Diverse Client Selection for Federated Learning: Submodularity and Convergence Analysis

Ravikumar Balakrishnan [*1] **Tian Li** [*2] **Tianyi Zhou** [*3] **Nageen Himayat** [1] **Virginia Smith** [2] **Jeffrey Bilmes** [3]

Abstract

In every communication round of federated learning, each client communicates its model updates back to the server which then aggregates them all. The incurred communication cost and overhead between clients and server, however, can be a major bottleneck particularly when the number of clients is large. We, in this paper, propose to select only a small diverse subset of clients, namely those carrying representative gradient information, and we transmit only these updates to the server. Our aim is for updating via only a subset to approximate updating via aggregating all client information. We achieve this by choosing a subset that maximizes a submodular facility location function defined over gradient space. We introduce "federated averaging with diverse client selection (`DivFL`)". We provide a thorough analysis of its convergence in the heterogeneous settings and apply it both to synthetic and to real datasets. Empirical results show our approach improves learning efficiency and encourages more uniform (i.e., fair) performance across clients.

1. Introduction

Federated learning (FL) studies the training of machine learning models on a sever for the sake of a swarm of clients each owning a limited amount of private local data. Recent approaches to this problem repeatedly alternate between device-local (stochastic) gradient descent steps and server-aggregation of the clients' model updates (McMahan et al., 2017). In cross-device settings, a server and its model usually serves billions of devices. Therefore, the communication between clients and the server can be costly and slow, forming an huge impediment to FL's viability.

One property of the collection of clients that can mitigate these problems, however, is often not exploited, and that is redundancy. Specifically, many clients might provide similar, and thus redundant, gradient information for updating the server model. Therefore, transmitting all such updates to the server is a waste of communication and computational resources. How best to select a representative and more informative client set while adhering to practical constraints in federated learning is still an open challenge. Although several selection criterion have been investigated in recent literature, e.g., sampling clients with probabilities proportional to their local dataset size (McMahan et al., 2017), sampling clients of larger update norm with higher probability (Chen et al., 2020), and selecting clients with higher losses (Cho et al., 2020), the redundancy and similarity of the clients' updates sent to the server is not represented and exploited in these approaches. In particular, communicating multiple clients' updates to the server may cause statistical and systems inefficiency if too many of them are too similar to each other. The commonly studied modular score/probability for each individual client is incapable of capturing information as a property over a *group* of clients. Ideally, a diverse set of clients would be selected, thereby increasing the impact of under-represented clients that contribute different information, and thereby improving fairness. This, in fact, is a topic of increasing interest (Mohri et al., 2019; Cho et al., 2020; Dennis et al., 2021).

In this paper, we introduce diversity to client selection in FL, namely a strategy to measure how a selected subset of clients can represent the whole when being aggregated on the server. Specifically, in each communication round, we aim to find a subset whose aggregated model update approximates the aggregated update over all clients. Inspired by the CRAIG method of coreset selection for efficient machine learning training (Mirzasoleiman et al., 2020), we derive an upper bound of the approximation error as a supermodular set function (in particular, the min-form of the facility location function (Cornuéjols et al., 1977)) evaluated on the selected subset. We can then apply submodular maximization (Fujishige, 2005; Iyer et al., 2013; Wei et al., 2014) on a complement submodular function to (approximately) minimize the error upper bound. We employ the greedy selection (Nemhauser et al., 1978) of a subset of

[*]Equal contribution [1]Intel Labs, USA [2]Machine Learning Department, Carnegie Mellon University [3]Department of Electrical and Computer Engineering, University of Washington, Seattle

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

clients according to the marginal gain of the submodular function to achieve a solution with provable approximation guarantee (Conforti and Cornuejols, 1984). By integrating the diverse client selection into the most commonly studied FL scheme, i.e., Federated Averaging (FedAvg) (McMahan et al., 2017), we propose `DivFL` that applies global model aggregation over a selected subset of clients after multiple local steps on every client. We present a theoretical convergence analysis of `DivFL` and show its tolerance to the heterogeneity of data distributions across clients and large number of local steps. In experiments, we compare `DivFL` with other client selection approaches on both synthetic dataset and FEMNIST, wherein our method excels on both the accuracy and fairness.

2. Background and Related Work

We consider a typical federated learning objective:

$$\min_{w} \; f(w) = \sum_{k=1}^{N} p_k F_k(w),$$

where for each device $k \in [N]$, p_k is a pre-defined weight (such that $\sum_{k=1}^{N} p_k = 1$) that can be set to $\frac{1}{N}$ or the fraction of training samples, and F_k is the client-specific empirical loss. While there are various possible modeling approaches, we consider this canonical objective of fitting a single global model to the non-identically distributed data across all devices (McMahan et al., 2017).

