Equitable Client Selection in Federated Learning via Truncated Submodular Maximization

Andrés C. Castillo J., Ege C. Kaya, Lintao Ye, and Abolfazl Hashemi

***Abstract*— In a typical Federated Learning paradigm, a random subset of clients is selected at every round for training. This randomly chosen subset often does not perform well when evaluated in terms of fairness as the final model's performance often varies greatly between clients. This lack of a balanced and fair performance can be detrimental in sensitive applications, such as disease diagnosis in healthcare settings. Such issues may be exacerbated by emerging performance-centric client sampling procedures. This paper proposes a new equitable client selection method, SUBTRUNC, that addresses the shortcomings of random selection via a modification of the well-known facility location problem through submodular function maximization. This new approach entailing submodular functions incorporates using the information of loss values of each client to ensure a more balanced and thus more fair performance of the final model. Additionally, strong theoretical guarantees on the convergence of the resulting FL algorithm are established under mild assumptions. The algorithm's performance is evaluated on heterogeneous scenarios with a clear improvement in fairness when being observed under the scope of a client dissimilarity metric.**

I. INTRODUCTION

The Federated Learning (FL) paradigm involves the collaborative training of a centralized machine learning model using edge devices, commonly referred to as clients. This setting allows for models to be trained using localized data from these devices without the need to transmit the data to a centralized location. Updates to the model are accumulated from these clients via periodic communication rounds and aggregated at the central location resulting in an improved machine learning model.

Federated Learning is of particular use within the healthcare sector as privacy regulations typically prevent information from being shared. By treating healthcare sites as clients, and sharing only model updates with a centralized location, machine learning models that enhance patient care can be trained and deployed. Traditionally, randomly selecting a subset of clients has been the de facto approach for this paradigm [1]. However, previous work has found that oftentimes, this random selection approach does not perform well in terms of convergence properties and fairness, especially in heterogeneous settings where the data being held by each client may not necessarily come from the same distribution. This is especially evident in applications characterized by a high degree of data heterogeneity, where the need for a balanced and fair machine learning model is highly prioritized, such as disease diagnostics based on X-ray images, where images can come from X-ray machines manufactured by different suppliers. Because of this, client selection remains an open challenge within the field [2], [3]. Another motivating scenario relevant to control systems is the case of networked sensing [4], [5]. In networks comprising units, there is a common goal to devise an inference method that reduces the total estimation error. Nevertheless, in numerous scenarios, each unit must produce a dependable estimate to avoid impeding decision-making among other units in the network [6]. For instance, in autonomous vehicle settings, a unit with substantial estimation error might necessitate slowing down, thus influencing the behavior of other units [7]. Hence, there is an urgent need to minimize the collective mean-square estimation error across the entire network while ensuring consistent performance among individual units. Such considerations have further received attention in shared communication systems [8].

To address this, differing from previous client selection strategies, the idea of incorporating submodular set functions as a viable strategy for solving the client selection problem has been proposed [9]–[12]. A typical problem studied in submodular optimization literature is the maximization of a submodular function under a cardinality constraint [13]–[15]. In this problem, the task is to maximize the utility of the selection made from a ground set N, while making sure that the number of elements in the selection stays under a given integer cardinality constraint value κ. This can be formalized by the following problem [15]:

$$\max_{S \subseteq N} f(S) \qquad \text{s.t.} \ \ |S| \leq \kappa. \tag{1}$$

Where f is a submodular function, N is the ground set, and κ is a positive integer.

A. Contributions

This paper focuses on exploring a more balanced and thus fair selection criteria of clients by exploiting a modified facility location submodular function approach. To this end, we present SUBTRUNC, a submodular maximization-based client selection algorithm. By modifying the gradient similarity approach employed in DIVFL [9] to incorporate a truncated submodular function term, we obtain a modified objective where this addition acts as a balance-promoting regularization term, utilizing the loss value of each client in addition to gradient similarity metrics. This allows for an aggregated model that is more balanced in its performance

Andrés C. Castillo J., Ege C. Kaya, and Abolfazl Hashemi are with the Elmore Family School of Electrical and Computer Engineering at Purdue University, West Lafayette, IN, USA. Email: {castil12, kayae, abolfazl}@purdue.edu. Lintao Ye is with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China. Email: yelintao93@hust.edu.cn.

across all clients. This proposed method not only ensures that the most representative clients are chosen, but that they can participate in a fair manner, which results in a model that performs similarly, in terms of training and test performance, across all clients.

Additionally, this new method enjoys strong theoretical guarantees on its convergence properties under mild assumptions such as non-convexity. In particular, under smoothness, and without assuming the well-known Bounded Client Dissimilarity (BCD) assumption where one assumes that for some $G^2 \in \mathbb{R}$,

$$\|\nabla f_i(w) - \nabla f(w)\|^2 \leq G^2 \, \forall \, w \in \mathbb{R}^d, i \in N. \tag{2}$$

Our work shows that by relaxing (2), which in practice is a hard-to-verify assumption [16], the assumptions of a uniformly bounded gradient, and strong convexity required by prior works, our proposed algorithm enjoys strong theoretical guarantees that hold on the expected performance in nonconvex optimization problems. That is, our method needs $K = \mathcal{O}(1/\epsilon^2)$ rounds of communication in order to achieve $\mathbb{E}\big[\|\nabla f\left(w_{k^*}\right)\|^2\big] \leq \epsilon$ while doing away with the Bounded Client Dissimilarity Assumption under a smooth nonconvex scenario.

B. Related Works

Recently, by being able to model client selection through a submodular maximization problem, [9] was able to obtain better performance when compared with traditional client selection strategies like random sampling. This work does so by attempting to select a subset of clients whose gradient most closely resembles that of the full client set, by modeling the problem as a submodular maximization problem, which can be solved with greedy methods. Although our work, similar to [9], utilizes the concept of submodular maximization for client selection, it differs from it in that we employ a facility location submodular function modified by a truncated submodular function, making use of the loss value of each client for the truncation. This modification acts as a fairness-aware term that promotes a balanced model performance across all clients regardless of the distribution of the data these clients may hold, resulting in models whose performance does not drastically differ from client to client throughout the training process. Additionally, our theoretical analysis of the convergence of our method is significantly different from DIVFL and utilizes significantly milder assumptions.

Both [10], [11] have also explored the use of submodular maximization and their greedy solution as a means to solve the client selection problem. In particular, these works seek to create an optimal client schedule under computational and time constraints. This approach differs from ours in that instead of approximating the full client gradient by a subset and greedily selecting them, the problem is modeled as a Submodular-Cost Submodular-Knapsack problem where the selection is constrained by computational and timing metrics, whereas we look at our constraint through a truncated approach, which can be likened to the notion of the presence of a budget. Additionally, neither [10], [11] establish any convergence rate for the resultant FL method.

In tackling data heterogeneity and client selection schemes within Federated Learning, [12] also approaches the client selection problem as a submodular maximization problem that can be greedily solved under a knapsack constraint. This work seeks to maximize statistical performance under system performance constraints, like upload and communication time. This differs from our work in that we leverage the heterogeneity of the data in each client to create more diverse solutions by exploiting the loss at each client's dataset while approximating the solution set via the client's gradient. Additionally, we do not further constrain the problem under system heterogeneity metrics as [12].

II. Preliminaries and Background

This paper considers the standard FL setting comprised of a central server that acts as a model aggregator, n clients that can participate in training, and a model parameterized by $w \in \mathbb{R}^d$. Each client i in this setting has data coming from a distribution D_i and an objective function $f_i(w)$ which is the expected loss of the client concerning some loss function l over drawn data from D_i. The main objective is for the central server to optimize the average loss $f(w)$ over the $|N|$ clients:

$$f(w) := \frac{1}{|N|} \sum_{i \in N} f_i(w), \tag{3}$$

$$f_i(w) := \mathbb{E}_{x \sim D_i}[l(x, w)]. \tag{4}$$

When the data distributions across all clients are equal, the setting is considered independent and identically distributed (iid). If they differ across clients, then it is considered heterogeneous (non-iid).

At any given round of a typical FL setting, a random subset of clients is chosen to perform training. By carrying out a series of local gradient descent updates in the client's data and communicating these to the central server, the final model is constructed. This method of averaging out the client's updates is known as the FEDAVG algorithm [1].

However, client selection still remains an open challenge within FL [2], [3]. Recently, utilizing submodular functions for client selection by modeling the problem as a facility location problem was introduced [9]. This strategy aims to find a representative subset of clients whose aggregated update models what the overall aggregated update would look like if all clients participated in the training.

