactive client a should be abandoned if*

$$y^a J > f_1(T). \tag{8}$$

Assuming $C_\tau \approx \widetilde{C}_\tau = \tau C$ and $\gamma \approx \widetilde{\gamma}$, i.e., the removed client does not significantly affect the overall SGD variance and the degree of non-IID, then Equation (8) can be formulated as

$$y^a > O(\frac{C}{JTE}). \tag{9}$$

From Corollary 1, the more epochs the training on the local client, the more sensitive it is to the in-activeness.

4.2 Incomplete Client

Based on Theorem 1, the convergence bound is controlled by the expectation of p_τ^k and its functions. EFL allows clients to upload partial updates with adaptive weight $p_\tau^k = \frac{Ep^k}{s_\tau^k}$. It assigns a more significant aggregation coefficient to clients that complete fewer local epochs and turns out to guarantee the convergence in the non-IID setting. The resulting convergence bound follows $O(\frac{E^5 \sum_{t=0}^{\tau-1} \sum_k^N p^k \mathbb{E}[1/s_t^k] + E^2 \sum_{t=0}^{\tau-1} \sum_k^N (p^k \sigma_k)^2 \mathbb{E}[1/s_t^k]}{(\tau E+E^2)^2})$.

The reason for enlarging the aggregation coefficient lies in Equation (6) that increasing p_τ^k is equivalent to increasing the learning rate of client k, by assigning clients that complete fewer epochs a more significant aggregation coefficient, these clients effectively run further in each local step, compensating for fewer epochs.

4.3 Client Departure

If k-th client quits at $\tau_0 < T$, the server will receive no more updates and $s_\tau^k = 0$ for all $\tau > \tau_0$. As a result, the value of ratio $\frac{\mathbb{E}[p_\tau^k s_\tau^k]}{p^k}$ is different for different k, and $r_\tau = 1$ for all $\tau > \tau_0$. According to Theorem 1, $\omega_{\tau E}^G$ cannot converge to the global optimum ω^* as $H_T \geq T - \tau_0$. Intuitively, a client should contribute sufficiently many updates for its features to be captured by the trained model in the non-IID setting. After a client leaves, the remaining training steps will not keep much memory as it runs more rounds. Thus, the model may not apply to the leaving client, especially when it leaves early in training ($\tau_0 \ll T$), which indicates that we may discard the departing client if we cannot guarantee the trained model performs well on it. The earlier a client leaves, the more likely it should be discarded.

However, removing the departing client (the a client) from the training may push the original learning objective $F = \sum_{k=1}^{N} p^k \widetilde{F}_k$ towards the new one $\hat{F} = \sum_{k=1,k\neq a}^{N} p^k \widetilde{F}_k$. The optimal weight ω^* will shift to some $\hat{\omega}^*$ that minimizes $\hat{F}$. A gap exists between two optima, which adds a term to the convergence bound obtained in Theorem 1. Thus, more updates are required for $\omega_{\tau E}^{G}$ to converge to the new optimal $\hat{\omega}^*$.

4.4 Client Arrival

The same argument holds when a new client joins in the training, requiring changing the original global objective to include the new client's loss. The learning rate also needs to be increased when the objective changes. Intuitively, if the shift happens at a large time τ_0, where $\omega_{\tau E}^{G}$ approaches the old optimum ω^* and η_{τ_0} is close to zero, reducing the latest differences $||\omega_{\tau E}^{G} - \hat{\omega}^*||^2 \approx ||\omega^* - \hat{\omega}^*||^2$ with a small learning rate is inapplicable. Thus, a more significant learning rate should be adopted, which is equivalent to initiating a fresh start after the shift, and they still need more updating rounds to address the new client fully.

We also present the bounds of the additional term due to the objective shift as

Theorem 2 *For the global objective shift $F \to \hat{F}$ and $\omega^* \to \hat{\omega}^*$, let $\hat{D}_k = F_k(\hat{\omega}^*) - F_k^*$ quantify the degree of non-IID data for the new objective. If client a quits the system, we have*

$$||\omega^* - \hat{\omega}^*||^2 \leq \frac{8Ln_a^2\hat{D}_a}{\mu^2 n^2}. \tag{10}$$

If client a joins the system, we have

$$||\omega^* - \hat{\omega}^*||^2 \leq \frac{8Ln_a^2\hat{D}_a}{\mu^2 (n+n_a)^2}, \tag{11}$$

where n is the total number of samples before the shift.

We can conclude that the bound reduces when the data becomes more IID and the changed client owns fewer data samples.

5 Experiments

In this section, we first demonstrate the effectiveness and efficiency of EFL in the non-IID data setting and compare it with several baseline algorithms. Then, we show the robustness of EFL on the low participation rate challenge.

5.1 Experimental Settings

Both convex and non-convex models are evaluated on several benchmark datasets of FL. Specifically, we adopt MNIST (LeCun et al., 1998), EMNIST (Cohen et al., 2017) dataset with Resnet50 (He et al., 2016), CIFAR100 dataset (Krizhevsky et al., 2009) with VGG11 (Simonyan & Zisserman, 2014) network, Shakespeare dataset with an LSTM (McMahan et al., 2017a) to predict the next character, Sentiment140 dataset (Go et al., 2009) with an LSTM to classify sentiment, and synthetic dataset with a linear regression classifier.

All experiments are implemented using PyTorch (Paszke et al., 2019) and run on a cluster where each node is equipped with 4 Tesla P40 GPUs and 64 Intel(R) Xeon(R) CPU E5-2683 v4 cores @ 2.10GHz. For reference, statistics of datasets, and implementation details are summarized in Appendix A.3.

5.2 Effects of Non-IID Data

We run experiments with a simplified version of the well-studied 11-layer VGG11 network, which we train on the CIFAR100 dataset in an FL setup using 100 clients. We randomly split the training data into equally sized shards for the IID setting and assigned one shard to every clients. For the non-IID (m) setting, we assign every client sample

Table 2: Efficiency comparison between EFL, FedAvg and FedProx using EMNIST, CIFAR100 and Shakespeare datasets. Summary of average (standard deviation) testing accuracy is reported. Our method maintains competitive performance but needs fewer ($\sim 10\%$) communication intermediaries. we report an average value of 50 runs.

Dataset	Method	Testing Accuracy	# Comm. Rounds$\times 100$	Total Comm. Bits
EMNIST	EFL	$85.93(.27)$	41	4.526×10^3
	FedAvg	$73.21(.65)$	66	2.292×10^4
	FedProx	$78.31(.40)$	38	2.466×10^4
CIFAR100	EFL	$81.38(.33)$	163	3.258×10^8
	FedAvg	$2.27(.79)$	275	2.958×10^9
	FedProx	$80.16(.51)$	187	2.721×10^9
Shakespeare	EFL	$60.49(.38)$	254	1.694×10^9
	FedAvg	$51.35(.74)$	346	1.703×10^{10}
	FedProx	$54.28(.52)$	238	1.917×10^{10}

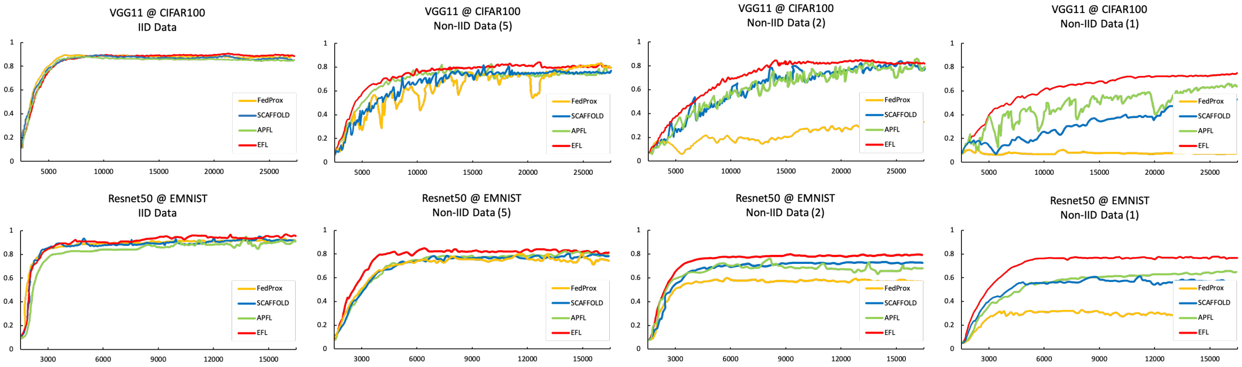

Figure 1: Testing Accuracy-Communication Rounds comparisons of VGG11 on CIFAR100 and Resnet50 on EMNIST for IID and non-IID settings. In the non-IID cases, every client only holds samples from exactly m classes in the dataset. All methods suffer from degraded convergence speed in the non-IID situation, but EFL is affected by far the least.

from exactly the m classes of the dataset. We also perform experiments with Resnet50, where we train on the EMNIST dataset under the same setup of the FL environment. Both models are trained using SGD.