Client Selection in Federated Learning. Client[1] sampling is a critical problem particularly for cross-device settings where it is prohibitive to communicate with all devices. Two common (or default) strategies are (a) sampling the clients based on the number of local data points and uniformly averaging the model updates, and (b) sampling the clients uniformly at random and aggregating the model updates with weights proportional to the local samples (Li et al., 2020). There is also recent work proposing advanced sampling techniques to incorporate dynamic systems constraints, accelerate the convergence of federated optimization, or to obtain a better model with higher accuracy (Nishio and Yonetani, 2019; Ribero and Vikalo, 2020; Cho et al., 2020; Lai et al., 2020). We investigate client selection through the lens of encouraging client diversity at each communication round that largely remains unexplored in previous work. The closest client selection method to ours is based on clustering (e.g., selecting representative clients from separate clusters (Dennis et al., 2021)). We note that performing (private) clustering in federated settings is still an open problem, and our method can be viewed as a soft version of dynamic clustering at each round (discussed in the next paragraph). The benefits of gradient (or model) diversity has been demonstrated in other related contexts, such as scaling up mini-batch stochastic gradient descent (SGD) (Yin et al., 2018). Enforcing sample or gradient diversity during optimization also implicitly places more emphasis on the underrepresented subpopulations of clients, and can promote fairness defined as representative disparity (Hashimoto et al., 2018). Similar to previous work (e.g., Cho et al., 2020), we observe our approach yields more fair solutions across the network in Section 5.

[1]Following conventions, we use the term 'client' for the problem of client selection. Throughout the paper, we use 'devices' and 'clients' interchangeably.

Diverse Subset Selection via Submodularity. Modular scores have been widely studied for subset selection in machine learning and federated learning, e.g., a utility score for each sample or client often measured by the loss. However, the diversity of a subset cannot be fully captured by such modular scores since there is no score interaction. Diversity is often well modeled by a diminishing return property, i.e., the (marginal) gain an element brings to a subset diminishes as more elements added to the subset. There exists a rich and expressive family of functions, all of which are natural for measuring diversity, and all having the diminishing returns property: given a finite ground set V of size n, and any subset $A \subseteq B \subseteq V$ and a $v \notin B$, a set function $F : 2^V \to \mathbb{R}$ is submodular if

$$F(v \cup A) - F(A) \geq F(v \cup B) - F(B). \qquad (1)$$

This implies v is no less valuable to the smaller set A than to the larger set B. The marginal gain of v conditioned on A is denoted $f(v|A) \triangleq f(v \cup A) - f(A)$ and reflects the importance of v to A. Submodular functions (Fujishige, 2005) have been widely used for diversity models (Lin and Bilmes, 2011; Batra et al., 2012; Prasad et al., 2014; Gillenwater et al., 2012; Bilmes and Bai, 2017).

Maximizing a submodular function usually encourages the diversity and reduces the redundancy of a subset. This property has been utilized for data selection in active learning (Guillory and Bilmes, 2011), curriculum learning (Zhou and Bilmes, 2018), mini-batch partitioning (Wang et al., 2019), gradient approximation (Mirzasoleiman et al., 2020), etc. Although the number of possible subsets A is $\binom{n}{k}$, enumerating them all to find the maximum is intractable. Thanks to submodularity, fast approximate algorithms (Nemhauser et al., 1978; Minoux, 1978; Mirzasoleiman et al., 2015) exist to find an approximately optimal A with provable bounds (Nemhauser et al., 1978; Conforti and Cornuejols, 1984). Despite its success in data selection, submodularity has not been explored for client selection in federated learning. Encouraging diversity amongst local gradients (or model updates) of selected clients can effectively reduce redundant communication and promote fairness. Moreover, it raises several new challenges in the

FL setting, e.g., (1) it is unclear which submodular function to optimize and in which space to measure the similarity/diversity between clients; (2) What convergence guarantee can be obtained under practical assumptions such as heterogeneity among clients, and (3) What are the effects of outdated client selection due to communication constraints?

3. Diverse Client Selection

In this section, we introduce "federated averaging with diverse client selection" (or `DivFL`), a method that incorporates diverse client selection into the most widely studied FL scheme, federated averaging (FedAvg). We will first derive a combinatorial objective for client selection via an approximation of the full communication from all clients, which naturally morphs into a facility location function in the gradient space that can be optimized by submodular maximization. We then present the standard greedy algorithm that optimizes the objective by selecting a diverse subset of clients at every communication round.