Our proposed method builds on the idea of leveraging submodular functions to create more representative client sets. It does so by finding a representative subset of clients while ensuring fair client usage by allowing those clients who may not be the most representative in a given round, to have more opportunity to participate in training. This fair client usage is introduced through a new regularization term that leverages the client's most recent loss value to design a judicious truncated function and adds that to the original modeling of the client selection problem via submodular functions.

Let us now introduce the concepts of marginal gain and submodularity [14].

Definition 1 (Marginal gain). *Given a set function $f: 2^N \rightarrow \mathbb{R}$ and $A, B \subseteq N$, we denote $f(B \cup A) - f(B)$, the marginal gain in f due to adding A to B, by $\Delta(A|B)$. When the set A is a singleton, i.e., $A = \{a\}$, we drop the curly brackets to adopt the short-hand notation $\Delta(a|B)$.*

We usually think of f as assigning a *utility score* to each subset $A \subseteq N$.

With Definition 1, we can now introduce the concept of submodularity in set functions:

Definition 2 (Submodularity). *A set function $f : 2^N \rightarrow \mathbb{R}$ is submodular if for every $A \subseteq B \subseteq N$ and $e \in N \setminus B$, it holds that*

$$\Delta(e|A) \geq \Delta(e|B). \tag{5}$$

This definition of submodularity gives a clear intuition about the nature of submodular functions, showcasing the diminishing marginal gains property, which can be exploited in the context of client selection [14].

Additionally, let us introduce monotone submodular functions.

Definition 3 (Monotonicity). *Let $f(S)$ be a submodular set function. If $f(S)$ satisfies the following: $\forall\, A \subseteq B \subseteq N, f(A) \leq f(B)$. It is said that f is a monotone submodular set function.*

It is well known that if $f(S)$ is a submodular function, $g(f(S))$ is also submodular for any concave function g [14]. This result leads to the following proposition which establishes the monotonicity and submodularity of *truncated* functions [17].

Proposition 1 (Truncation). *Let $g(S)$ be a monotone submodular set function composed of a non-decreasing concave function and let $b \in \mathbb{R}^+$. Then, g remains monotone submodular under truncation, i.e.,*

$$f(S) := \min\{g(S), b\}, \tag{6}$$

is a monotone submodular function.

We further have the following proposition [14].

Proposition 2 (Linear combination). *Nonnegative linear combinations of submodular functions preserve submodularity. More formally, let $g_1, g_2, ..., g_n : 2^N \rightarrow \mathbb{R}$ be submodular set functions. Let $\alpha_1, \alpha_2, ..., \alpha_n \geq 0$. Then, the positive linear combination*

$$f(S) := \sum_{i=1}^{n} \alpha_i g_i(S), \tag{7}$$

is also a submodular function.

Submodular functions and their maximization are very useful in practice, with a wide range of applications [18]–[20]. Due to their versatility and natural occurrence in practical and well-known settings their optimization has garnered interest in fields like optimization, control systems, and machine learning [21]–[23].

Modeling client selection as a modified facility location problem with a truncated submodular function regularization serves as the basis for our proposed method for a more balanced client selection.

III. SubTrunc: A Fairness-aware Client Selection Approach

To motivate our formulation of a modified facility location objective via a truncated submodular function, we first follow the outline of DivFL [9], which follows the logic found in [24]. Suppose there is a mapping $\sigma : N \rightarrow S$, in which N is the groundset of elements and S is the constructed set and where the gradient information $\nabla f_i(w)$ from client i is approximated by the gradient information from a selected client $\sigma(j) \in S$. For $j \in S$, $C_j := \{i \in N | \sigma(i) = j\}$ is the set approximated by client j and $\gamma_j := |C_j|$ results in the following formulation:

$$\begin{aligned} &\min_{S \subseteq N} \| \sum_{i \in N} \nabla f_i(w) - \sum_{j \in S} \gamma_j \nabla f_j(w) \| \leq \\ &\sum_{i \in N} \min_{j \in S} \| \nabla f_i(w) - \nabla f_j(w) \| := \bar{G}(S). \end{aligned} \tag{8}$$

That is, $\sum_{j \in S} \gamma_j \nabla f_j(w)$ can be viewed as the approximation of the global gradient $\sum_{i \in N} \nabla f_i(w)$. Therefore, the left-hand side of (8) can be interpreted as the approximation error, and the right-hand side of (8) provides an upper bound on the approximation error. Thus, to minimize the approximation error, DivFL aims to select a set of clients S that minimizes $\bar{G}(S)$ subject to a cardinality constraint on S. Upon defining $G(S) := \bar{G}(\emptyset) - \bar{G}(S)$ this task can be written as

$$\max_{S} G(S) \quad \text{s.t.} \quad |S| \leq \kappa, \tag{9}$$

where κ is a target bound on the number of participating clients in each communication round. Minimizing $\bar{G}(S)$ or equivalently maximizing $G(S)$ is the equivalent of maximizing the well-known facility location monotone submodular function [14]. However, this problem under a cardinality constraint, is NP-hard in general which requires efficient approximation algorithms to provide a near-optimal solution. The greedy algorithm and its randomized variants are among the canonical methods for such optimization problems [25]–[27].

Finding the most representative set at any given round may not always provide a model that performs in a balanced and similar fashion across all clients thereby leading to potential unfair behaviours in FL-based model training. As a result of this observation, we propose a fairness regularization term utilizing the truncation of a judicious submodular function:

$$H(S) := \lambda \min(b, F(S)), \tag{10}$$

with

$$F(S) := \sum_{i \in S} \phi(f_i(w)), \tag{11}$$

where $f_i(w)$ is the expected loss of client i with respect to some loss function l as defined in (4), $\lambda \geq 0$ and $b \in \mathbb{R}^+$ are input parameters aiming to explore the inherent trade-off between performance and fairness; $\phi(\cdot)$ can be any monotone nondecreasing function with a bounded Lipschitz constant L. When a function with $L < 1$ is used, $F(S)$ effectively becomes an attenuation term, suppressing the difference between the client losses. Whereas, when $L > 1$, $F(S)$ enhances this difference. For the purpose of this work, we use $\phi(x) = \ln(1 + x)$, which has a Lipschitz constant $L < 1$. However, the choice of ϕ could be a potential avenue of further research as finding a ϕ that judiciously tunes this attenuation or enhancement effect could be of interest. Here, we assume without loss of generality that the clients' loss functions are nonnegative. The combination of this fairness-aware term with the original objective results in the following optimization for client selection:

$$\max_{S \subseteq N} W(S) := G(S) + H(S) \quad \text{s.t.} \quad |S| \leq \kappa. \tag{12}$$

A desirable property of the proposed formulation is the preservation of monotonicity and submodularity, which is indeed the case, as demonstrated next.

Proposition 3. *The set function $W(S)$ in* (12) *is monotone and submodular.*

Proof. Note that $F(S)$ is a modular function and hence monotone and submodular [14]. By Proposition 1, $H(S)$ is therefore a monotone submodular function. Finally, from Proposition 2, it can be seen that any linear combination of submodular functions with the same ground set remains submodular. Since both $G(S)$ and $H(S)$ in (10) are monotone submodular, their nonnegative linear combination in(12) is also monotone submodular. ■

The fairness-aware term as defined in (10) is the combination of tuneable parameters in the form of both λ and b, where λ represents a weighting on the fairness term and b acts as a truncation parameter. From this formulation, we can see that if we set $\lambda = 0$, then we obtain the original DIVFL formulation [9]. On the other hand, increasing λ puts more weight on the proposed regularization term, thereby prioritizing fairness over performance. Additionally, varying b, as stated previously, will allow *training participation* to those clients who have not participated in previous rounds, or whose participation has been minimal when compared to other clients. In particular, consider a scenario where b is very large. Then, the minimum in the definition of $H(S)$ will typically equate to $F(S) := \sum_{i \in S} \phi(f_i(w))$ which is maximized by selecting the worst-performing clients, according to their local loss functions, who are suffering the most as a result of learning a collaborative model. On the other hand, if b is very small, then the minimum in the definition of $H(S)$ will typically equate to b which is independent of S, and this effectively makes the client selection independent of local performance and can be thought of as a global performance-centric formulation.