Figure 1 shows the convergence comparison in gradient evaluations for the two models using different algorithms. FedProx (Li et al., 2018) incorporates a proximal term in local objective to improve the model performance on the non-IID data, SCAFFOLD (Karimireddy et al., 2020b) adopts control variate to alleviate the effects of data heterogeneity, and APFL (Deng et al., 2020) learns personalized local models to mitigate heterogeneous data on clients.

We observe that while all methods achieve comparably fast convergence in terms of gradient evaluations on IID data, they suffer considerably in the non-IID setting. From left to right, as data becomes more non-IID, convergence becomes worse for FedProx, and it can diverge in some cases. SCAFFOLD and APFL exhibit their ability in alleviating the data heterogeneity but are not stable during training. As this trend can also be observed for Resnet50 on EMNIST case, we can concluded that the performance loss that is originated from the non-IID data is not unique to some functions. We defer the experiments using different base models (ResNet50 on CIFAR100 and VGG11 on EMNIST datasets) to Appendix A.5.

We subsequently characterize EFL's performance efficiency. Regarding the communication costs, We report results in Table 2 to show that our method maintains better performances but needs fewer ($\sim 10\%$) communicated bits.

Aiming to illustrate the proposed algorithm's effectiveness better, we further evaluate and compare EFL with the state-of-the-art algorithms, including FedGATE (Haddadpour et al., 2021), VRL-SGD (Liang et al., 2019) (methods with regularization terms to alleviate non-IID data), APFL (Deng et al., 2020), and PFedMe (T Dinh et al., 2020) (personalized algorithms to mitigate the heterogeneous issue) on MNIST, CIFAR100, Sentiment140, and Shakespeare datasets. The performance of all the methods is evaluated by the best mean testing accuracy (BMTA) in percentage,

Table 3: BMTA comparisons for the non-IID data setting

Methods	MNIST	CIFAR100	Sentiment140	Shakespeare
FedAvg	98.30	2.27	59.14	51.35
FedGATE	**99.15**	80.94	68.84	54.71
VRL-SGD	98.86	2.81	68.62	52.33
APFL	98.49	77.19	68.81	55.27
PFedMe	99.06	81.17	**69.01**	58.42
EFL	99.10	**81.38**	68.95	**60.49**

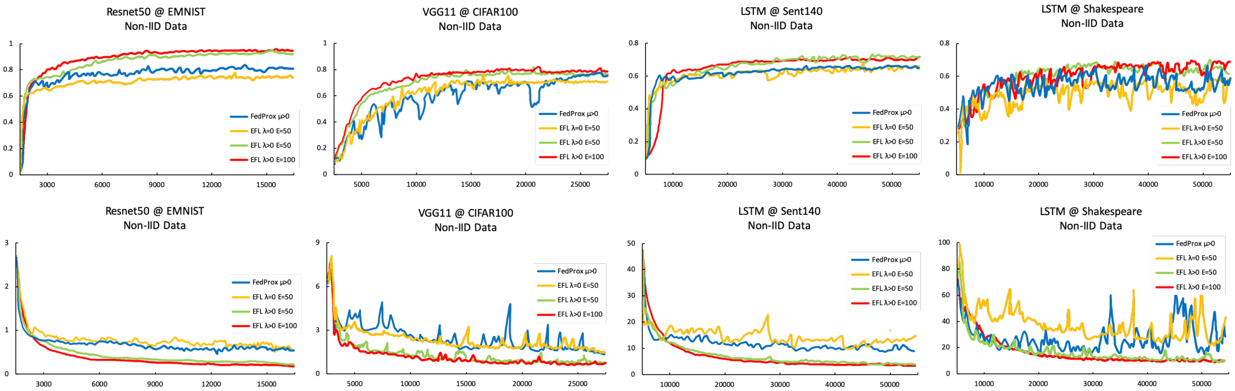

Figure 2: The first row shows Testing Accuracy-Communication Rounds comparison, and the second row shows Training Loss-Communication Rounds comparison in non-IID settings. EFL with elastic term stabilizes and improves the convergence of the algorithm.

where the mean testing accuracy is the average of the testing accuracy on all participants. For each of the datasets, we apply a non-IID data setting.

Table 3 shows the BMTA of all the methods under non-IID data setting, which is not easy for vanilla algorithm FedAvg. On the challenging CIFAR100 dataset, VRL-SGD is unstable and performs catastrophically because the models are destroyed such that the customized gradient updates in the method can not tune it up. APFL and PFedMe train personalized models to alleviate the non-IID data. However, the performance of APFL is still damaged by unstable training. FedGATE, PFedMe, and EFL achieve comparably good performance on all datasets. An insight implicit in this set of experiments is that while personalized FL emerges as prevalent solutions for heterogeneous settings, well-designed algorithms with regularization like EFL can achieve competitive performances regarding testing accuracy.

5.3 Ablation Studies: λ, E, and q

EFL utilizes the incomplete local updates, which indicates that clients may perform a different amount of local epochs s_τ^k. This parameter and the elastic term scaled by λ affect the algorithm's performance. However, s_τ^k is determined by its constraints, i.e., it is a client-specific parameter. EFL can only set the maximum number of local epochs to prevent local models from drifting too far away from the global model and tuning the best λ. Intuitively, a proper λ choice restricts the optimization trajectory by limiting the most informative parameters' change and guarantees convergence.

We explore the impacts of the elastic term by setting different values of λ and investigate whether the maximum number of local epochs influences the convergence behavior of the algorithm. Figure 2 shows the performance comparison on several datasets using different models. We compare EFL with $\lambda = 0$ and EFL with best λ. We can observe that the appropriate λ stabilizes methods' convergence and force divergent algorithms to converge for all datasets. It also increases the accuracy in most cases. As a result, setting $\lambda \geq 0$ is particularly useful in the non-IID setting, indicating the EFL benefits realistic federated environments.

We also conduct an additional ablation over different values of sparsity parameter (q) and report these results in Table 4. We found a strong correlation between the sparsity indicator q and performances. Increasing q (less sparse updates)

Table 4: Comparisons between EFL with different sparsity parameters (q) on EMNIST, CIFAR100, and Shakespeare datasets are reported. Average (standard deviation) testing accuracy and the total communicated bits when the algorithm converges are summarized.

Dataset	Method (EFL)	Testing Accuracy	Total Comm. Bits
EMNIST	$q = .1$	$80.75(.30)$	3.749×10^3
	$q = .3$	$85.93(.27)$	4.526×10^3
	$q = .6$	$86.04(.25)$	7.183×10^3
CIFAR100	$q = .1$	$78.84(.48)$	1.429×10^8
	$q = .3$	$81.38(.33)$	3.258×10^8
	$q = .6$	$81.52(.29)$	6.310×10^8
Shakespeare	$q = .1$	$53.07(.53)$	8.398×10^8
	$q = .3$	$55.15(.49)$	1.241×10^9
	$q = .6$	$60.49(.38)$	1.694×10^9

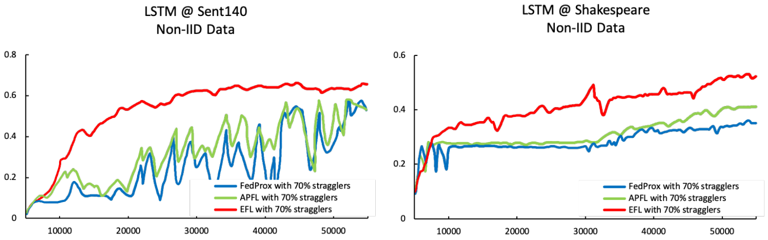

Figure 3: Testing Accuracy-Communication Rounds comparisons among different algorithms in the low participation rate. EFL utilizes incomplete updates from stragglers and is robust to the low participation rate.

leads to a more accurate model and needs more communicated bits (less efficient training). However, enlarging q will not always bring the same returns on the testing accuracy, i.e., a more significant q contributes less to the performance increase. At the same time, it needs more communicated bits on EMNIST and CIFAR100 datasets. One possible reason is that the updates in EFL may be somewhat noisy, especially during earlier federation rounds when client local models are not sufficiently trained. Simply enlarging q beyond a threshold will not be as helpful as doing so when q is small.