3.1. Approximation of Full Communication

We aim to find a subset S of clients whose aggregated gradient can approximate the full aggregation over all the N clients $V = [N]$. To formulate this problem, we start by following the logic in (Mirzasoleiman et al., 2020). Given a subset S, we define a mapping $\sigma : V \to S$ such that the gradient information $\nabla F_k(v^k)$ from client k is approximated by that from a selected client $\sigma(k) \in S$. For $i \in S$, let $C_i \triangleq \{k \in V | \sigma(k) = i\}$ be the set of clients approximated by client-i and $\gamma_i \triangleq |C_i|$. The full aggregated gradient can be written as

$$\sum_{k\in[N]} \nabla F_k(v^k) = \sum_{k\in[N]} \left[\nabla F_k(v^k) - \nabla F_{\sigma(k)}(v^{\sigma(k)})\right] + \sum_{k\in S} \gamma_k \nabla F_k(v^k). \quad (2)$$

Subtracting the second term from both sides, taking the norms, and applying triangular inequality, we can obtain an upper bound for the approximation to the aggregated gradient by S, i.e.,

$$\left\| \sum_{k\in[N]} \nabla F_k(v^k) - \sum_{k\in S} \gamma_k \nabla F_k(v^k) \right\| \leq \sum_{k\in[N]} \left\| \nabla F_k(v^k) - \nabla F_{\sigma(k)}(v^{\sigma(k)}) \right\|. \quad (3)$$

The above inequality holds for any feasible mapping σ since the left hand side does not depend on σ. So we can take the minimum of the right hand side w.r.t. $\sigma(k),\ \forall k \in [N]$, i.e.,

$$\left\| \sum_{k\in[N]} \nabla F_k(v^k) - \sum_{k\in S} \gamma_k \nabla F_k(v^k) \right\| \leq \sum_{k\in[N]} \min_{i\in S} \left\| \nabla F_k(v^k) - \nabla F_i(v^i) \right\| \triangleq G(S). \quad (4)$$

The right hand side provides a relaxed objective $G(S)$ for minimizing the approximation error in the left hand. Minimizing $G(S)$ (or maximizing $\bar{G}$, a constant minus its negation) equals maximizing a well-known submodular function, i.e., the facility location function (Cornuéjols et al., 1977). To restrict the communication cost, we usually limit the number of selected clients to be no greater than K, i.e., $|S| \leq K$. This resorts to a submodular maximization problem under cardinality constraint, which is NP-hard but an approximation solution with $1 - e^{-1}$ bound can be achieved via the greedy algorithm (Nemhauser et al., 1978).

3.2. Greedy Selection of Clients

The naïve greedy algorithm for minimizing the upper bound of gradient approximation starts from $S \leftarrow \emptyset$, and adds one client $k \in V \backslash S$ with the greatest marginal gain to S in every step, i.e.,

$$S \leftarrow S \cup k^*, \quad k^* \in \operatorname*{argmax}_{k\in V\backslash S} [\bar{G}(S) - \bar{G}(\{k\} \cup S)] \quad (5)$$

until $|S| = K$. Although it requires to evaluate the marginal gain for all clients $k \in V \backslash S$ in every step, there exists several practical accelerated algorithms (Minoux, 1978; Mirzasoleiman et al., 2015) to substantially reduce the number of clients participating in the evaluation. To incorporate the client selection into any federated learning algorithm, we simply apply the greedy algorithm in each aggregation round and only aggregate the model updates over selected clients. The complete procedure is given in Algorithm 1 assuming the base algorithm is Federated Averaging (FedAvg) (McMahan et al., 2017).

In the left hand side of Eq. (3)-(4), we aim at approximating the full communication by a weighted sum over selected clients in S with weights $\{\gamma_i\}_{i\in S}$. However, since we relax the problem to minimizing its upper bound and the greedy solution does not guarantee to achieve the global minimum of the relaxed objective, the weight associated with the greedy solution S, i.e., $\gamma_i = |C_i|$ with $C_i = \{k \in V | i \in \operatorname{argmin}_{j\in S} \left\| \nabla F_k(v^k) - \nabla F_j(v^j) \right\|\}$, is sub-optimal. In fact, given S, the optimal weight $\{\gamma_i\}_{i\in S}$ can be achieved by directly minimizing the left hand side of Eq. (3)-(4) but it is infeasible because the full aggregation $\sum_{k\in[N]} \nabla F_k(v^k)$ is not available in our setting. Though there might exist better options, we find that simple uniform weights work promisingly in all evaluated scenarios of our

Algorithm 1 DivFL

Input: T, E, η, w_0

1 **for** $t = 0, \cdots, T-1$ **do**

2 Server selects a subset of K clients S_t by greedy algorithm in Eq. (5), and sends w_t to them.

3 **for** *device* $k \in S_t$ *in parallel* **do**

4 $w^k \leftarrow w_t$

5 Solve the local sub-problem of client-k inexactly by updating w^k for E local mini-batch SGD steps:
$$w^k = w^k - \eta \nabla F_k(w^k)$$

6 Send $\Delta_t^k := w_t^k - w_t$ back to Server

7 **end**

8 Server aggregates $\{\Delta_t^k\}$:
$$w_{t+1} \leftarrow w_t + \frac{1}{|S_t|} \sum_{k \in S_t} \Delta_t^k$$

9 **end**

10 **return** w_T

experiments. The greedy selection in line 2 of Algorithm 1 requires collecting the local gradients of all clients, which might be expensive in communication cost. In practice, we can reduce the cost by querying the gradients every m communication rounds. In experiments, we evaluate the performance under different m.