Algorithm 1 SUBTRUNC

Input: Truncation and regularization parameters $b, \lambda \in \mathbb{R}$, communication rounds K, local steps E, participating clients κ, initial weight vector w_0, learning rate η
Output: w_k, weights for trained model

Server initializes w_0
for $k = 1, \dots, K$ **do**
  Subset S_i of size κ is selected by the server via the
  stochastic variant of the naïve greedy algorithm,
  following the formulation of (12).
  for client $i \in S_i$ **do**
    $w_{k,0}^{(i)} \leftarrow w_k$
    for $\tau = 1, \dots, E$ **do**
      Select random batch from client i:
      $\mathcal{B}_{k,\tau}^{(i)}$ compute stochastic gradient $\tilde{f}_i$ at $w_{k,\tau}^{(i)}$
      over $\mathcal{B}_{k,\tau}^{(i)}$
      $w_{k,\tau+1}^{(i)} \leftarrow w_{k,\tau}^{(i)} - \eta \nabla \tilde{f}_i(w_{k,\tau}^{(i)}; \mathcal{B}_{k,\tau}^{(i)})$
    end for
    $w_k^{(i)} \leftarrow w_k - w_{k,E}^{(i)}$
  end for
  $w_{k+1} \leftarrow w_k - \frac{1}{\kappa} \sum_{i \in S_i} w_k^{(i)}$
end for
return w_K ▷ Final model weights

Finding a solution for the best client that can maximize the utility score of (12) can be done with the naïve greedy algorithm, due to the submodular nature of (12) as given by Proposition 3 [25].

A naïve greedy approach starts at round i with an empty set, $S_i \leftarrow \emptyset$, and adds the element $e \in N \setminus S_i$ which provides the highest marginal gain, $\Delta(e|S_i)$.

$$S_{i+1} \leftarrow S_i \cup \{\arg\max_{e \in N \setminus S_i} (\Delta(e|S_i))\}. \tag{13}$$

Nonetheless, if the cardinality of the ground set $|N|$ is too large, searching through this space for the desired elements may prove to be too expensive. Because of this, stochastic variants of the naïve greedy algorithm can be employed [26]–[29] to effectively reduce this search cost while still maintaining high-confidence in the solution provided. This is done by randomly sampling a smaller subset R of size r and searching through this reduced space. That is:

$$S_{i+1} \leftarrow S_i \cup \{\arg\max_{e \in R \setminus S_i} (\Delta(e|S_i))\}. \tag{14}$$

Applying the selection strategy utilizing 12, and using the FEDAVG algorithm as the core method for aggregating client model updates results in our proposed algorithm SUBTRUNC which is summarized as Algorithm 1.

IV. THEORETICAL CONVERGENCE ANALYSIS

In this section, we analyze the convergence properties of the proposed algorithm under standard assumptions in nonconvex FL. Our analysis utilizes more relaxed assumptions compared to [9], complementing our proposed fairness-promoting formulation.

TABLE I: Comparison of Client Selection Methods on non-iid MNIST and CIFAR10.

	MNIST			CIFAR10		
Method	Training Loss	Testing Accuracy [%]	Client Dissimilarity [%]	Training Loss	Testing Accuracy [%]	Client Dissimilarity [%]
DIVFL	$\mathbf{0.16 \pm 0.02}$	82.16 ± 0.91	8.89 ± 0.59	$\mathbf{0.88 \pm 0.03}$	35.40 ± 1.28	12.62 ± 0.95
SUBTRUNC	0.21 ± 0.01	$\mathbf{83.72 \pm 0.81}$	$\mathbf{7.96 \pm 0.62}$	0.89 ± 0.03	$\mathbf{35.53 \pm 1.33}$	$\mathbf{12.46 \pm 0.94}$
RANDOM	0.18 ± 0.02	82.99 ± 0.99	9.16 ± 0.18	0.92 ± 0.01	32.55 ± 0.94	14.42 ± 0.18

A. Assumptions and Related Concepts

The definitions and assumptions used to analyze the performance of the algorithm are listed below. These are standard and ubiquitous in the analysis of training algorithms for nonconvex machine learning and FL problems [16], [30]–[33].

Definition 4 (L-smoothness)**.** *A function $f : \Omega \to \mathbb{R}$ is considered to be L-smooth if $\forall\, w, w' \in \Omega$, $\|\nabla f(w) - \nabla f(w')| \leq L\|w - w'\|$. Additionally, if f can be differentiated twice, then $\forall\, w, w' \in \Omega$, $f(w') \leq f(w) + \langle \nabla f(w), w' - w\rangle + \frac{L}{2}\|w' - w\|^2$.*

Assumption 1 (L-smoothness)**.** *For all x, we assume $l(x, w)$ to be L-smooth with respect to w. Then, each $f_i(w)$ defined in* (4) *where $i \in [n]$ is L-smooth, and so is $f(w)$ defined in* (3).

Assumption 2 (Nonnegativity)**.** *Each $f_i(w)$ is non-negative, therefore, $f_i^* := \min_w f_i(w) \geq 0$.*

The nonnegativity assumption holds without loss of generality as any function bounded from below can be shifted to satisfy this assumption. Furthermore, almost all loss functions of interest in FL are nonnegative.

Assumption 3 (Bounded bias)**.** *Let τ be the local update steps, with $\tau \in \{0, \ldots, E - 1\}$, K be the communication rounds with $k \in \{0, ..., K - 1\}$, $\hat{u}_{k,\tau}^{(i)}$ be the stochastic gradient of client i at communication round k and local update step τ and let $b_{k,\tau}$ be the approximation error for the true gradient at communication round k and local update step τ. We assume that the gradient of the subset constructed by our objective function* (12) *is the gradient of the global loss function plus an approximation error, $b_{k,\tau}$ which we refer to as bias. That is:*

$$\frac{1}{\kappa} \sum_{i \in S_k} \hat{u}_{k,\tau}^{(i)} = \hat{u}_{k,\tau} + b_{k,\tau}. \tag{15}$$

We assume that this bias, $b_{k,\tau}$ is bounded by: $b_{k,\tau} \leq \gamma$.

This assumption is also used in prior work [9], [24]. Similar to [9], we assume that the gradient of the subset constructed by our objective function approximates the full gradient of the overall set, with the addition of an approximation error which we model as bias.

B. Main Theoretical Results

Theorem 1 states the convergence properties of the proposed SUBTRUNC method summarized as Algorithm 1.

Theorem 1. *Let Assumptions 1, 2, and 3 hold. Set $\eta_k = \frac{1}{LE}\sqrt{\frac{1}{K}}\,\forall\, k \in \{0, ..., K - 1\}$. Let $\mathbb{P}$ be a distribution such that $\mathbb{P}(k) = \frac{(1+\zeta)^{(K-1-k)}}{\sum_{k'=0}^{K-1}(1+\zeta)^{(K-1-k')}}$, where $\zeta := \eta^2 L^2 E^2 \left(\frac{9\eta LE}{4}\right)$. Let $k^* \sim \mathbb{P}$. Then, for $K \geq 9$:*

$$\begin{gathered}\mathbb{E}\left[\|\nabla f\left(w_{k^*}\right)\|^2\right] \leq \frac{12Lf\left(w_0\right)}{\sqrt{K}} + \\ \left(\frac{2}{EnK} + \frac{3}{K} + \frac{4}{En\sqrt{K}}\right)\sigma^2 + \left(4 + \frac{4}{\sqrt{K}}\right)\gamma.\end{gathered} \tag{16}$$

That is, there exists a learning rate and a nonuniform distribution on the iterates such that if the output is generated according to that distribution, the expected performance satisfies:

$$\mathbb{E}\left[\|\nabla f\left(w_{k^*}\right)\|^2\right] = \mathcal{O}\left(\frac{1}{\sqrt{K}} + \frac{\sigma^2}{En\sqrt{K}} + \gamma\right). \tag{17}$$

A complete proof of Theorem 1 can be found in: `https://abolfazlh.github.io/files/subtrunc.pdf`. Theorem 1 establishes a bound on the so-called approximate first-order stationary of the global model parameters. In particular, we can see that the convergence error consists of three terms: the first term quantifies the impact of the initialization. The second term captures the impact of the statistical noise in the local stochastic gradients utilized by the clients. Finally, the last term captures the impact of bias that arises from using the proposed client selection strategy (12), where the bias is defined in Assumption 3. Theorem 1, hence, indicates that as long as $K = \Omega(1/\epsilon^2)$, it holds that $\mathbb{E}\left[\|\nabla f\left(w_{k^*}\right)\|^2\right] \leq \mathcal{O}(\epsilon + \gamma)$. Thus, if the bias parameter satisfies $\gamma = \mathcal{O}(\epsilon)$, Algorithm 1 (SUBTRUNC) can identify an ϵ-accurate first-order stationary solution.

Unlike prior work, we analyze the proposed method without making the Bounded Client Dissimilarity assumption, by leveraging the smoothness of the clients' loss functions and a nonuniform sampling technique to output a solution. Additionally, compared to [9] which assumes the restrictive assumption of strong convexity, our analysis applies to general nonconvex problems that arise frequently in practical and large-scale application of FL in the big data setting.

Finally, note that adopting a distribution over the iterates to output a global model is a standard approach to state the theoretical convergence results for FL and optimization algorithms [16], [31], [34]. In practice, however, the latest global model is used for inference.