5.4 Robustness of EFL

Finally, Figure 3 demonstrates that EFL is robust to the low participation rate. In particular, we track the convergence speed of LSTM trained on Sentiment140 and Shakespeare datasets. We can observe that reducing the participation rate negatively affects all methods. However, the causes for these adverse effects are different: In FedAvg, the actual participation rate is determined by the number of clients that finish the complete training process because it does not include the incomplete updates. It steers the optimization process away from the minimum and might even cause catastrophic forgetting. On the other hand, a low participation rate reduces the convergence speed of EFL by causing the clients' residuals to go out of sync and increasing the gradient staleness. The more rounds a client has to wait before it is selected to participate in training again, the more outdated the accumulated gradients become.

6 Conclusion

This paper proposes EFL as an unbiased FL algorithm that can adapt to the heterogeneous statistical issue by making the most informative parameters less volatile. EFL can be understood as an extension paradigm of lifelong learning, which tackles bias originating from the non-IID data and the low participation rate in FL. Theoretically, we provide convergence guarantees for EFL training on the non-IID data at the low participation rate. Empirically, experiments support the competitive performance of the algorithm on robustness and efficiency.

Acknowledgments

The work was supported in part by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund, and by Guangdong Research Project No. 2017ZT07X152 and 2021A1515011825.

References

Naman Agarwal, Ananda Theertha Suresh, Felix Yu, Sanjiv Kumar, and H Brendan Mcmahan. cpSGD: Communication-efficient and differentially-private distributed SGD. *arXiv preprint arXiv:1805.10559*, 2018.

Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. *arXiv preprint arXiv:1704.05021*, 2017.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Communication-efficient SGD via gradient quantization and encoding. In *Advances in Neural Information Processing Systems*, pp. 1709–1720, 2017.

Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-SGD: Distributed SGD with quantization, sparsification, and local computations. *arXiv preprint arXiv:1906.02367*, 2019.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. signSGD: Compressed optimisation for non-convex problems. *arXiv preprint arXiv:1802.04434*, 2018.

Tianyi Chen, Georgios Giannakis, Tao Sun, and Wotao Yin. Lag: Lazily aggregated gradient for communication-efficient distributed learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. EMNIST: Extending MNIST to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Xinyan Dai, Xiao Yan, Kaiwen Zhou, Han Yang, Kelvin KW Ng, James Cheng, and Yu Fan. Hyper-sphere quantization: Communication-efficient SGD for federated learning. *arXiv preprint arXiv:1911.04655*, 2019.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Timothy J Draelos, Nadine E Miner, Christopher C Lamb, Jonathan A Cox, Craig M Vineyard, Kristofor D Carlson, William M Severa, Conrad D James, and James B Aimone. Neurogenesis deep learning: Extending deep networks to accommodate new classes. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 526–533. IEEE, 2017.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. In *International Conference on Artificial Intelligence and Statistics*, pp. 2350–2358. PMLR, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Lifelong learning via progressive distillation and retrospection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 437–452, 2018.

Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pp. 4387–4398. PMLR, 2020.

Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020a.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020b.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. *arXiv preprint arXiv:2102.02079*, 2021.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of FedAvg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

Zhize Li, Dmitry Kovalev, Xun Qian, and Peter Richtárik. Acceleration for compressed gradient descent in distributed and federated optimization. *arXiv preprint arXiv:2002.11364*, 2020b.

Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Variance reduced local SGD with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.

Tao Lin, Sebastian U Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local SGD. *arXiv preprint arXiv:1808.07217*, 2018.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pp. 1273–1282. PMLR, 2017a.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017b.

German I Parisi, Jun Tani, Cornelius Weber, and Stefan Wermter. Lifelong learning of spatiotemporal representations with dual-memory recurrent self-organization. *Frontiers in neurorobotics*, 12:78, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2021–2031. PMLR, 2020.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Anit Kumar Sahu, Tian Li, Maziar Sanjabi, Manzil Zaheer, Ameet Talwalkar, and Virginia Smith. On the convergence of federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 3, 2018.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. In *Advances in neural information processing systems*, pp. 2990–2999, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Andrea Soltoggio. Short-term plasticity as cause–effect hypothesis testing in distal reward learning. *Biological cybernetics*, 109(1):75–94, 2015.

Sebastian U Stich. Local SGD converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

Yusuke Tsuzuku, Hiroto Imachi, and Takuya Akiba. Variance-based gradient compression for efficient distributed deep learning. *arXiv preprint arXiv:1802.06058*, 2018.

Hao Wang, Zakhary Kaplan, Di Niu, and Baochun Li. Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, pp. 1698–1707. IEEE, 2020.

Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

Mikhail Yurochkin, Mayank Agarwal, Soumya Ghosh, Kristjan Greenewald, Nghia Hoang, and Yasaman Khazaeni. Bayesian nonparametric federated learning of neural networks. In *International Conference on Machine Learning*, pp. 7252–7261. PMLR, 2019.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Fan Zhou and Guojing Cong. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. *arXiv preprint arXiv:1708.01012*, 2017.

Hangyu Zhu, Jinjin Xu, Shiqing Liu, and Yaochu Jin. Federated learning on non-iid data: A survey. *Neurocomputing*, 465:371–390, 2021.

Wennan Zhu, Peter Kairouz, Brendan McMahan, Haicheng Sun, and Wei Li. Federated heavy hitters discovery with differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pp. 3837–3847. PMLR, 2020.

A Appendix

A.1 Heterogeneity in Federated Learning

Federated learning (FL) usually requires many devices/data silos to collaboratively accomplish a learning task, which poses a unique challenge, namely heterogeneity (Li et al., 2020a). More specifically, the heterogeneity can be attributed to two major aspects: (i) One is from hardware specifications and the running environment of clients (called systemic heterogeneity), e.g., different CPU, RAM, battery life, dynamic network conditions, etc. Clients with limited computing and communication resources may fail to communicate with the server, i.e., be inactive, or return incomplete updates. These undesirable behaviors are proved to degrade the system's performance since they magnify the discrepancies between irregular models and the global model (Sahu et al., 2018; Chen et al., 2018).

Additionally, (ii) since clients are end users' devices, the issue of statistical heterogeneity naturally arises in FL, which refers to non-IID data, i.e., data are highly skewed, extremely imbalanced, and vary across clients (McMahan et al., 2017a). From a statistical perspective, it leads to distribution shifts, which raises the difficulties of model convergence. In reality, statistical heterogeneity is prevalent in recommender systems and personalized advertisement placement. For example, mobile phone users who read news articles may be interested in different news categories like politics, sports, or fashion; advertisement platforms might need to send different types of ads to different groups of customers.

A.2 Complete Proofs

A.2.1 Proof of Theorem 1

Equivalent Representation

For ease of the analysis, we introduce for each client k and each global round τ a sequence of virtual variables $\alpha^k_{\tau E}, \alpha^k_{\tau E+1}, ..., \alpha^k_{(\tau+1)E-1}$, where each $\alpha^k_t \in \{0,1\}$ and $\sum_{j=0}^{E} \alpha^k_{\tau E+j} = s^k_\tau$. If s^k_τ is a random variable, then α^k_t is also a random variable, and the distribution of α^k_t determines the distribution of s^k_τ. For example, if $\alpha^k_t \sim \text{Bernoulli}(p)$, then $s^k_\tau \sim \text{Bin}(E,p)$. In general, we do not assume any distributions and correlations of α^k_t, and thus the result is valid for any distributions.

Equation (4) and (5) and be rewritten as:

$$\omega^k_{\tau E+j+1} = \omega^k_{\tau E+j} - \eta_\tau g^k_{\tau E+j} \alpha^k_{\tau E+j}, \tag{12}$$

$$\omega^G_{(\tau+1)E} = \omega^G_{\tau E} + R^G_{(\tau-1)E} - \sum_{k=1}^{N} p^k_\tau \sum_{j=0}^{E} \eta_\tau g^k_{\tau E+j} \alpha^k_{\tau E+j}, \tag{13}$$

and ω^G_t is aggregated and synchronized when t is a multiple of E. In order to generalize the rule to arbitrary t, we define $\overline{\omega}_t$ such that $\overline{\omega}_0 = \omega^G_0$, and the residual R can be absorbed into gradient itself, so we have

$$\overline{\omega}_{\tau E+j+1} = \overline{\omega}_{\tau E+j} - \eta_\tau \sum_{k=1}^{N} p^k_\tau g^k_{\tau E+j} \alpha^k_{\tau E+j}, \tag{14}$$

Lemma 1 *For any τ, $\overline{\omega}_{\tau E} = \omega^G_{\tau E}$*

Proof 1 *By definition, we have $\overline{\omega}_0 = \omega^G_0$, suppose $\overline{\omega}_{\tau E} = \omega^G_{\tau E}$, then*

$$\begin{aligned} \overline{\omega}_{(\tau+1)E} &= \overline{\omega}_{(\tau+1)E-1} - \eta_\tau \sum_{k=1}^{N} p^k_\tau g^k_{(\tau+1)E-1} \alpha^k_{(\tau+1)E-1} \\ &= ... = \overline{\omega}_{\tau E} - \sum_{j=0}^{E-1} \eta_\tau \sum_{k=1}^{N} p^k_\tau g^k_{\tau E+j} \alpha^k_{\tau E+j} \\ &= \omega^G_{\tau E} - \sum_{k=1}^{N} p^k_\tau \sum_{j=0}^{E-1} \eta_\tau g^k_{\tau E+j} \alpha^k_{\tau E+j} = \omega^G_{(\tau+1)E}. \end{aligned} \tag{15}$$

We adopt $\overline{\omega}_t$ to denote the global weight in the following analysis.