4. Convergence Analysis

In this section, we provide a novel theoretical analysis of the convergence behavior of Algorithm 1 for strongly convex problems under practical assumptions of non-identically distributed data, partial device participation, and local updating. Although the current analysis only holds for the proposed client selection algorithm applied to FedAvg, we believe that it can be extended to other federated learning methods as well in the future. We note that this FL analysis is new, it did not appear before as far as we know.

As discussed in Section 3.1, we draw connections between full gradient approximation and submodular function maximization. By solving a submodular maximization problem in the client selection procedure, we effectively guarantee that the approximation error is small (see Eq. (4)). We state an assumption on this below.

Assumption 1 (Gradient approximation error)**.** *At each communication round t, we assume the server selects a set S_t of devices such that their aggregated gradients (with weights $\{\gamma_k\}_{k \in S_t}$) is a good approximation of the full gradients on all devices with error ϵ, i.e.,*

$$\left\| \frac{1}{N} \sum_{k \in S_t} \gamma_k \nabla F_k(v_t^k) - \frac{1}{N} \sum_{k \in [N]} \nabla F_k(v_t^k) \right\| \leq \epsilon.$$

The same assumption has been studied in previous works on coreset selection for mini-batch SGD (Mirzasoleiman et al., 2020). Next, we state other assumptions used in our proof, which are standard in the federated optimization literature (e.g., Li et al., 2019).

Assumption 2. *Each F_k ($k \in [N]$) is L-smooth.*

Assumption 3. *Each F_k ($k \in [N]$) is μ-strongly convex.*

Assumption 4. *For $k \in [N]$, in-device variance of stochastic gradients on random samples ζ are bounded, i.e., $\mathbb{E}[\|\nabla F_k(w_t^k, \zeta) - \nabla F_k(w_t^k)\|^2] \leq \sigma^2$.*

Assumption 5. *For $k \in [N]$, the stochastic gradients on random samples ζ are uniformly bounded, i.e., $\|\nabla F_k(w_t^k, \zeta)\|^2 \leq G^2$.*

Assumption 6 (Bounded heterogeneity)**.** *Statistical heterogeneity defined as $F^* - \sum_{i \in [N]} p_k F_k^*$ is bounded by C, where $F^* := \min_w f(w)$ and $F_k^* := \min_v F_k(v)$.*

Let $w^* \in \operatorname{argmin}_w f(w)$ and $v_k^* \in \operatorname{argmin}_v F_k(v)$ for $k \in [N]$. Note that under Assumption 3 (μ-strongly convexity), Assumption 6 implies that $\| \sum_{k \in [N]} p_k v_k^* - w^* \|$ is bounded by a constant (which we denote as M) observing that

$$\left\| \sum_{k \in [N]} p_k v_k^* - w^* \right\| \leq \sum_{k \in [N]} p_k \|w^* - v_k^*\|$$

$$\leq \sum_{k \in [N]} p_k \left(1 + \|w^* - v_k^*\|^2\right) \leq 1 + \frac{2}{\mu} \left(F^* - \sum_{k \in [N]} p_k F_k^* \right).$$

Setup. Following Li et al. (2019), we flatten local SGD iterations at each communication round, and index gradient evaluation steps with t (slightly abusing notation). We define *virtual* sequences $\{v_t^k\}_{k \in [N]}$ and $\{w_t^k\}_{k \in [N]}$ where

$$v_{t+1}^k = w_t^k - \eta_t \nabla F_k(w_t^k)$$

$$w_{t+1}^k = \begin{cases} v_{t+1}^k, \text{ if not aggregate,} \\ \text{select } S_{t+1} \text{ and average } \{v_{t+1}^k\}_{k \in S_{t+1}}, \text{ otherwise.} \end{cases}$$

While all devices virtually participate in the updates of $\{v_t^k\}$ at each *virtual* iteration t, the effective updating rule of $\{w_t^k\}$ is the same as that in Algorithm 1. Further, let

$$\overline{v}_t := \sum_{k \in [N]} p_k v_k^t, \quad \overline{w}_t := \sum_{k \in [N]} p_k w_t^k.$$

Therefore,

$$\overline{w}_t = \begin{cases} \overline{v}_t & \text{if not aggregate,} \\ \frac{1}{K} \sum_{k \in S_t} v_t^k & \text{otherwise.} \end{cases}$$

Denote the aggregated stochastic gradient over all clients as g_t, i.e.,

$$\overline{v}_{t+1} = \overline{w}_t - \eta_t \left(\sum_{k \in [N]} p_k \nabla F_k(w_t^k, \zeta_t^k) \right) := \overline{w}_t - \eta_t g_t.$$

We aim at approximating $\overline{v}_{t+1}$ by $\overline{w}_{t+1}$ (when aggregating) and next state a main lemma bounding $\|\overline{w}_{t+1} - \overline{v}_{t+1}\|$.