V. EXPERIMENTAL RESULT

The performance of SUBTRUNC is evaluated on the MNIST dataset of handwritten digits as well as on the

TABLE II: Effects of a varying λ under a non-iid setting within MNIST and CIFAR10.

	MNIST		CIFAR10	
λ-Value	Testing Accuracy [%]	Client Dissimilarity [%]	Testing Accuracy [%]	Client Dissimilarity [%]
0.10	83.13 ± 1.04	8.38 ± 0.56	35.47 ± 1.33	12.58 ± 0.92
0.25	83.60 ± 0.78	8.18 ± 0.55	35.47 ± 1.29	12.47 ± 0.96
0.50	83.76 ± 0.42	8.14 ± 0.34	35.53 ± 1.33	12.46 ± 0.94
0.75	83.94 ± 0.33	8.07 ± 0.41	35.55 ± 1.32	12.46 ± 1.08
0.95	83.72 ± 0.81	7.96 ± 0.62	35.50 ± 1.33	12.45 ± 1.05

CIFAR10 dataset, both under a non-iid data distribution. We benchmark the performance of SubTrunc to DivFL as well as the random sampling of clients without replacement which is a simple but standard benchmark for client sampling. The experiments are simulated on a pool of $|N| = 100$ clients, but in order to avoid scanning through the whole set, we employ a stochastic greedy search with a subset R of $r = 10$ clients with $\kappa = 10$, on the LeNet architecture.

Both datasets are partitioned in such a way that each client had 3 equally partitioned distinct classes. Unless otherwise stated, the truncation factor is set at a value of $b = 1.10$, which seemed to be the best value for these particular datasets.

Our evaluation utilizes three metrics: training loss, test accuracy, and client dissimilarity. The first aims to characterize the convergence property of the proposed method while the second aims to characterize its generalization power. The last metric aims to characterize how balanced or fair the global model is. Client dissimilarity is calculated by taking the difference of the final model's performance on the client's test dataset, across all clients. Measuring model dissimilarity this way allows us to better express the divergence in model performance across different clients.

A. Results on Non-IID MNIST & CIFAR10

Under both non-iid scenarios, SubTrunc outperforms the baselines by achieving a lower overall client dissimilarity score, indicating that the final model's performance is more consistent across clients. Additionally SubTrunc converges at a similar rate to the baseline methods as evidenced by the training loss while maintaining a comparable performance under testing accuracy. These results can be seen in Table I, where the bolded entries represent the best performance.

The dissimilarity evolution of our method demonstrates that our algorithm SubTrunc is effective in ensuring a more balanced and thus fair performance across all clients even under the presence of high data heterogeneity throughout the course of training.

B. Tuning Lambda

Seeking to better understand the effect of λ on our fairness-aware term, we simulate results under the same non-iid conditions and setting as described at the beginning of Section V, while varying λ.

It can be seen from Table II that as the values for λ get closer to zero, i.e. as the method approaches DivFL, the client dissimilarity score degrades, whereas as λ increases, the client dissimilarity improves, while still maintaining a comparable testing accuracy across the gamut of λ values. This highlights the important trade-off between performance-centric models versus balanced models.

VI. Conclusion

In this paper, we proposed the inclusion of a fairness-aware term to the submodular maximization approach of solving the client selection problem within the FL setting resulting in our algorithm, SubTrunc. Our proposed algorithm is able to obtain models that perform in a more balanced fashion across clients and highlights the trade-off between accuracy and model uniformity by judiciously tuning both b, the truncation parameter, and λ, the weighting of the fairness-aware term. Our method's main benefit of reduced client dissimilarity lends itself to practical applications in sectors such as healthcare and control systems where a balanced and accurate performance across highly non-iid clients is required to perform in a consistent manner regardless of client data. Additionally, we showed theoretically that our method needs $K = \mathcal{O}(1/\epsilon^2)$ rounds of communication to achieve $\mathbb{E}\left[\|f\left(w_{k^*}\right)\|^2\right] \leq \epsilon$ and it does so under significantly milder assumptions than prior work. Through our experimental results, we showed that our algorithm results in more balanced models when compared to both the random selection strategy and DivFL under the scope of a client dissimilarity metric. The addition of this fairness-aware term makes our algorithm an easy-to-implement solution that can enjoy a more balanced and thus fair model without compromising on performance in federated settings.

Acknowledgments

The authors would like to thank Dr. Vijay Gupta (Elmore Family School of Electrical and Computer Engineering, Purdue University) for his guidance and contributions to this work.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. International Conference on Artificial Intelligence and Statistics*, vol. 54, pp. 1273–1282, PMLR, 2017.

[2] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, *et al.*, "Advances and open problems in federated learning," *Foundations and trends® in machine learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[3] J. Wang, Z. Charles, Z. Xu, G. Joshi, H. B. McMahan, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data, *et al.*, "A field guide to federated optimization," *arXiv preprint arXiv:2107.06917*, 2021.

[4] S. Grime and H. F. Durrant-Whyte, "Data fusion in decentralized sensor networks," *Control engineering practice*, vol. 2, no. 5, pp. 849–863, 1994.

[5] R. Olfati-Saber, "Distributed kalman filtering for sensor networks," in *Proc. IEEE Conference on Decision and Control*, pp. 5492–5498, 2007.

[6] A. Hashemi, O. F. Kilic, and H. Vikalo, "Near-optimal distributed estimation for a network of sensing units operating under communication constraints," in *Proc. IEEE Conference on Decision and Control*, pp. 2890–2895, 2018.

[7] N. Lu, N. Cheng, N. Zhang, X. Shen, and J. W. Mark, "Connected vehicles: Solutions and challenges," *IEEE Internet of Things Journal*, vol. 1, no. 4, pp. 289–299, 2014.

[8] R. Jurdi, J. G. Andrews, and R. W. Heath, "Scheduling observers over a shared channel with hard delivery deadlines," *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 133–148, 2020.

[9] R. Balakrishnan, T. Li, T. Zhou, N. Himayat, V. Smith, and J. Bilmes, "Diverse client selection for federated learning via submodular maximization," in *Proc. International Conference on Learning Representations*, 2022.

[10] L. Ye and V. Gupta, "Client Scheduling for Federated Learning over Wireless Networks: A Submodular Optimization Approach," in *Proc. IEEE Conference on Decision and Control*, pp. 63–68, 12 2021.

[11] T. Omori and K. Kashima, "Combinatorial optimization approach to client scheduling for federated learning," *IEEE Control Systems Letters*, 2023.

[12] J. Zhang, J. Wang, Y. Li, F. Xin, F. Dong, J. Luo, and Z. Wu, "Addressing heterogeneity in federated learning with client selection via submodular optimization," *ACM Transactions on Sensor Networks*, vol. 20, no. 2, pp. 1–32, 2024.

[13] A. Asadpour, H. Nazerzadeh, and A. Saberi, "Stochastic submodular maximization," in *Internet and Network Economics* (C. Papadimitriou and S. Zhang, eds.), (Berlin, Heidelberg), pp. 477–489, Springer Berlin Heidelberg, 2008.

[14] A. Krause and D. Golovin, "Submodular Function Maximization," in *Tractability*, pp. 71–104, Cambridge University Press, 2 2014.

[15] Niv Buchbinder, Moran Feldman, Joseph (Seffi) Naor, and Roy Schwartz, "Submodular maximization with cardinality constraints," in *Proc. Annual ACM-SIAM symposium on Discrete algorithms*, pp. 1433–1452, 1 2014.

[16] R. Das, A. Acharya, A. Hashemi, S. Sanghavi, I. S. Dhillon, and U. Topcu, "Faster non-convex federated learning via global and local momentum," in *Proc. Uncertainty in Artificial Intelligence*, pp. 496–506, PMLR, 2022.

[17] M.-F. Balcan and N. J. A. Harvey, "Submodular Functions: Learnability, Structure, and Optimization," *SIAM Journal on Computing*, vol. 47, pp. 703–754, 1 2018.

[18] Andreas Krause, H. Brendan McMahan, Carlos Guestrin, and Anupam Gupta, "Robust Submodular Observation Selection," *Journal of Machine Learning Research*, vol. 9, pp. 2761–2801, 2008.

[19] M. Shamaiah, S. Banerjee, and H. Vikalo, "Greedy sensor selection: Leveraging submodularity," in *Proc. IEEE Conference on Decision and Control*, pp. 2572–2577, 12 2010.

[20] Andreas Krause and Carlos Guestrin, "Near-optimal observation selection using submodular functions," *Proc. International Conference on Artificial Intelligence*, vol. 2, pp. 1650–1654, 7 2007.