Important Lemma

Lemma 2

$$\mathbb{E}_\xi || \sum_{k=1}^{N} p_\tau^k (g_t^k - \overline{g}_t^k)||^2 \leq \sum_{k=1}^{N} (p_\tau^k)^2 \sigma_k^2. \tag{16}$$

Proof 2

$$|| \sum_{k=1}^{N} p_\tau^k (g_t^k - \overline{g}_t^k)||^2 = \sum_{k=1}^{N} ||p_\tau^k (g_t^k - \overline{g}_t^k)||^2 + \sum_{i \neq k} ||p_\tau^k p_\tau^i (g_t^k - \overline{g}_t^k)(g_t^i - \overline{g}_t^i)||^2. \tag{17}$$

Since each client runs independently, the covariance

$$\mathbb{E}_\xi ||(g_t^k - \overline{g}_t^k)(g_t^i - \overline{g}_t^i)||^2 = 0, \tag{18}$$

and we have

$$\mathbb{E}_\xi || \sum_{k=1}^{N} p_\tau^k (g_t^k - \overline{g}_t^k)||^2 = \sum_{k=1}^{N} \mathbb{E}_\xi ||p_\tau^k (g_t^k - \overline{g}_t^k)||^2 \leq \sum_{k=1}^{N} (p_\tau^k)^2 \sigma_k^2. \tag{19}$$

Lemma 3

$$\mathbb{E}_\xi [\sum_{k=1}^{N} p_\tau^k ||\overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k||^2] \leq (E-1) G^2 \eta_\tau^2 (\sum_{k=1}^{N} p_\tau^k s_\tau^k + (\sum_{k=1}^{N} p_\tau^k - 2) + \sum_{k=1}^{N} \frac{(p_\tau^k)^2 s_\tau^k}{p^k}). \tag{20}$$

Proof 3 *Because* $\omega_{\tau E}^k = \overline{\omega}_{\tau E}$ *for all k, we have*

$$\begin{aligned} ||\overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k||^2 &= ||(\overline{\omega}_{\tau E+j} - \overline{\omega}_{\tau E}) - (\omega_{\tau E+j}^k - \overline{\omega}_{\tau E})||^2 \\ &= ||\overline{\omega}_{\tau E+j} - \overline{\omega}_{\tau E}||^2 - 2 < \overline{\omega}_{\tau E+j} - \overline{\omega}_{\tau E}, \omega_{\tau E+j}^k - \overline{\omega}_{\tau E} > \\ &+ ||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E}||^2. \end{aligned} \tag{21}$$

According to Equation (12) and (14), we have

$$\begin{aligned} \sum_{k=1}^{N} p_\tau^k \omega_{\tau E+j}^k &= \sum_{k=1}^{N} p_\tau^k \omega_{\tau E+j-1}^k - \eta_\tau \sum_{k=1}^{N} p_\tau^k g_{\tau E+j-1}^k \alpha_{\tau E+j-1}^k \\ &= \sum_{k=1}^{N} p_\tau^k \omega_{\tau E+j-1}^k + \overline{\omega}_{\tau E+j} - \overline{\omega}_{\tau E+j-1} \\ &= ... = \sum_{k=1}^{N} p_\tau^k \omega_{\tau E}^k + \overline{\omega}_{\tau E+j} - \overline{\omega}_{\tau E}. \end{aligned} \tag{22}$$

Thus,

$$\begin{aligned} &-2 \sum_{k=1}^{N} p_\tau^k < \overline{\omega}_{\tau E+j} - \overline{\omega}_{\tau E}, \omega_{\tau E+j}^k - \overline{\omega}_{\tau E} > \\ &= -2 < \overline{\omega}_{\tau E+j} - \overline{\omega}_{\tau E}, \sum_{k=1}^{N} p_\tau^k \omega_{\tau E}^k + \overline{\omega}_{\tau E+j} - \overline{\omega}_{\tau E} - \sum_{k=1}^{N} p_\tau^k \overline{\omega}_{\tau E} > \\ &= -2 ||\overline{\omega}_{\tau E+j} - \overline{\omega}_{\tau E}||^2, \end{aligned} \tag{23}$$

and

$$\sum_{k=1}^{N} p_\tau^k ||\overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k||^2 = (\sum_{k=1}^{N} p_\tau^k - 2) ||\overline{\omega}_{\tau E+j} - \overline{\omega}_{\tau E}||^2 + \sum_{k=1}^{N} p_\tau^k ||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E}||^2. \tag{24}$$

We also have

$$\begin{aligned}
||\overline{\omega}_{\tau E+j} - \overline{\omega}_{\tau E}||^2 &= ||\sum_{i=0}^{j-1} \eta_\tau \sum_{k=1}^{N} p_\tau^k g_{\tau E+i}^k \alpha_{\tau E+i}^k||^2 \\
&= ||\eta_\tau \sum_{k=1}^{N} p_\tau^k (\sum_{i=0}^{j-1} g_{\tau E+i}^k \alpha_{\tau E+i}^k)||^2 \\
&= \eta_\tau^2 || \sum_{k=1}^{N} p^k (\frac{p_\tau^k}{p^k} \sum_{i=0}^{j-1} g_{\tau E+i}^k \alpha_{\tau E+i}^k)||^2 \\
&\leq \eta_\tau^2 \sum_{k=1}^{N} \frac{(p_\tau^k)^2}{p^k} || \sum_{i=0}^{j-1} g_{\tau E+i}^k \alpha_{\tau E+i}^k||^2,
\end{aligned} \tag{25}$$

where

$$\begin{aligned}
||\sum_{i=0}^{j-1} g_{\tau E+i}^k \alpha_{\tau E+i}^k||^2 &= \sum_{i=0}^{j-1} ||g_{\tau E+i}^k \alpha_{\tau E+i}^k||^2 + 2 \sum_{p<q} < g_{\tau E+p}^k \alpha_{\tau E+p}^k, g_{\tau E+q}^k \alpha_{\tau E+q}^k > \\
&\leq \sum_{i=0}^{j-1} ||g_{\tau E+i}^k \alpha_{\tau E+i}^k||^2 + 2 \sum_{p<q} ||g_{\tau E+p}^k \alpha_{\tau E+p}^k|| ||g_{\tau E+q}^k \alpha_{\tau E+q}^k|| \\
&\leq \sum_{i=0}^{j-1} ||g_{\tau E+i}^k \alpha_{\tau E+i}^k||^2 + \sum_{p<q} (||g_{\tau E+p}^k \alpha_{\tau E+p}^k||^2 + ||g_{\tau E+q}^k \alpha_{\tau E+q}^k||^2) \\
&= j \sum_{i=0}^{j-1} ||g_{\tau E+i}^k \alpha_{\tau E+i}^k||^2.
\end{aligned} \tag{26}$$

As a result, we have

$$\mathbb{E}_\xi || \sum_{i=0}^{j-1} g_{\tau E+i}^k \alpha_{\tau E+i}^k||^2 \leq jG^2 \sum_{i=0}^{j-1} \alpha_{\tau E+i}^k \leq (E-1)G^2 s_\tau^k. \tag{27}$$

Then plug Equation (27) into Equation (25),

$$\mathbb{E}_\xi ||\overline{\omega}_{\tau E+j} - \overline{\omega}_{\tau E}||^2 \leq (E-1)G^2 \eta_\tau^2 \sum_{k=1}^{N} \frac{(p_\tau^k)^2}{p^k} s_\tau^k. \tag{28}$$

Similarly, we have

$$\mathbb{E}_\xi \sum_{k=1}^{N} p_\tau^k ||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j}||^2 = \mathbb{E}_\xi \sum_{k=1}^{N} p_\tau^k ||\eta_\tau \sum_{i=0}^{j-1} g_{\tau E+i}^k \alpha_{\tau E+i}^k||^2 \leq (E-1)G^2 \eta_\tau^2 \sum_{k=1}^{N} p_\tau^k s_\tau^k. \tag{29}$$