Lemma 1. *For any virtual iteration t, under Algorithm 1 and Assumptions 1-6, we have*

$$\begin{aligned}\|\overline{w}_{t+1} - \overline{v}_{t+1}\| \leq & LGE(E-1)\left(1+\frac{E-1}{t+\gamma-(E-1)}\right)^2\eta_t^2 \\ & + E\epsilon\left(1+\frac{E-1}{t+\gamma-(E-1)}\right)\eta_t.\end{aligned}$$

Proof. We only need to consider the aggregation step t. Let the last time of aggregation happens at step $t_0 = t + 1 - E$, when we select a subset S (associated with weights $\{\gamma_k\}_{k\in S}$) using the greedy algorithm. Under the updating rule, the approximation in the E steps between t_0 and t fulfills

$$\|\overline{w}_{t+1} - \overline{v}_{t+1}\| \leq \sum_{\tau=t_0}^{t}\left\|\frac{1}{N}\sum_{k\in S}\gamma_k\Delta v_\tau^k - \frac{1}{N}\sum_{k\in[N]}\Delta v_\tau^k\right\|,$$

where $\Delta v_\tau^k \triangleq -\eta_\tau \nabla F_k(v_\tau^k)$. Under Assumption 1, we have that the approximation of the full gradients at each communication round t_0 satisfies

$$\left\|\frac{1}{N}\sum_{k\in S}\gamma_k\nabla F_k(v_{t_0}^k) - \frac{1}{N}\sum_{k\in[N]}\nabla F_k(v_{t_0}^k)\right\| \leq \epsilon,$$

For every local step $\tau \in (t_0, t]$, we use the same S to approximate the full gradient because we only communicate the local gradients every E local steps. Note

$$\begin{aligned}&\left\|\frac{1}{N}\sum_{k\in S}\gamma_k\nabla F_k(v_\tau^k) - \frac{1}{N}\sum_{k\in[N]}\nabla F_k(v_\tau^k)\right\| \\ \leq &\left\|\frac{1}{N}\sum_{k\in S}\gamma_k\nabla F_k(v_\tau^k) - \frac{1}{N}\sum_{k\in S}\gamma_k\nabla F_k(v_{t_0}^k)\right\| + \\ &\left\|\frac{1}{N}\sum_{k\in S}\gamma_k\nabla F_k(v_{t_0}^k) - \frac{1}{N}\sum_{k\in[N]}\nabla F_k(v_{t_0}^k)\right\| + \\ &\left\|\frac{1}{N}\sum_{k\in[N]}\nabla F_k(v_\tau^k) - \frac{1}{N}\sum_{k\in[N]}\nabla F_k(v_{t_0}^k)\right\| \\ \leq & 2LG\sum_{\nu=t_0}^{\tau}\eta_\nu + \epsilon,\end{aligned}$$

where the first and the third term on the right hand side are bounded using the L-smoothness of $F_k(\cdot)$ and G-bounded norm of its stochastic gradient. Hence,

$$\begin{aligned}&\|\overline{w}_{t+1} - \overline{v}_{t+1}\| \leq \sum_{\tau=t_0}^{t}\left\|\frac{1}{N}\sum_{k\in S}\gamma_k\Delta v_\tau^k - \frac{1}{N}\sum_{k\in[N]}\Delta v_\tau^k\right\| \\ =&\sum_{\tau=t_0}^{t}\eta_\tau\left\|\frac{1}{N}\sum_{k\in S}\gamma_k\nabla F_k(v_\tau^k) - \frac{1}{N}\sum_{k\in[N]}\nabla F_k(v_\tau^k)\right\| \\ \leq & LGE(E-1)\eta_{t_0}^2 + E\epsilon\eta_{t_0} \\ = & LGE(E-1)\left(1+\frac{E-1}{t+\gamma-(E-1)}\right)^2\eta_t^2 + \\ & E\epsilon\left(1+\frac{E-1}{t+\gamma-(E-1)}\right)\eta_t\end{aligned}$$

□

With Lemma 1, we state our convergence results as follows.

Theorem 1 (Convergence of Algorithm 1)**.** *Under Assumptions 1-6, we have*

$$\mathbb{E}[\|w^* - w_t\|^2] \leq O(1/t) + O(\epsilon).$$

In experiments, we observe that `DivFL` allows us to achieve faster convergence (empirically) at the cost of additional solution bias (a non-diminishing term dependent on ϵ).