[21] L. Ye and S. Sundaram, "Distributed Maximization of Submodular and Approximately Submodular Functions," in *Proc. IEEE Conference on Decision and Control*, pp. 2979–2984, 12 2020.

[22] J. P. Hespanha and D. Garagic, "Optimal Sensor Selection for Binary Detection based on Stochastic Submodular Optimization," in *Proc. IEEE Conference on Decision and Control*, pp. 3464–3470, 12 2020.

[23] J. Edmonds, *Submodular Functions, Matroids, and Certain Polyhedra*, pp. 11–26. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003.

[24] B. Mirzasoleiman, J. Bilmes, and J. Leskovec, "Coresets for data-efficient training of machine learning models," in *Proc. International Conference on Machine Learning*, vol. 119, pp. 6950–6960, PMLR, 2020.

[25] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Mathematical Programming*, vol. 14, pp. 265–294, 12 1978.

[26] B. Mirzasoleiman, A. Badanidiyuru, A. Karbasi, J. Vondrak, and A. Krause, "Lazier Than Lazy Greedy," in *Proc. the AAAI Conference on Artificial Intelligence*, vol. 29, 2015.

[27] A. Hashemi, M. Ghasemi, H. Vikalo, and U. Topcu, "Randomized Greedy Sensor Selection: Leveraging Weak Submodularity," *IEEE Transactions on Automatic Control*, vol. 66, pp. 199–212, 1 2021.

[28] A. Hashemi, M. Ghasemi, H. Vikalo, and U. Topcu, "A Randomized Greedy Algorithm for Near-Optimal Sensor Scheduling in Large-Scale Sensor Networks," in *Proc. American Control Conference*, pp. 1027–1032, IEEE, 2018.

[29] A. Hashemi, A. Acharya, R. Das, H. Vikalo, S. Sanghavi, and I. S. Dhillon, "On the Benefits of Multiple Gossip Steps in Communication-Constrained Decentralized Federated Learning," *IEEE Transactions on Parallel and Distributed Systems*, pp. 1–1, 2022.

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

[31] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. International Conference on Machine Learning*, vol. 119, pp. 5132–5143, PMLR, 2020.

[32] S. U. Stich, "Local SGD converges fast and communicates little," in *Proc. International Conference on Learning Representations*, 2019.

[33] A. Acharya, A. Hashemi, P. Jain, S. Sanghavi, I. S. Dhillon, and U. Topcu, "Robust training in high dimensions via block coordinate geometric median descent," in *Proc. International Conference on Artificial Intelligence and Statistics*, vol. 151, pp. 11145–11168, PMLR, 2022.

[34] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in *Proc. International Conference on Artificial Intelligence and Statistics*, vol. 108, pp. 2021–2031, PMLR, 2020.

APPENDIX

A. Convergence Analysis for SubTrunc

We will provide convergence results for SubTrunc, without the Bounded Client Dissimilarity Assumption. Following a similar outline to [16], we use Lemma 1 and exploit the L-smoothness of f to obtain a bound on the per-round progress of the algorithm. Additionally both Lemmas 2, 3 are used to help bound terms deriving from the analysis of Lemma 1, mainly dealing with bounding the gradient error at any given round as well as providing a bound on the expected value of the gradient for each client at any given communication round k and local step τ. Different from [16], we carefully account for the bias terms that arise from utilizing the submodular client selection method by leveraging Assumption 3.

Now, we present the proof of Theorem 1 which states the convergence properties of SubTrunc.

Proof. Using the results from Lemma 1, for $\eta_k LE \leq \frac{1}{3}$, the per-round progress can be bounded as follows:

$$\begin{gathered}\mathbb{E}\left[f\left(w_{k+1}\right)\right] \leq \mathbb{E}\left[f\left(w_k\right)\right] - \frac{\eta_k E}{4}\mathbb{E}\left[\|\nabla f\left(w_k\right)\|^2\right] + \\ \eta_k^2 LE^2\left(\frac{9\eta_k LE}{4}\right)\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\left[\|\nabla f_i\left(w_k\right)\|^2\right] + \\ \eta_k^2 LE\left(\frac{\eta_k LE}{2}\left(\frac{1}{n}+\frac{3E}{4}\right)+\frac{1}{n}\right)\sigma^2 + \eta_k\left(1+\eta_k LE\right)E\gamma.\end{gathered}$$

Now, by utilizing f_i's attributes of L-smoothness and non-negativity, we obtain the following:

$$\begin{aligned}&\sum_{i\in[n]}\|\nabla f_i\left(w_k\right)\|^2 \leq \sum_{i\in[n]} 2L\left(f_i\left(w_k\right)-f_i^*\right)\\ &\leq 2nLf\left(w_k\right) - 2L \leq \sum_{i\in[n]} f_i^* \leq 2nLf\left(w_k\right).\end{aligned}$$

Plugging the above to the result of Lemma 1 and setting a constant learning rate $\eta_k = \eta$, we get:

$$\begin{gathered}\mathbb{E}\left[f\left(w_{k+1}\right)\right] \leq \left(1+\eta^2 L^2 E^2\left(\frac{9\eta LE}{4}\right)\right)\mathbb{E}\left[f\left(w_k\right)\right] - \\ \frac{\eta E}{4}\mathbb{E}\left[\|\nabla f\left(w_k\right)\|^2\right] + \eta^2 LE\left(\frac{\eta LE}{2}\left(\frac{1}{n}+\frac{3E}{4}\right)+\frac{1}{n}\right)\sigma^2 + \\ \eta\left(1+\eta LE\right)E\gamma.\end{gathered} \tag{18}$$

Let us define the following $\zeta := \eta^2 L^2 E^2\left(\frac{9\eta LE}{4}\right)$, $\zeta_2 := \left(\frac{\eta LE}{2}\left(\frac{1}{n}+\frac{3E}{4}\right)+\frac{1}{n}\right)$ and finally $\zeta_3 := \left(1+\eta LE\right)$. Then substituting the above, and unfolding the recursion of (18), we obtain:

$$\begin{gathered}\mathbb{E}\left[f\left(w_K\right)\right] \leq \left(1+\zeta\right)^K f\left(w_0\right) - \\ \frac{\eta E}{4}\sum_{k=0}^{K-1}\left(1+\zeta\right)^{(K-1-k)}\mathbb{E}\left[\|\nabla f\left(w_k\right)\|^2\right] + \\ \eta^2 LE\zeta_2\sigma^2\sum_{k=0}^{K-1}\left(1+\zeta\right)^{(K-1-k)} + \eta\zeta_3 E\gamma\sum_{k=0}^{K-1}\left(1-\zeta\right)^{(K-1-k)}.\end{gathered} \tag{19}$$

Define $\mathbb{P}\left(k\right) := \frac{\left(1-\zeta\right)^{(K-1-k)}}{\sum_{k'=0}^{K-1}\left(1+\zeta\right)^{(K-1-k')}}$. Then, by re-arranging equation (19) and using $\mathbb{E}\left[f\left(w_K\right)\right] \geq 0$, we get the following:

$$\begin{gathered}\sum_{k=0}^{K-1} p_k\mathbb{E}\left[\|\nabla f\left(w_k\right)\|^2\right] \leq \frac{4\left(1+\zeta\right)^K f\left(w_0\right)}{\eta E\sum_{k'=0}^{K-1}\left(1+\zeta\right)^{k'}} + \\ 4\eta L\zeta_2\sigma^2 + 4\zeta_3\gamma\end{gathered} \tag{20}$$

$$= \frac{4\zeta f\left(w_0\right)}{\eta E\left(1-\left(1+\zeta\right)^{-K}\right)} + 4\eta L\zeta_2\sigma^2 + 4\zeta_3\gamma. \tag{21}$$

This last step follows from $\sum_{k'=0}^{K-1}\left(1+\zeta\right)^{k'} = \frac{(1+\zeta)^K-1}{\zeta}$. Now, let us show:

$$(1+\zeta)^{-K} < 1-\zeta K+\zeta^2\frac{K(K+1)}{2} < 1-\zeta K+\zeta^2K^2$$
$$\implies 1-(1+\zeta)^{-K} > \zeta K(1-\zeta K).$$

Using the above on equation (21), we have that for $\zeta K < 1$:

$$\sum_{k=0}^{K-1} p_k\mathbb{E}\left[\|\nabla f(w_k)\|^2\right] \le \frac{4f(w_0)}{\eta EK(1-\zeta K)}+ \\ 4\eta L\left(\frac{\eta LE}{2}\left(\frac{1}{n}+\frac{3E}{4}\right)+\frac{1}{n}\right)\sigma^2+4(1+\eta LE)\gamma. \tag{22}$$

Now, let us pick $\eta=\frac{1}{LE}\sqrt{\frac{1}{K}}$. We need to make sure that $\eta LE\le\frac{1}{3}$; this happens for $K\ge 9$. Plugging $\eta LE=\frac{1}{LE}\sqrt{\frac{1}{K}}$ in (22), and using $1-\zeta K\ge\frac{1}{3}$, we get:

$$\sum_{k=0}^{K-1} p_k\mathbb{E}\left[\|\nabla f(w_k)\|^2\right] \le \frac{12Lf(w_0)}{\sqrt{K}}+ \\ \left(\frac{2}{EnK}+\frac{3}{K}+\frac{4}{En\sqrt{K}}\right)\sigma^2+\left(4+\frac{4}{\sqrt{K}}\right)\gamma. \tag{23}$$

This finishes the proof. ■

Next, we provide the intermediate lemmas used in the proof of Theorem 1.
Using the following Lemma 1 we will be able to bound the per-round progress of the algorithm.