Substitute Equation (24) with Equation (28) and (29),

$$\mathbb{E}_\xi [\sum_{k=1}^{N} p_\tau^k ||\overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k||^2] \leq (E-1)G^2 \eta_\tau^2 (\sum_{k=1}^{N} p_\tau^k s_\tau^k + (\sum_{k=1}^{N} p_\tau^k - 2) + \sum_{k=1}^{N} \frac{(p_\tau^k)^2}{p^k} s_\tau^k). \tag{30}$$

Bounding $||\overline{\omega}_{\tau E+j+1} - \omega^*||^2$

$$\begin{aligned}
||\overline{\omega}_{\tau E+j+1} - \omega^*||^2 = {} & ||\overline{\omega}_{\tau E+j} - \eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k g_{\tau E+j}^k - \omega^* - \eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k \\
& + \eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k ||^2 \\
= {} & \underbrace{||\overline{\omega}_{\tau E+j} - \omega^* - \eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k ||^2}_{A_1} + \eta_\tau^2 || \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\overline{g}_{\tau E+j}^k - g_{\tau E+j}^k)||^2 \qquad (31) \\
& + \underbrace{2\eta_\tau < \overline{\omega}_{\tau E+j} - \omega^* - \eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k, \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\overline{g}_{\tau E+j}^k - g_{\tau E+j}^k) >}_{A_2}.
\end{aligned}$$

Since $\mathbb{E}_\xi[g_{\tau E+j}^k] = \overline{g}_{\tau E+j}^k$, we have $\mathbb{E}_\xi[A_2] = 0$, and

$$\begin{aligned}
A_1 & = ||\overline{\omega}_{\tau E+j} - \omega^* - \eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k ||^2 \\
& = ||\overline{\omega}_{\tau E+j} - \omega^*||^2 \underbrace{-2\eta_\tau < \overline{\omega}_{\tau E+j} - \omega^*, \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k >}_{B_1} \qquad (32) \\
& \quad \underbrace{+ \eta_\tau^2 || \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k ||^2}_{B_2}.
\end{aligned}$$

Because $\widetilde{F}_k$ is L-smooth, we have

$$||\alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k||^2 \leq 2L(\widetilde{F}_k(\omega_{\tau E+j}^k) - \widetilde{F}_k^*)\alpha_{\tau E+j}^k, \qquad (33)$$

and based on the convexity of l_2 norm, it can be derived that

$$\begin{aligned}
B_2 & = \eta_\tau^2 || \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k ||^2 = \eta_\tau^2 || \sum_{k=1}^{N} p^k (\frac{p_\tau^k}{p^k} \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k)||^2 \\
& \leq \eta_\tau^2 \sum_{k=1}^{N} \frac{(p_\tau^k)^2}{p^k} ||\alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k||^2 \leq 2L\theta\eta_\tau^2 \sum_{k=1}^{N} p_\tau^k (\widetilde{F}_k(\omega_{\tau E+j}^k) - \widetilde{F}_k^*)\alpha_{\tau E+j}^k.
\end{aligned} \qquad (34)$$

$$\begin{aligned}
B_1 = -2\eta_\tau < \overline{\omega}_{\tau E+j} - \omega^*, \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k > & = -2\eta_\tau \sum_{k=1}^{N} p_\tau^k < \overline{\omega}_{\tau E+j} - \omega^*, \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k > \\
= -2\eta_\tau \sum_{k=1}^{N} p_\tau^k < \overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k, \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k > & -2\eta_\tau \sum_{k=1}^{N} p_\tau^k < \omega_{\tau E+j}^k - \omega^*, \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k >,
\end{aligned} \qquad (35)$$

where

$$\begin{aligned}
-2 < \overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k, \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k > & \leq 2|| < \overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k, \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k > || \\
& \leq 2\alpha_{\tau E+j}^k ||\overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k|| || \overline{g}_{\tau E+j}^k|| \\
& \leq (\frac{1}{\eta_\tau} ||\overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k||^2 + \eta_\tau ||\overline{g}_{\tau E+j}^k||^2)\alpha_{\tau E+j}^k.
\end{aligned} \qquad (36)$$

Because $\widetilde{F}_k$ is μ-strong convex, we have

$$< \omega_{\tau E+j}^k - \omega^*, \alpha_{\tau E+j}^k \overline{g}_{\tau E+j}^k > \geq (\widetilde{F}_k(\omega_{\tau E+j}^k) - \widetilde{F}_k^* + \frac{\mu}{2} ||\omega_{\tau E+j}^k - \omega^*||^2)\alpha_{\tau E+j}^k. \qquad (37)$$

Substitute Equation (35) with Equation (36)(37), we have

$$B_1 \leq \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (||\overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k||^2 + \eta_\tau^2 ||\overline{g}_{\tau E+j}^k||^2 - 2\eta_\tau (\widetilde{F}_k(\omega_{\tau E+j}^k) - \widetilde{F}_k^* + \frac{\mu}{2}||\omega_{\tau E+j}^k - \omega^*||^2)), \tag{38}$$

and plug Equation (34)(38) to Equation (32),

$$\begin{aligned} A_1 \leq\ & ||\overline{\omega}_{\tau E+j} - \omega^*||^2 + 2L\theta\eta_\tau^2 \sum_{k=1}^{N} p_\tau^k (\widetilde{F}_k(\omega_{\tau E+j}^k) - \widetilde{F}_k^*)\alpha_{\tau E+j}^k \\ & + \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (||\overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k||^2 + \eta_\tau^2 ||\overline{g}_{\tau E+j}^k||^2 - 2\eta_\tau (\widetilde{F}_k(\omega_{\tau E+j}^k) - \widetilde{F}_k^* + \frac{\mu}{2}||\omega_{\tau E+j}^k - \omega^*||^2)) \\ \leq\ & ||\overline{\omega}_{\tau E+j} - \omega^*||^2 - \mu\eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k ||\omega_{\tau E+j}^k - \omega^*||^2 + \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k ||\overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k||^2 \\ & + \underbrace{2(1+\theta)L\eta_\tau^2 \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\omega_{\tau E+j}^k) - \widetilde{F}_k^*) - 2\eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\omega_{\tau E+j}^k) - \widetilde{F}_k^*)\,.}_{C_1} \end{aligned} \tag{39}$$

$$\begin{aligned} ||\omega_{\tau E+j}^k - \omega^*||^2 &= ||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j} + \overline{\omega}_{\tau E+j} - \omega^*||^2 \\ &= ||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j}||^2 + ||\overline{\omega}_{\tau E+j} - \omega^*||^2 + 2 < \omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j}, \overline{\omega}_{\tau E+j} - \omega^* > \\ &\geq ||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j}||^2 + ||\overline{\omega}_{\tau E+j} - \omega^*||^2 - 2||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j}|| ||\overline{\omega}_{\tau E+j} - \omega^*|| \\ &\geq ||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j}||^2 + ||\overline{\omega}_{\tau E+j} - \omega^*||^2 - (2||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j}||^2 + \frac{1}{2}||\overline{\omega}_{\tau E+j} - \omega^*||^2) \\ &= \frac{1}{2}||\overline{\omega}_{\tau E+j} - \omega^*||^2 - ||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j}||^2. \end{aligned} \tag{40}$$

Thus,

$$A_1 \leq (1 - \frac{1}{2}\mu\eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k)||\overline{\omega}_{\tau E+j} - \omega^*||^2 + (1 + \mu\eta_\tau) \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k)||\overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k||^2 + C_1. \tag{41}$$

Let $\gamma_\tau = 2\eta_\tau(1 - (1+\theta)L\eta_\tau)$, and assume $\eta_\tau \leq \frac{1}{2(1+\theta)L}$, we have

$$\begin{aligned} C_1 &= -2\eta_\tau(1 - (1+\theta)L\eta_\tau) \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\omega_{\tau E+j}^k) - \widetilde{F}_k^*) + 2\eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\omega^*) - \widetilde{F}_k^*) \\ &= -\gamma_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\omega_{\tau E+j}^k) - \widetilde{F}_k^* + \widetilde{F}_k(\omega^*) - \widetilde{F}_k(\omega^*)) + 2\eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\omega^*) - \widetilde{F}_k^*) \\ &= -\gamma_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\omega_{\tau E+j}^k) - \widetilde{F}_k(\omega^*)) + (2\eta_\tau - \gamma_\tau) \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\omega^*) - \widetilde{F}_k^*) \\ &\leq \underbrace{-\gamma_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\omega_{\tau E+j}^k) - \widetilde{F}_k(\omega^*))}_{C_2} + 2(1+\theta)L\eta_\tau^2 \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k D_k. \end{aligned} \tag{42}$$