We provide a sketch of the proof here and defer complete analysis to Appendix A. Examine the distances between $\overline{w}_{t+1}$ and w^*,

$$\begin{aligned}\|\overline{w}_{t+1} - w^*\|^2 = & \|\overline{w}_{t+1} - \overline{v}_{t+1}\|^2 + \|\overline{v}_{t+1} - w^*\|^2 + \\ & 2\langle\overline{w}_{t+1} - \overline{v}_{t+1}, \overline{v}_{t+1} - w^*\rangle.\end{aligned}$$

If iteration t is not an aggregation step, $\overline{w}_{t+1} = \overline{v}_{t+1}$ and

$$\|\overline{w}_{t+1} - w^*\|^2 = \|\overline{v}_{t+1} - w^*\|^2,$$

which we can bound with Lemma 1 in Li et al. (2019):

$$\mathbb{E}[\|\overline{v}_{t+1} - w^*\|^2] \leq (1-\eta_t\mu)\mathbb{E}[\|\overline{w}_t - w^*\|^2] + \eta_t^2 B \tag{6}$$

for some constant B. If t is an aggregation step, we need to bound

$$\begin{aligned}\mathbb{E}[\|\overline{w}_{t+1} - \overline{v}_{t+1}\|^2] + \mathbb{E}[\|\overline{v}_{t+1} - w^*\|^2] + \\ 2\mathbb{E}[\langle\overline{w}_{t+1} - \overline{v}_{t+1}, \overline{v}_{t+1} - w^*\rangle].\end{aligned}$$

The second term can be bounded by Eq. (6), which contains $(1-\eta_t\mu)\mathbb{E}[\|\overline{w}_t - w^*\|^2]$. Therefore, combined with Lemma 1, with a decaying step size, we can obtain a recursion on $\mathbb{E}[\|\overline{w}_{t+1} - w^*\|^2]$ which leads to Theorem 1. We provide the complete proof in Appendix A.

5. Experiments

Setup. We evaluate the diverse client selection approach utilizing both synthetic and real federated data—Federated Extended MNIST (FEMNIST) (Cohen et al., 2017). The synthetic data enable us to control the heterogeneity across clients for evaluation. In the case of FEMNIST, we only utilize the 10 lowercase character classes of handwritten alphanumeric characters and assign each client the data drawn from 2 classes. We consider two baselines: a) random sampling without replacement, and b) the power-of-choice approach (Cho et al., 2020) where we first sample a random subset of clients, and then select the devices with the largest training losses. For DivFL, we employ the heuristic of querying every device in the network for the gradients every m communication rounds to estimate the dissimilarity, and choose m to be 1 or 10. While the former provides the upper bound on the performance, the latter approach and other variants are amenable in more realistic settings. We describe the results on datasets below.

5.1. Results on the Synthetic Dataset

We generate synthetic data following the setup described in Li et al. (2020). The parameters and data are generated from Gaussian distributions and the model is logistic regression. $y = \text{argmax}(\text{softmax}(W^T X + b))$. We consider a total of 30 clients where the local dataset sizes for each client follows the power law. We set the mini batch-size to 10 and the learning rate $\eta = 0.01$.

For all methods, we effectively select $K = 5$ clients at each communication round. For power-of-choice, we first randomly sample 24 devices, and further select 5 out of them with the largest training losses. We report training loss and test accuracy versus the number of communication rounds in Figure 1 for the synthetic IID setting. We observe two key benefits of DivFL compared to random sampling and power-of-choice approaches. On the one hand, DivFL achieves a significant convergence speedup ($\sim 10\times$ faster) to reach the same loss and accuracy relative to random sampling and power-of-choice. Furthermore, DivFL also achieve the lowest loss and highest accuracy among the client selection approaches. As one would expect, the choice of m affects the convergence as well as the overall loss/accuracy with a larger value of m resulting in slightly higher loss/lower accuracy (still outperforming the baselines). More results are presented in Appendix B.1.

We report the training loss and testing accuracy for the synthetic non-IID dataset in Figure 2. In this case, we note that both the power-of-choice approach and DivFL outperform random sampling on the two metrics. However, DivFL is more robust to outdated client set selection when increasing the local SGD steps E. Specifically, we note that as we increase the number of local epochs (5 in this case), the two metrics degrade for both power-of-choice and DivFL but the former suffers more degeneration: DivFL continues to outperform random sampling but the power-of-choice approach becomes worse than random sampling. As shown in Appendix B.2, DivFL provides the best trade-off between mean and variance of the test accuracy. A comprehensive set of results for different numbers of local epochs τ and different K are given in Appendix B.2.

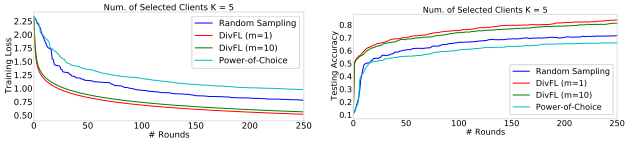

Figure 1. Training loss and test accuracy of DivFL compared with random sampling and power-of-choice on the synthetic IID data. For DivFL, we collect the gradients from all devices every m communication round to estimate the dissimilarity matrix. We see that DivFL achieves faster convergence and converges to more accurate solutions than all baselines in terms of convergence.