Lemma 1. *With $\eta_k LE\le\frac{1}{3}$, we have:*

$$\mathbb{E}[f(w_{k+1})]\le\mathbb{E}[f(w_k)]-\frac{\eta_k E}{4}\mathbb{E}\left[\|\nabla f(w_k)\|^2\right]+ \\ \eta_k^2LE^2\left(\frac{9\eta_k LE}{4}\right)\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\left[\|\nabla f_i(w_k)\|^2\right] \\ +\eta_k^2LE\left(\frac{\eta_k LE}{2}\left(\frac{1}{n}+\frac{3E}{4}\right)+\frac{1}{n}\right)\sigma^2+ \\ \eta_k(1+\eta_k LE)E\gamma. \tag{24}$$

Proof. Define:

$$\hat{u}_{k,\tau}^{(i)} := \nabla\tilde{f}_i\left(w_{k,\tau}^{(i)};\mathcal{B}_{k,\tau}^{(i)}\right),$$
$$\hat{u}_{k,\tau} := \frac{1}{n}\sum_{i\in[n]}\hat{u}_{k,\tau}^{(i)},$$
$$u_{k,\tau} := \frac{1}{n}\sum_{i\in[n]}\nabla f_i\left(w_{k,\tau}^{(i)}\right),$$
$$\frac{1}{\kappa}\sum_{i\in S_k}\hat{u}_{k,\tau}^{(i)}=\hat{u}_{k,\tau}+b_{k,\tau},$$
$$\bar{w}_{k,\tau} := \frac{1}{n}\sum_{i\in[n]}w_{k,\tau}^{(i)},\ \tilde{e}_{k,\tau}^{(i)}=\nabla f_i\left(w_{k,\tau}^{(i)}\right)-\nabla f_i(\bar{w}_{k,\tau}).$$

Then:

$$w_{k+1}=w_k-\eta_k\sum_{\tau=0}^{E-1}\left(\frac{1}{r}\sum_{i\in S_k}\hat{u}_{k,\tau}^{(i)}\right), \tag{25}$$

$$\bar{w}_{k,\tau}=w_k-\eta_k\sum_{t=0}^{\tau-1}\hat{u}_{k,t}, \tag{26}$$

$$\mathbb{E}_{\{\mathcal{B}_{k,\tau}^{(i)}\}_{i=1}^n}\left[\hat{u}_{k,\tau}\right]=u_{k,\tau}, \tag{27}$$

$$\mathbb{E}\left[\|\sum_{t=0}^{\tau-1}\hat{u}_{k,t}\|^2\right] \leq \tau\sum_{t=0}^{\tau-1}\mathbb{E}\left[\|u_{k,t}\|^2\right] + \frac{\tau\sigma^2}{n}, \tag{28}$$

$$\mathbb{E}\left[\|\sum_{t=0}^{\tau-1}\hat{u}_{k,t}^{(i)}\|^2\right] \leq \tau\sum_{t=0}^{\tau-1}\mathbb{E}\left[\|\nabla f_i\left(w_{k,t}^{(i)}\right)\|^2\right] + \tau\sigma^2, \tag{29}$$

$$\sum_{\tau=0}^{E-1}\mathbb{E}\left[\|b_{k,\tau}\|^2\right] \leq E\gamma. \tag{30}$$

Now, by using f's L-smoothness and (25), we obtain the following:

$$\begin{gathered}\mathbb{E}[f(w_{k+1})] \leq \mathbb{E}[f(w_k)] - \\ \mathbb{E}\left[\left\langle\nabla f(w_k), \eta_k\sum_{\tau=0}^{E-1}\left(\frac{1}{\kappa}\sum_{i\in S_k}\hat{u}_{k,\tau}^{(i)}\right)\right\rangle\right] + \\ \frac{L}{2}\mathbb{E}\left[\|\eta_k\sum_{\tau=0}^{E-1}\left(\frac{1}{\kappa}\sum_{i\in S_k}\hat{u}_{k,\tau}^{(i)}\right)\|^2\right],\end{gathered} \tag{31}$$

$$\begin{gathered}= \mathbb{E}[f(w_k)] - \mathbb{E}\left[\left\langle\nabla f(w_k), \sum_{\tau=0}^{E-1}\eta_k\left(\hat{u}_{k,\tau} + b_{k,\tau}\right)\right\rangle\right] + \\ \frac{\eta_k^2 L}{2}\mathbb{E}\left[\|\sum_{\tau=0}^{E-1}\hat{u}_{k,\tau} + b_{k,\tau}\|^2\right],\end{gathered} \tag{32}$$

$$\begin{gathered}= \mathbb{E}[f(w_k)] - \\ \mathbb{E}\left[\underbrace{\left\langle\nabla f(w_k), \sum_{\tau=0}^{E-1}\eta_k u_{k,\tau}\right\rangle}_{A} + \underbrace{\left\langle\nabla f(w_k), \sum_{\tau=0}^{E-1}\eta_k b_{k,\tau}\right\rangle}_{B}\right] + \\ \frac{\eta_k^2 L}{2}\underbrace{\mathbb{E}\left[\|\sum_{\tau=0}^{E-1}\hat{u}_{k,\tau} + b_{k,\tau}\|^2\right]}_{C}.\end{gathered} \tag{33}$$

We can now use that for 2 vectors x, y, the following holds: $\langle x, y\rangle = \frac{1}{2}\left(\|x\|^2 + \|y\|^2 - \|x-y\|^2\right)$. Leveraging this for term A, we obtain the following:

$$\begin{gathered}\left\langle\nabla f(w_k), \sum_{\tau=0}^{E-1}u_{k,\tau}\right\rangle = \sum_{\tau=0}^{E-1}\langle\nabla f(w_k), u_{k,\tau}\rangle = \\ \frac{1}{2}\sum_{\tau=0}^{E-1}\left(\|\nabla f(w_k)\|^2 + \|u_{k,\tau}\|^2 - \|\nabla f(w_k) - u_{k,\tau}\|^2\right).\end{gathered} \tag{34}$$

Additionally, we also have that for 2 vectors, x, y, we have that $\langle x, y\rangle \leq \frac{\lambda}{2}\|x\|^2 + \frac{1}{2\lambda}\|x\|^2$. Using $\lambda = \frac{1}{2}$ on B, we obtain:

$$\begin{aligned}-\left\langle\nabla f(w_k), \sum_{\tau=0}^{E-1}\eta_k b_{k,\tau}\right\rangle &= \eta_k\sum_{\tau=0}^{E-1}\langle\nabla f(w_k), -b_{k,\tau}\rangle \\ &\leq \eta_k\sum_{\tau=0}^{E-1}\left(\frac{1}{4}\|\nabla f(w_k)\|^2 + \|b_{k,\tau}\|^2\right), \\ &\leq \frac{\eta_k E}{4}\|\nabla f(w_k)\|^2 + \eta_k\sum_{\tau=0}^{E-1}\|b_{k,\tau}\|^2.\end{aligned} \tag{35}$$

Now, we have the following: $\mathbb{E}\left[\|x+y\|^2\right] \leq 2\mathbb{E}\left[\|x\|^2\right] + 2\mathbb{E}\left[\|y\|^2\right]$ and $\mathbb{E}\left[\|\sum_{i\in R}x_i\|^2\right] \leq |R|\sum_{i\in R}\mathbb{E}\left[\|x_i\|^2\right]$. Using these as well as (28) on term C, we obtain the following:

$$\mathbb{E}\left[\|\sum_{\tau=0}^{E-1}\hat{u}_{k,\tau}+b_{k,\tau}\|^2\right]\leq 2\mathbb{E}\left[\|\sum_{\tau=0}^{E-1}\hat{u}_{k,\tau}\|^2\right]+ 2\mathbb{E}\left[\|\sum_{\tau=0}^{E-1}b_{k,\tau}\|^2\right], \tag{36}$$

$$\leq 2E\sum_{\tau=0}^{E-1}\mathbb{E}\left[\|u_{k,\tau}\|^2\right]+\frac{2E\sigma^2}{n}+2E\sum_{\tau=0}^{E-1}\mathbb{E}\left[\|b_{k,\tau}\|^2\right]. \tag{37}$$