Then, we bound C_2,

$$\begin{aligned}
&\sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\omega_{\tau E+j}^k) - \widetilde{F}_k(\omega^*)) = \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\omega_{\tau E+j}^k) - \widetilde{F}_k(\overline{\omega}_{\tau E+j}))\\
&+ \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\overline{\omega}_{\tau E+j}) - \widetilde{F}_k(\omega^*))\\
&\geq \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k < \nabla \widetilde{F}_k(\overline{\omega}_{\tau E+j}), \omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j} > + \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\overline{\omega}_{\tau E+j}) - \widetilde{F}_k(\omega^*))\\
&\geq - \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k ||\nabla \widetilde{F}_k(\overline{\omega}_{\tau E+j})|| ||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j}|| + \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\overline{\omega}_{\tau E+j}) - \widetilde{F}_k(\omega^*))\\
&\geq -\frac{1}{2} \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\eta_\tau ||\nabla \widetilde{F}_k(\overline{\omega}_{\tau E+j})||^2 + \frac{1}{\eta_\tau} ||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j}||^2) + \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\overline{\omega}_{\tau E+j}) - \widetilde{F}_k(\omega^*))\\
&\geq - \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\eta_\tau L(\widetilde{F}_k(\overline{\omega}_{\tau E+j}) - \widetilde{F}_k^*) + \frac{1}{2\eta_\tau} ||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j}||^2) + \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\overline{\omega}_{\tau E+j}) - \widetilde{F}_k(\omega^*)).
\end{aligned} \tag{43}$$

Thus, we have

$$\begin{aligned}
C_1 &\leq \gamma_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\eta_\tau L(\widetilde{F}_k(\overline{\omega}_{\tau E+j}) - \widetilde{F}_k^*) + \frac{1}{2\eta_\tau} ||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j}||^2)\\
&- \gamma_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\overline{\omega}_{\tau E+j}) - \widetilde{F}_k(\omega^*)) + 2(1+\theta) L \eta_\tau^2 \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k D_k\\
&= \gamma_\tau (\eta_\tau L - 1) \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\overline{\omega}_{\tau E+j}) - \widetilde{F}_k(\omega^*)) + \frac{\gamma_\tau}{2\eta_\tau} \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k ||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j}||^2\\
&+ 2(1+\theta) L \eta_\tau^2 \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k D_k + \gamma_\tau \eta_\tau L \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k D_k\\
&\leq \gamma_\tau (\eta_\tau L - 1) \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\overline{\omega}_{\tau E+j}) - \widetilde{F}_k(\omega^*)) + \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k ||\omega_{\tau E+j}^k - \overline{\omega}_{\tau E+j}||^2\\
&+ 2(2+\theta) L \eta_\tau^2 \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k D_k.
\end{aligned} \tag{44}$$

Substitute Equation (41) with Equation (44),

$$\begin{aligned}
A_1 &\leq ||\overline{\omega}_{\tau E+j} - \omega^*||^2 - \mu \eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k ||\omega_{\tau E+j}^k - \omega^*||^2 + 2 \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k ||\overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k||^2\\
&+ 2(2+\theta) L \eta_\tau^2 \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k D_k + \gamma_\tau (\eta_\tau L - 1) \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\overline{\omega}_{\tau E+j}) - \widetilde{F}_k(\omega^*)),
\end{aligned} \tag{45}$$

and plug Equation (45) to Equation (31), we have

$$\begin{aligned}
||\overline{\omega}_{\tau E+j+1} - \omega^*||^2 &\leq (1 - \frac{1}{2} \mu \eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k) ||\overline{\omega}_{\tau E+j} - \omega^*||^2\\
&+ \eta_\tau^2 || \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\overline{g}_{\tau E+j}^k - g_{\tau E+j}^k)||^2 + (2 + \mu \eta_\tau) \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k ||\overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k||^2\\
&+ 2(2+\theta) L \eta_\tau^2 \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k D_k + \gamma_\tau (1 - \eta_\tau L) \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\omega^*) - \widetilde{F}_k(\overline{\omega}_{\tau E+j})).
\end{aligned} \tag{46}$$

Let

$$
\begin{aligned}
B_{\tau E+j} = (2 + \frac{\mu}{2(1+\theta)L}) \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k ||\overline{\omega}_{\tau E+j} - \omega_{\tau E+j}^k||^2 \\
+ ||\sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\overline{g}_{\tau E+j}^k - g_{\tau E+j}^k)||^2 + 2(2+\theta)L \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k D_k,
\end{aligned}
\tag{47}
$$

we have

$$
\begin{aligned}
||\overline{\omega}_{\tau E+j+1} - \omega^*||^2 \leq (1 - \frac{1}{2}\mu\eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k)||\overline{\omega}_{\tau E+j} - \omega^*||^2 + \eta_\tau^2 B_{\tau E+j} \\
+ 2\eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k (\widetilde{F}_k(\omega^*) - \widetilde{F}_k(\overline{\omega}_{\tau E+j})).
\end{aligned}
\tag{48}
$$

Apply the lemma, we have

$$
\begin{aligned}
\mathbb{E}_\xi[B_{\tau E+j}] \leq \sum_{k=1}^{N} (p_\tau^k)^2 \alpha_{\tau E+j}^k \sigma_k^2 + 2(2+\theta)L \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k D_k \\
+ (2 + \frac{\mu}{2(1+\theta)L})(E-1)G^2(\sum_{k=1}^{N} p_\tau^k s_\tau^k + (\sum_{k=1}^{N} p_\tau^k - 2) + \sum_{k=1}^{N} \frac{(p_\tau^k)^2 s_\tau^k}{p^k}).
\end{aligned}
\tag{49}
$$

Let $\Delta_{\tau E+j} = ||\overline{\omega}_{\tau E+j} - \omega^*||^2$, and $\overline{\Delta}_{\tau E+j} = \mathbb{E}[\Delta_{\tau E+j}]$, where the expectation is taken over all random variables up to $\tau E + j$

Bounding $||\overline{\omega}_{\tau E} - \omega^*||^2$

Summing Δ_t from $t = \tau E$ to $(\tau+1)E$, we have

$$
\sum_{j=1}^{E} \Delta_{\tau E+j} \leq \sum_{j=0}^{E-1}(1 - \frac{1}{2}\mu\eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k)\Delta_{\tau E+j} + \eta_\tau^2 B_\tau + 2\eta_\tau \sum_{k=1}^{N} p_\tau^k s_\tau^k (\widetilde{F}_k(\omega^*) - \widetilde{F}_k(\overline{\omega}_{\tau E+j})),
\tag{50}
$$

where $B_\tau = \sum_{j=0}^{E-1} B_{\tau E+j}$, and $\overline{\omega}_{\tau E+j} = argmin_{\omega_{\tau E+j}} \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k \widetilde{F}_k(\overline{\omega}_{\tau E+j})$.

Equation (50) can be reorganized to

$$
\Delta_{(\tau+1)E} \leq \Delta_{\tau E} - \frac{1}{2}\mu\eta_\tau \sum_{j=0}^{E-1}\sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k \Delta_{\tau E+j} + \eta_\tau^2 B_\tau + 2\eta_\tau \sum_{k=1}^{N} p_\tau^k s_\tau^k (\widetilde{F}_k(\omega^*) - \widetilde{F}_k(\overline{\omega}_{\tau E+j})).
\tag{51}
$$

Then, we seek to find a lower bound for $\Delta_{\tau E+j}$.

$$
\begin{aligned}
\sqrt{\Delta_{\tau E+j+1}} &= ||\overline{\omega}_{\tau E+j+1} - \omega^*|| = ||\overline{\omega}_{\tau E+j+1} - \overline{\omega}_{\tau E+j} + \overline{\omega}_{\tau E+j} - \omega^*|| \\
&\leq ||\overline{\omega}_{\tau E+j+1} - \overline{\omega}_{\tau E+j}|| + \sqrt{\Delta_{\tau E+j}} \\
&= ||\eta_\tau \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k g_{\tau E+j}^k|| + \sqrt{\Delta_{\tau E+j}}.
\end{aligned}
\tag{52}
$$

Let $h_{\tau E+j} = ||\sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k g_{\tau E+j}^k||$, we have

$$
\begin{aligned}
\sqrt{\Delta_{(\tau+1)E}} &\leq \sqrt{\Delta_{(\tau+1)E-1}} + \eta_\tau h_{(\tau+1)E-1} \\
&\leq ... \leq \sqrt{\Delta_{\tau E+j}} + \sum_{i=j}^{E-1} \eta_\tau h_{\tau E+j},
\end{aligned}
\tag{53}
$$

$$
\begin{aligned}
\Delta_{(\tau+1)E} &\leq \Delta_{\tau E+j} + 2\sqrt{\Delta_{\tau E+j}}(\sum_{i=j}^{E-1} \eta_\tau h_{\tau E+j}) + (\sum_{i=j}^{E-1} \eta_\tau h_{\tau E+j})^2 \\
&\leq 2\Delta_{\tau E+j} + 2(\sum_{i=j}^{E-1} \eta_\tau h_{\tau E+j})^2,
\end{aligned}
\tag{54}
$$

and

$$\Delta_{\tau E+j} \geq \frac{1}{2}\Delta_{(\tau+1)E} - (\sum_{i=j}^{E-1} \eta_\tau h_{\tau E+j})^2 \geq \frac{1}{2}\Delta_{(\tau+1)E} - (\sum_{j=0}^{E-1} \eta_\tau h_{\tau E+j})^2. \tag{55}$$