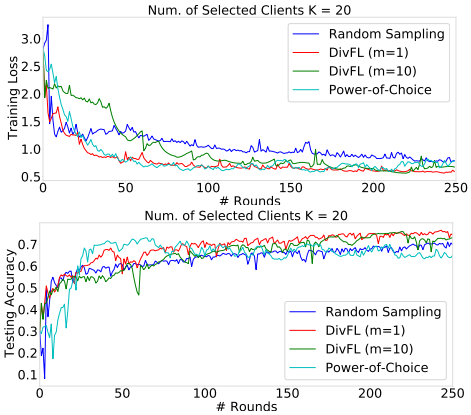

Figure 2. Training loss and test accuracy of DivFL compared with random sampling and power-of-choice on synthetic non-IID data. DivFL converges to more accurate solutions than all baselines.

5.2. Results on FEMNIST

We also evaluate DivFL on a real dataset FEMNIST with 500 clients. In Figure 3, we show the training loss and test accuracy of different approaches where $K = 10$ clients are selected in each round. We train a CNN model with two 5×5-convolutional and 2×2-maxpooling (with a stride of 2) layers followed by a dense layer with 128 activations.

As in the case of synthetic data, DivFL clearly achieves a higher accuracy and lower loss than the random sampling approach. Comparing to the power-of-choice approach, its test accuracy is marginally higher and the loss is comparable. Moreover, DivFL converges faster than the baselines. Under the choice of $K = 10$, there is no significant difference between $m = 1$ and $m = 10$, indicating the robustness/tolerance to the outdated client selection. DivFL

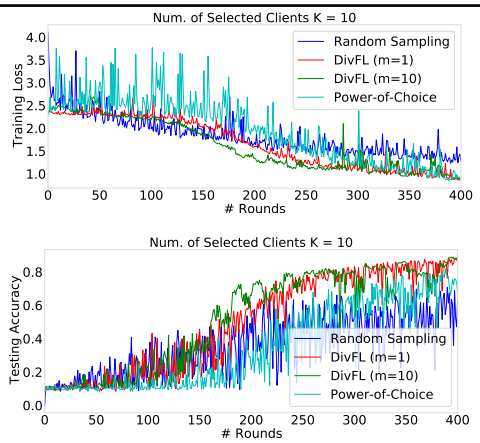

Figure 3. Training loss and test accuracy of `DivFL` compared with random sampling and power-of-choice on FEMNIST. `DivFL` converges faster to more accurate solutions than all baselines.

also achieves lower variance of test accuracy than the baselines (see Figure 9 in Appendix B.3). We also test different choices of number of selected clients K, and observe the consistent improvement of `DivFL`. The additional results are presented in Appendix B.3.

References

D. Batra, P. Yadollahpour, A. Guzman-Rivera, and G. Shakhnarovich. Diverse m-best solutions in markov random fields. In *ECCV*, pages 1–16, 2012.

J. A. Bilmes and W. Bai. Deep submodular functions. *CoRR*, abs/1701.08939, 2017. URL http://arxiv.org/abs/1701.08939.

W. Chen, S. Horvath, and P. Richtarik. Optimal client sampling for federated learning, 2020.

Y. J. Cho, J. Wang, and G. Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies, 2020.

G. Cohen, S. Afshar, J. Tapson, and A. van Schaik. Emnist: an extension of mnist to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017.

M. Conforti and G. Cornuejols. Submodular set functions, matroids and the greedy algorithm: Tight worst-case bounds and some generalizations of the rado-edmonds theorem. *Discrete Applied Mathematics*, 7(3):251–274, 1984.

G. Cornuéjols, M. Fisher, and G. Nemhauser. On the uncapacitated location problem. *Annals of Discrete Mathematics*, 1:163–177, 1977.

D. K. Dennis, T. Li, and V. Smith. Heterogeneity for the win: One-shot federated clustering. *arXiv preprint arXiv:2103.00697*, 2021.

S. Fujishige. *Submodular functions and optimization*. Annals of discrete mathematics. Elsevier, 2005.

J. Gillenwater, A. Kulesza, and B. Taskar. Near-optimal map inference for determinantal point processes. In *NeurIPS*, pages 2735–2743, 2012.

A. Guillory and J. Bilmes. Active semi-supervised learning using submodular functions. In *Uncertainty in Artificial Intelligence (UAI)*, Barcelona, Spain, July 2011. AUAI.

T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 2018.