Using the results from (34) (35) (37) and plugging them on (33), we obtain the following:

$$\begin{gathered}\mathbb{E}\left[f\left(w_{k+1}\right)\right]\leq\mathbb{E}\left[f\left(w_k\right)\right]-\frac{\eta_k E}{4}\mathbb{E}\left[\|\nabla f\left(w_k\right)\|^2\right]+\\ \underbrace{\frac{\eta_k}{2}\sum_{\tau=0}^{E-1}\mathbb{E}\left[\|\nabla f\left(w_k\right)-u_{k,\tau}\|^2\right]}_{D}-\frac{\eta_k}{2}\left(1-\eta_k LE\right)\sum_{\tau=0}^{E-1}\mathbb{E}\left[\|u_{k,\tau}\|^2\right]\\ +\eta_k\left(1+\eta_k LE\right)\sum_{\tau=0}^{E-1}\mathbb{E}\left[\|b_{k,\tau}\|^2\right]+\frac{\eta_k^2 LE\sigma^2}{n}.\end{gathered} \tag{38}$$

Now, term D is upperbounded by Lemma 2, plugging this result on equation (38) and using (30), we get:

$$\begin{gathered}\mathbb{E}\left[f\left(w_{k+1}\right)\right]\leq\mathbb{E}\left[f\left(w_k\right)\right]-\frac{\eta_k E}{4}\mathbb{E}\left[\|\nabla f\left(w_k\right)\|^2\right]-\\ \frac{\eta_k}{2}\underbrace{\left(1-2\eta_k LE-\eta_k^2L^2E^2\right)}_{E}\sum_{\tau=0}^{E-1}\mathbb{E}\left[\|u_{k,\tau}\|^2\right]\\ +\eta_k^2LE^2\left(\frac{9\eta_k LE}{4}\right)\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\left[\|\nabla f_i\left(w_k\right)\|^2\right]\\ +\eta_k^2LE\left(\eta_k LE\left(\frac{1}{n}+\frac{3E}{4}\right)+\frac{1}{n}\right)\sigma^2\\ +\eta_k\left(1+\eta_k LE\right)\sum_{\tau=0}^{E-1}\mathbb{E}\left[\|b_{k,\tau}\|^2\right].\end{gathered} \tag{39}$$

We can see that term $E\geq 0$ for $\eta_k LE\leq\frac{1}{3}$. Due to the above, we drop this term and for $\eta_k LE\leq\frac{1}{3}$:

$$\begin{gathered}\mathbb{E}\left[f\left(w_{k+1}\right)\right]\leq\mathbb{E}\left[f\left(w_k\right)\right]-\frac{\eta_k E}{4}\mathbb{E}\left[\|\nabla f\left(w_k\right)\|^2\right]+\\ \eta_k^2LE^2\left(\frac{9\eta_k LE}{4}\right)\frac{1}{n}\sum_{i\in[n]}\mathbb{E}\left[\|\nabla f_i\left(w_k\right)\|^2\right]\\ +\eta_k^2LE\left(\frac{\eta_k LE}{2}\left(\frac{1}{n}+\frac{3E}{4}\right)+\frac{1}{n}\right)\sigma^2+\eta_k\left(1+\eta_k LE\right)\gamma.\end{gathered} \tag{40}$$

■

We can now use this result to obtain convergence guarantees as stated in Theorem 1. The following Lemma 2 helps us bound term D in Lemma 1, which we can think of as a bound on the expected error between the gradient at w_k with respect to the overall gradient computed at a given communication round k and local step τ.

Lemma 2. *With $\eta_k LE\leq\frac{1}{3}$, we get that:*

$$\begin{gathered}\sum_{\tau=0}^{E-1}\mathbb{E}\left[\|\nabla f\left(w_k\right)-u_{k,\tau}\|^2\right]\leq\eta_k^2L^2E^2\sum_{\tau=0}^{E-1}\mathbb{E}\left[\|u_{k,\tau}\|^2\right]+\\ \frac{9\eta_k^2L^2E^3}{2n}\sum_{i\in[n]}\mathbb{E}\left[\|\nabla f_i\left(w_k\right)\|^2\right]+\eta_k^2L^2E^2\left(\frac{1}{n}+\frac{3E}{4}\right)\sigma^2.\end{gathered} \tag{41}$$

Proof. Knowing that:

$$\mathbb{E}\left[\|\nabla f\left(w_k\right)-u_{k,\tau}\|^2\right]= \\ \mathbb{E}\left[\|\nabla f\left(w_k\right)-\nabla f\left(\bar{w}_{k,\tau}\right)+\nabla f\left(\bar{w}_{k,\tau}-u_{k,\tau}\right)\|^2\right], \tag{42}$$

$$\leq 2\mathbb{E}\left[\|\nabla f\left(w_k\right)-\nabla f\left(\bar{w}_{k,\tau}\right)\|^2\right]+ \\ 2\mathbb{E}\left[\|\nabla f\left(\bar{w}_{k,\tau}\right)-u_{k,\tau}\|^2\right], \tag{43}$$

$$\leq 2L^2\mathbb{E}\left[\|w_k-\bar{w}_{k,\tau}\|^2\right]+ \\ 2\mathbb{E}\left[\|\frac{1}{n}\sum_{i\in[n]}\underbrace{\left(\nabla f_i\left(\bar{w}_{k,\tau}\right)-\nabla f_i\left(w_{k,\tau}^{(i)}\right)\right)}_{=-\tilde{e}_{k,\tau}^{(i)}}\|^2\right], \tag{44}$$

$$\leq 2\eta_k^2L^2\mathbb{E}\left[\|\sum_{t=0}^{\tau-1}\hat{u}_{k,t}\|^2\right]+\frac{2}{n}\sum_{i\in[n]}\mathbb{E}\left[\|\tilde{e}_{k,\tau}^{(i)}\|^2\right], \tag{45}$$

$$\leq 2\eta_k^2L^2\left(\tau\sum_{t=0}^{\tau-1}\mathbb{E}\left[\|u_{k,t}\|^2\right]+\frac{\tau\sigma^2}{n}\right)+ \\ \frac{2L^2}{n}\sum_{i\in[n]}\mathbb{E}\left[\|w_{k,\tau}^{(i)}-\bar{w}_{k,\tau}^{(i)}\|^2\right]. \tag{46}$$

Where (44) comes about by f's L-smoothness and from the way $u_{k,\tau}$ is defined. Equation (45) comes about from equation (26), and equation (46) results from f_i's L-smoothness as well as equation (28).

However:

$$\sum_{i\in[n]}\mathbb{E}\left[\|w_{k,\tau}^{(i)}-\bar{w}_{k,\tau}\|^2\right]= \\ \sum_{i\in[n]}\mathbb{E}\left[\|\left(w_{k,0}^{(i)}-\eta_k\sum_{t=0}^{\tau-1}\hat{u}_{k,t}^{(i)}\right)-\left(\bar{w}_{k,0}-\eta_k\sum_{t=0}^{\tau-1}\hat{u}_{k,t}\right)\|^2\right], \tag{47}$$

$$=\eta_k^2\sum_{i\in[n]}\mathbb{E}\left[\|\sum_{t=0}^{\tau-1}\hat{u}_{k,t}-\sum_{t=0}^{\tau-1}\hat{u}_{k,t}^{(i)}\|^2\right], \tag{48}$$

$$\leq \eta_k^2\tau\sum_{i\in[n]}\sum_{t=0}^{\tau-1}\mathbb{E}\left[\|\hat{u}_{k,t}-\hat{u}_{k,t}^{(i)}\|^2\right], \tag{49}$$

$$=\eta_k^2\tau\sum_{i\in[n]}\sum_{t=0}^{\tau-1}\mathbb{E}\left[\|\hat{u}_{k,t}\|^2+\|\hat{u}_{k,t}^{(i)}\|^2-2\left\langle\hat{u}_{k,t},\hat{u}_{k,t}^{(i)}\right\rangle\right]. \tag{50}$$