Plug Equation (55) to Equation (51), we have

$$\begin{aligned}(1 + \frac{1}{4}\mu\eta_\tau \sum_{k=1}^{N} p_\tau^k s_\tau^k)\Delta_{(\tau+1)E} \leq \Delta_{\tau E} + \frac{1}{2}\mu\eta_\tau^3 \sum_{k=1}^{N} p_\tau^k s_\tau^k (\sum_{j=0}^{E-1} h_{\tau E+j})^2 + \eta_\tau^2 B_\tau \\ + 2\eta_\tau \sum_{k=1}^{N} p_\tau^k s_\tau^k (\widetilde{F}_k(\omega^*) - \widetilde{F}_k(\overline{\omega}_{\tau E+j}))\end{aligned} \tag{56}$$

Apply Lemma 2 and Lemma 3, we have

$$\begin{aligned}\mathbb{E}_\xi[h_{\tau E+j}^2] &= \mathbb{E}_\xi || \sum_{k=1}^{N} p_\tau^k \alpha_{\tau E+j}^k g_{\tau E+j}^k ||^2 \\ &\leq \sum_{k=1}^{N} \frac{(p_\tau^k)^2}{p^k} \mathbb{E}_\xi ||\alpha_{\tau E+j}^k g_{\tau E+j}^k||^2 \leq \sum_{k=1}^{N} \frac{(p_\tau^k)^2}{p^k} G^2 \alpha_{\tau E+j}^k,\end{aligned} \tag{57}$$

$$\mathbb{E}_\xi[(\sum_{j=0}^{E-1} h_{\tau E+j})^2] \leq \mathbb{E}_\xi[E \sum_{j=0}^{E-1} h_{\tau E+j}^2] \leq EG^2 \sum_{k=1}^{N} \frac{(p_\tau^k)^2}{p^k} s_\tau^k, \tag{58}$$

and

$$\begin{aligned}\mathbb{E}_\xi[B_\tau] = \sum_{j=0}^{E-1} \mathbb{E}_\xi[B_{\tau E+j}] = \sum_{k=1}^{N} (p_\tau^k)^2 s_\tau^k \sigma_k^2 + 2(2+\theta)L \sum_{k=1}^{N} p_\tau^k s_\tau^k D_k \\ + (2 + \frac{\mu}{2(1+\theta)L})E(E-1)G^2(\sum_{k=1}^{N} p_\tau^k s_\tau^k + \theta(\sum_{k=1}^{N} p_\tau^k - 2) + \sum_{k=1}^{N} p_\tau^k s_\tau^k).\end{aligned} \tag{59}$$

Let $\overline{\Delta}_{\tau E+j} = \mathbb{E}_\xi[\Delta_{\tau E+j}]$, $\overline{B}_\tau = \mathbb{E}_\xi[B_\tau]$, $\overline{H}_\tau = \mathbb{E}_\xi[(\sum_{j=0}^{E-1} h_{\tau E+j})^2]$, we have

$$\begin{aligned}(1 + \frac{1}{4}\mu\eta_\tau \sum_{k=1}^{N} p_\tau^k s_\tau^k)\overline{\Delta}_{(\tau+1)E} \leq \overline{\Delta}_{\tau E} + \frac{1}{2}\mu\eta_\tau^3 \sum_{k=1}^{N} p_\tau^k s_\tau^k \overline{H}_\tau + \eta_\tau^2 \overline{B}_\tau \\ + 2\eta_\tau \mathbb{E}_\xi \sum_{k=1}^{N} p_\tau^k s_\tau^k (\widetilde{F}_k(\omega^*) - \widetilde{F}_k(\overline{\omega}_{\tau E+j})).\end{aligned} \tag{60}$$

Let $r_\tau = 0$ indicate the event that for all k, $\mathbb{E}[p_\tau^k s_\tau^k] = c_\tau p^k$ for some constant c_τ that does not depend on k, otherwise $r_\tau = 1$. Note that if $r_\tau = 0$, then $\sum_{k=1}^{N} p_\tau^k s_\tau^k (\widetilde{F}_k(\omega^*) - \widetilde{F}_k(\overline{\omega}_{\tau E+j})) = c_\tau(\widetilde{F}_k(\omega^*) - \widetilde{F}_k(\overline{\omega}_{\tau E+j})) \leq 0$. We have

$$\begin{aligned}\sum_{k=1}^{N} p_\tau^k s_\tau^k (\widetilde{F}_k(\omega^*) - \widetilde{F}_k(\overline{\omega}_{\tau E+j})) &= \sum_{k=1}^{N} p_\tau^k s_\tau^k (\widetilde{F}_k(\omega^*) - \widetilde{F}_k^* + \widetilde{F}_k^* - \widetilde{F}_k(\overline{\omega}_{\tau E+j})) \\ &\leq \sum_{k=1}^{N} p_\tau^k s_\tau^k D_k.\end{aligned} \tag{61}$$

Thus,

$$\sum_{k=1}^{N} p_\tau^k s_\tau^k (\widetilde{F}_k(\omega^*) - \widetilde{F}_k(\overline{\omega}_{\tau E+j})) \leq r_\tau \sum_{k=1}^{N} p_\tau^k s_\tau^k D_k. \tag{62}$$

Assuming $\eta_\tau \leq \frac{4}{\mu E\theta} \leq \frac{4}{\mu\sum_{k=1}^N p_\tau^k s_\tau^k}$, Equation (60) can be written as

$$\begin{aligned}\overline{\Delta}_{(\tau+1)E} &\leq (1 - \frac{\frac{1}{4}\mu\eta_\tau\sum_{k=1}^N p_\tau^k s_\tau^k}{1+\frac{1}{4}\mu\eta_\tau\sum_{k=1}^N p_\tau^k s_\tau^k})\overline{\Delta}_{\tau E} + 2\eta_\tau^2\overline{H}_\tau + \eta_\tau^2\overline{B}_\tau \\ &+ 2\eta_\tau r_\tau\sum_{k=1}^N p_\tau^k s_\tau^k D_k \\ &\leq (1-\frac{1}{8}\mu\eta_\tau\sum_{k=1}^N p_\tau^k s_\tau^k)\overline{\Delta}_{\tau E} + \eta_\tau^2(\overline{B}_\tau + 2\overline{H}_\tau) + 2\eta_\tau r_\tau\sum_{k=1}^N p_\tau^k s_\tau^k D_k,\end{aligned} \tag{63}$$

and take expectation over p_τ^k and s_τ^k, we have

$$\mathbb{E}[\overline{\Delta}_{(\tau+1)E}] \leq (1-\frac{1}{8}\mu\eta_\tau\mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k])\overline{\Delta}_{\tau E} + \eta_\tau^2\mathbb{E}[\overline{B}_\tau + 2\overline{H}_\tau] + 2\eta_\tau r_\tau\sum_{k=1}^N\mathbb{E}[p_\tau^k s_\tau^k]D_k. \tag{64}$$