Now, equation (48) results from $w_{k,0}^{(i)}=w_k\forall\, i\in[n]$, because of this $\bar{w}_{k,0}=w_k$. Additionally, $\hat{u}_{k,t}=\frac{1}{n}\sum_{i\in[n]}\hat{u}_{k,\tau}^{(i)}$, (50) simplifies to:

$$\sum_{i\in[n]}\mathbb{E}\left[\|w_{k,\tau}^{(i)}-\bar{w}_{k,\tau}\|^2\right] \\ \leq \eta_k^2\tau\sum_{i\in[n]}\sum_{t=0}^{\tau-1}\left(\mathbb{E}\left[\|\hat{u}_{k,\tau}^{(i)}\|\right]-\mathbb{E}\left[\|\hat{u}_{k,t}\|^2\right]\right), \tag{51}$$

$$\leq \eta_k^2\tau\sum_{i\in[n]}\sum_{t=0}^{\tau-1}\mathbb{E}\left[\|\hat{u}_{k,t}^{(i)}\|^2\right], \tag{52}$$

$$\leq \eta_k^2 \tau \sum_{i\in[n]} \sum_{t=0}^{\tau-1} \left(\mathbb{E}\left[\|\nabla f_i \left(w_{k,t}^{(i)} \right) \|^2 \right] + \sigma^2 \right). \tag{53}$$

Using the result from Lemma 3, we get:

$$\begin{aligned} &\sum_{i\in[n]} \mathbb{E}\left[\|w_{k,\tau}^{(i)} - w_{k,\tau}\|^2 \right] \\ &\leq \frac{9\eta_k^2\tau^2}{8} \sum_{i\in[n]} \left(2\mathbb{E}\left[\|\nabla f_i (w_k)\|^2 \right] + \sigma^2 \right). \end{aligned} \tag{54}$$

Plugging the result from (54) into (46), we obtain:

$$\begin{aligned} \mathbb{E}\left[\|\nabla f(w_k) - u_{k,\tau}\|^2 \right] \leq 2\eta_k^2 L^2 \left(\tau \sum_{t=0}^{\tau-1} \mathbb{E}\left[\|u_{k,t}\|^2 \right] + \frac{\tau\sigma^2}{n} \right) + \\ \frac{9\eta_k^2 L^2 \tau^2}{4n} \sum_{i\in[n]} \left(2\mathbb{E}\left[\|\nabla f_i (w_k)\|^2 \right] + \sigma^2 \right), \end{aligned} \tag{55}$$

$$\begin{aligned} = 2\eta_k^2 L^2 \tau \sum_{t=0}^{\tau-1} \mathbb{E}\left[\|u_{k,t}\|^2 \right] + \\ \frac{9\eta_k^2 L^2 \tau^2}{2n} \sum_{i\in[n]} \mathbb{E}\left[\|\nabla f_i (w_k)\|^2 \right] + \eta_k^2 L^2 \tau \sigma^2 \left(\frac{2}{n} + \frac{9\tau}{4} \right). \end{aligned} \tag{56}$$

Now, summing equation (56) for $\tau \in \{0, ..., E-1\}$, we get:

$$\begin{aligned} \sum_{\tau=0}^{E-1} \mathbb{E}\left[\|\nabla f(w_k) - u_{k,\tau}\|^2 \right] \leq \eta_k^2 L^2 E^2 \sum_{\tau=0}^{E-1} \mathbb{E}\left[\|u_{k,\tau}\|^2 \right] + \\ \frac{9\eta_k^2 L^2 E^3}{2n} \sum_{i\in[n]} \mathbb{E}\left[\|\nabla f_i (w_k)\|^2 \right] + \\ \eta_k^2 L^2 E^2 \left(\frac{1}{n} + \frac{3E}{4} \right) \sigma^2. \end{aligned} \tag{57}$$

■

We can now use this bound on term D from Lemma 1 and continue with our analysis. However, in stating Lemma 2, we needed the results from Lemma 3 which provide an upper bound on the expected value of the client's gradients at a given communication round k and local step τ.

Lemma 3. *With $\eta_k LE \leq \frac{1}{3}$, we get that:*

$$\sum_{t=0}^{\tau-1} \mathbb{E}\left[\|\nabla f_i \left(w_{k,t}^{(i)} \right) \|^2 \right] \leq \frac{\tau}{8} \left(18\mathbb{E}\left[\|\nabla f_i (w_k)\|^2 \right] + \sigma^2 \right). \tag{58}$$

Proof. We have that:

$$\begin{aligned} \mathbb{E}\left[\|\nabla f_i \left(w_{k,t}^{(i)} \right) \|^2 \right] = \\ \mathbb{E}\left[\|\nabla f_i \left(w_{k,t}^{(i)} \right) - \nabla f_i (w_k) + \nabla f_i (w_k)\|^2 \right], \end{aligned} \tag{59}$$

$$\leq 2\mathbb{E}\left[\|\nabla f_i (w_k)\|^2 \right] + 2\mathbb{E}\left[\|\nabla f_i \left(w_{k,t}^{(i)} \right) - \nabla f_i (w_k)\|^2 \right], \tag{60}$$

$$\leq 2\mathbb{E}\left[\|\nabla f_i (w_k)\|^2 \right] + 2L^2 \mathbb{E}\left[\|w_{k,t}^{(i)} - w_k\|^2 \right]. \tag{61}$$

But:

$$
\begin{aligned}
\mathbb{E}\left[\|w_k - w_{k,t}^{(i)}\|^2\right] &= \mathbb{E}\left[\|\eta_k \sum_{t'=0}^{t-1} \nabla \tilde{f}_i\left(w_{k,t'}^{(i)}; \mathcal{B}_{k,t'}^{(i)}\right)\|^2\right] \\
&\leq \eta_k^2 t \sum_{t'=0}^{t-1} \mathbb{E}\left[\|\eta_k \sum_{t'=0}^{t-1} \nabla \tilde{f}_i\left(w_{k,t'}^{(i)}; \mathcal{B}_{k,t'}^{(i)}\right)\|^2\right] \\
&\leq \eta_k^2 t \sum_{t'=0}^{t-1} \left(\mathbb{E}\left[\|\nabla f_i\left(w_{k,t'}^{(i)}\right)\|^2\right] + \sigma^2\right).
\end{aligned} \tag{62}
$$

Plugging this on (61), we obtain:

$$
\begin{aligned}
\mathbb{E}\left[\|\nabla f_i\left(w_{k,t}^{(i)}\right)\|^2\right] &\leq 2\mathbb{E}\left[\|\nabla f_i\left(w_k\right)\|^2\right] + \\
&2\eta_k^2 L^2 t \sum_{t'=0}^{t-1} \left(\mathbb{E}\left[\|\nabla f_i\left(w_{k,t'}^{(i)}\right)\|^2\right] + \sigma^2\right).
\end{aligned} \tag{63}
$$

Now, summing equation (63) by $t \in \{0, ..., \tau - 1\}$, we obtain:

$$
\begin{aligned}
\sum_{t=0}^{\tau-1} \mathbb{E}\left[\|\nabla f_i\left(w_{k,t}^{(i)}\right)\|^2\right] &\leq 2\tau\left(\mathbb{E}\left[\|\nabla f_i\left(w_k\right)\|^2\right]\right) + \\
&2\eta_k^2 L^2 \sum_{t=0}^{\tau-1} \tau \sum_{t'=0}^{t-1} \left(\mathbb{E}\left[\|\nabla f_i\left(w_{k,t'}^{(i)}\right)\|^2\right] + \sigma^2\right),
\end{aligned} \tag{64}
$$

$$
\begin{aligned}
&\leq 2\tau\left(\mathbb{E}\left[\|\nabla f_i\left(w_k\right)\|^2\right]\right) + \\
&\eta_k^2 L^2 \tau^2 \sum_{t=0}^{\tau-1} \left(\mathbb{E}\left[\|\nabla f_i\left(w_{k,t}^{(i)}\right)\|^2\right] + \sigma^2\right).
\end{aligned} \tag{65}
$$

Set $\eta_k LE \leq \frac{1}{3}$. Then:

$$
\begin{aligned}
\sum_{t=0}^{\tau-1} \mathbb{E}\left[\|\nabla f_i\left(w_{k,t}^{(i)}\right)\|^2\right] &\leq 2\tau\left(\mathbb{E}\left[\|\nabla f_i\left(w_k\right)\|^2\right]\right) + \\
&\frac{1}{9} \sum_{t=0}^{\tau-1} \left(\mathbb{E}\left[\|\nabla f_i\left(w_{k,t}^{(i)}\right)\|^2\right] + \sigma^2\right).
\end{aligned} \tag{66}
$$

Simplifying and re-arranging equation (66), we obtain:

$$
\sum_{t=0}^{\tau-1} \mathbb{E}\left[\|\nabla f_i\left(w_{k,t}^{(i)}\right)\|^2\right] \leq \frac{\tau}{8}\left(18\mathbb{E}\left[\|\nabla f_i\left(w_k\right)\|^2\right] + \sigma^2\right). \tag{67}
$$

We can now use the results from Lemma 3 to help provide a bound for Lemma 2. ■