Proof of Theorem 1

The Theorem is proved by induction. Let $\eta_\tau = \frac{16E}{\mu((\tau+1)E+\gamma)\mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k]}$. Initially, $\frac{C_0}{\gamma^2} \geq \mathbb{E}[\overline{\Delta}_0]$. Suppose $\mathbb{E}[\overline{\Delta}_{\tau E}] \leq \frac{C_\tau}{(\tau E+\gamma)^2} + \frac{H_\tau J}{\tau E+\gamma}$, then

$$\begin{aligned}\mathbb{E}[\overline{\Delta}_{(\tau+1)E}] &\leq \frac{\tau E+\gamma-E}{(\tau+1)E+\gamma}(\frac{C_\tau}{(\tau E+\gamma)^2} + \frac{H_\tau J}{\tau E+\gamma}) + \frac{(16E)^2\mathbb{E}[\overline{B}_\tau + 2\overline{H}_\tau]}{(\mu\mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k])^2((\tau+1)E+\gamma)^2} \\ &+ \frac{r_\tau J}{(\tau+1)E+\gamma} \\ &\leq \frac{\tau E+\gamma-E}{(\tau E+\gamma)^2-E^2}\frac{C_\tau}{(\tau+1)E+\gamma} + \frac{(\tau E+\gamma-E)H_\tau J}{(\tau E+\gamma)^2-E^2} \\ &+ \frac{(16E)^2\mathbb{E}[\overline{B}_\tau + 2\overline{H}_\tau]}{(\mu\mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k])^2((\tau+1)E+\gamma)^2} + \frac{r_\tau J}{(\tau+1)E+\gamma} \\ &\leq \frac{C_\tau}{((\tau+1)E+\gamma)^2} + \frac{H_\tau J}{(\tau+1)E+\gamma} + \frac{(16E)^2\mathbb{E}[\overline{B}_\tau + 2\overline{H}_\tau]}{(\mu\mathbb{E}[\sum_{k=1}^N p_\tau^k s_\tau^k])^2((\tau+1)E+\gamma)^2} \\ &+ \frac{r_\tau J}{(\tau+1)E+\gamma} \\ &= \frac{C_{\tau+1}}{((\tau+1)E+\gamma)^2} + \frac{H_{\tau+1}J}{(\tau+1)E+\gamma}.\end{aligned} \tag{65}$$

Thus, $\overline{\Delta}_{(\tau+1)E} \leq \frac{C_{\tau+1}}{((\tau+1)E+\gamma)^2} + \frac{H_{\tau+1}J}{(\tau+1)E+\gamma}$.

A.2.2 Proof of Theorem 2

Client Departure

Let $\hat{n} = n - n_a$, we have

$$\begin{aligned}
||\omega^* - \hat{\omega}^*||^2 &\leq \frac{4}{\mu^2}||\nabla F(\hat{\omega}^*)||^2 = \frac{4}{\mu^2}||\nabla F(\hat{\omega}^*) - \nabla \hat{F}(\hat{\omega}^*)||^2 \\
&= \frac{4}{\mu^2}||\sum_{k \neq a}(p^k - \hat{p}^k)\nabla \widetilde{F}_k(\hat{\omega}^*) + p^a \nabla \widetilde{F}_a(\hat{\omega}^*)||^2 \\
&= \frac{4}{\mu^2}||\sum_{k \neq a}(\frac{n_k}{n} - \frac{n_k}{n - n_a})\nabla \widetilde{F}_k(\hat{\omega}^*) + p^a \nabla \widetilde{F}_a(\hat{\omega}^*)||^2 \\
&= \frac{4}{\mu^2}|| - \sum_{k \neq a} \frac{n_a n_k}{n(n - n_a)}\nabla \widetilde{F}_k(\hat{\omega}^*) + p^a \nabla \widetilde{F}_a(\hat{\omega}^*)||^2 \\
&= \frac{4}{\mu^2}|| - p^a \sum_{k \neq a} \hat{p}^k \nabla \widetilde{F}_k(\hat{\omega}^*) + p^a \nabla \widetilde{F}_a(\hat{\omega}^*)||^2 = \frac{4(p^a)^2}{\mu^2}||\nabla \widetilde{F}_a(\hat{\omega}^*)||^2 \\
&\leq \frac{8Ln_a^2\hat{D}_a}{\mu^2 n^2}.
\end{aligned} \tag{66}$$

Client Arrival

Let $\hat{n} = n + n_a$, we have

$$\begin{aligned}
||\omega^* - \hat{\omega}^*||^2 &\leq \frac{4}{\mu^2}||\nabla F(\omega^*)||^2 = \frac{4}{\mu^2}||\nabla \hat{F}(\omega^*) - \nabla F(\omega^*)||^2 \\
&= ... = \frac{4}{\mu^2}|| - \hat{p}^a \sum_{k \neq a} p^k \nabla \widetilde{F}_k(\omega^*) + \hat{p}^a \nabla \widetilde{F}_a(\omega^*)||^2 \\
&\leq \frac{8Ln_a^2\hat{D}_a}{\mu^2(n + n_a)^2}.
\end{aligned} \tag{67}$$

A.3 Datasets and Corresponding Models

In this section, we describe details on datasets and corresponding models used in the experiments.

MNIST It contains handwritten digits which have been size-normalized and centered in a fixed-size image. For the IID setting, we split the training data randomly into equally sized shards and assign81one shard to everyone of the clients. For the non-IID (m) setting, we assign every client samples from exactly m classes of the dataset. The data splits are non-overlapping and balanced, such that every client ends up with the same number of data points. The corresponding model is multinomial logistic regression, where the input is a flattened image and the output is a class label.

Shakespeare The dataset is from *The Complete Works of William Shakespeare*. We construct a participant dataset for each speaking role and sample 132 speaking roles with at least five lines. A stacked character-level LSTM language model is used, which after reading each character in a line, to predict the next character. The model takes a series of characters as input and embeds each of these into a learned 8 dimensional space. The embedded characters are then processed through 2 LSTM layers, each with 256 nodes. Finally the output of the second LSTM layer is sent to a softmax output layer with one node per character.

Sentiment140 The dataset is collected from tweets. It is a binary classification task that classifies the sentiment of the text. The input is a 25-word sequence and embeds each word into a 300-dimensional space using Glove. The output is a binary label after two LSTM layers and one densely-connected layer.

EMNIST The EMNIST dataset is a set of handwritten character digits derived from the NIST Special Database 19 and converted to a 28×28 pixel image format and dataset structure that directly matcheshe MNIST database. For the IID setting, we split the training data randomly into equally sized shards and assign one shard to everyone of the clients. For the non-IID (m) setting, we assign every client samples from exactly m classes of the dataset. The data splits are non-overlapping and balanced, such that every client ends up with the same number of data points. The corresponding model is Resnet-50, where the input is a flattened image and the output is a class label.

CIFAR100 It consists of colour images in 100 classes, For the IID partition, we split the training data randomly into equally sized shards and assign one shard to everyone of the clients. For the non-IID (m) setting, we partition the

Table 5: Statistics of datasets. The number of devices, the number of samples, the mean and the standard deviation of data samples on each device are summarized.

Dataset	# Devices	# Samples	Mean	SD
MNIST	100	58,254	583	146
CIFAR100	100	59,137	591	32
Shakespeare	132	359,016	2,719	204
Sentiment140	1,503	90,110	60	41
EMNIST	500	131,600	263	93

training data by sorting the dataset by labels, and we assign every client samples from exactly m classes of the dataset. The VGG11 model is a simplified version of the original VGG11 architecture described in Simonyan & Zisserman (2014), where all dropout and batch normalization layers are removed and the number of convolutional filters and size of all fully-connected layers is reduced by a factor of 2.

The statistics of datasets are summarized in Table 5.

A.4 Hyperparameters setting

The dataset is split into 80% for training, 10% for testing and 10% for validation. Each experiment is conducted 50 times and reports the mean and the standard deviation. In each round, we randomly sample 10 participants.

A.5 Experiments on Different Base Models

We report experiments on different base models (ResNet50 on CIFAR100 and VGG11 on EMNIST datasets) and summarize the results in Table 6. While all methods achieve comparably testing accuracy on IID data, we observed that they suffer considerably in the non-IID setting. As this trend can be observed in both cases, we can conclude that the performance loss originating from the non-IID data is not unique to some functions. Compared with using different base models (VGG11 on CIFAR100 and ResNet50 on EMNIST datasets) in Figure 1, Section 5.2, the testing accuracy of new experiments has deteriorated, indicating that a proper choice of base models is also essential to FL's performance.

Table 6: Summary of average (standard deviation) *testing accuracy* after 50 runs of ResNet50 on CIFAR100 and VGG11 on EMNIST datasets. Both IID and non-IID(m) cases are reported. All methods suffer from degraded performances in the non-IID situation, but EFL is affected by far the least.

Methods	ResNet50 on CIFAR100				VGG11 on EMNIST			
	IID	Non-IID(5)	Non-IID(2)	Non-IID(1)	IID	Non-IID(5)	Non-IID(2)	Non-IID(1)
FedProx	82.48(.15)	80.14(.26)	34.25(.72)	21.87(.90)	83.12(.19)	70.18(.39)	56.94(.73)	35.26(.97)
SCAFFOLD	81.04(.18)	78.72(.29)	68.21(.47)	48.04(.68)	84.70(.17)	73.89(.31)	68.93(.56)	59.15(.72)
APFL	82.35(.21)	79.19(.32)	70.81(.51)	60.19(.57)	84.33(.26)	78.54(.49)	65.82(.67)	61.04(.83)
EFL	**83.59(.09)**	**81.23(.17)**	**77.36(.34)**	**75.35(.40)**	**87.62(.11)**	**81.94(.24)**	**76.19(.35)**	**75.27(.41)**