R. Iyer, S. Jegelka, and J. A. Bilmes. Fast semidifferential-based submodular function optimization. In *ICML*, 2013.

F. Lai, X. Zhu, H. V. Madhyastha, and M. Chowdhury. Oort: Efficient federated learning via guided participant selection. *arxiv. org/abs/2010.06081*, 2020.

T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2020.

X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

H. Lin and J. Bilmes. A class of submodular functions for document summarization. In *ACL*, pages 510–520, 2011.

B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *International Conference on Artificial Intelligence and Statistics*, 2017.

M. Minoux. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques*, volume 7 of *Lecture Notes in Control and Information Sciences*, chapter 27, pages 234–243. Springer Berlin Heidelberg, 1978.

B. Mirzasoleiman, A. Badanidiyuru, A. Karbasi, J. Vondrák, and A. Krause. Lazier than lazy greedy. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 1812–1818, 2015.

B. Mirzasoleiman, J. Bilmes, and J. Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, 2020.

M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, 2019.

G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions. *Mathematical Programming*, 14(1):265–294, 1978.

T. Nishio and R. Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019-2019 IEEE International Conference on Communications*, 2019.

A. Prasad, S. Jegelka, and D. Batra. Submodular meets structured: Finding diverse subsets in exponentially-large structured item sets. In *NeurIPS*, pages 2645–2653, 2014.

M. Ribero and H. Vikalo. Communication-efficient federated learning via optimal client sampling. *arXiv preprint arXiv:2007.15197*, 2020.

S. Wang, W. Bai, C. Lavania, and J. Bilmes. Fixing mini-batch sequences with hierarchical robust partitioning. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89, pages 3352–3361, 2019.

K. Wei, R. Iyer, and J. Bilmes. Fast multi-stage submodular maximization. In *ICML*, 2014.

D. Yin, A. Pananjady, M. Lam, D. Papailiopoulos, K. Ramchandran, and P. Bartlett. Gradient diversity: a key ingredient for scalable distributed learning. In *International Conference on Artificial Intelligence and Statistics*, 2018.

T. Zhou and J. Bilmes. Minimax curriculum learning: Machine teaching with desirable difficulties and scheduled diversity. In *ICLR*, 2018.

Appendix

A. Complete Convergence Analysis

Below shows the convergence analysis. Many steps make a distinction between if we are doing an aggregating steps (from the clients to the server), or not (when the clients do not communicate). We assume that we aggregate every E time steps. Define virtual sequences $\{v_t^k\}_{k\in[N]}$ and $\{w_t^k\}_{k\in[N]}$ where for all $k \in [N]$,

$$v_{t+1}^k = w_t^k - \eta_t \nabla F_k(w_t^k) \tag{7}$$

$$w_{t+1}^k = \begin{cases} v_{t+1}^k & \text{if not aggregating,} \\ \text{sample } S_{t+1} \text{ and average } \{v_{t+1}^k\}_{k\in S_{t+1}} & \text{otherwise.} \end{cases} \tag{8}$$

Let

$$\overline{v}_t := \sum_{k\in[N]} p_k v_k^t, \tag{9}$$

$$\overline{w}_t := \sum_{k\in[N]} p_k w_t^k. \tag{10}$$

where $p_k \geq 0$ is the given weight of the k^{th} client and w.l.o.g., we assume $\sum_k p_k = 1$. Therefore,

$$\overline{w}_t = \begin{cases} \overline{v}_t & \text{if not aggregating, i.e., when } t \neq \ell E \text{ for some integer } \ell, \\ \frac{1}{K}\sum_{l\in S_t} v_t^l & \text{otherwise,} \end{cases} \tag{11}$$

and

$$\overline{v}_{t+1} = \overline{w}_t - \eta_t \left(\sum_{k\in[N]} p_k F_k(w_t^k, \zeta_t^k) \right) := \overline{w}_t - \eta_t g_t. \tag{12}$$

We have

$$\|\overline{w}_{t+1} - w^*\|^2 = \|\overline{w}_{t+1} - \overline{v}_{t+1} + \overline{v}_{t+1} - w^*\|^2 \tag{13}$$

$$= \|\overline{w}_{t+1} - \overline{v}_{t+1}\|^2 + \|\overline{v}_{t+1} - w^*\|^2 + 2\langle \overline{w}_{t+1} - \overline{v}_{t+1}, \overline{v}_{t+1} - w^* \rangle. \tag{14}$$

If not aggregating,

$$\overline{w}_{t+1} = \overline{v}_{t+1}. \tag{15}$$

Hence

$$\|\overline{w}_{t+1} - w^*\|^2 = \|\overline{v}_{t+1} - w^*\|^2. \tag{16}$$

Using Lemma 1 in (Li et al., 2019), we know $\mathbb{E}[\|\overline{v}_{t+1} - w^*\|^2] \leq (1 - \eta_t\mu)\mathbb{E}[\|\overline{w}_t - w^*\|^2] + \eta_t^2 B$ holds for some constant B. If we are aggregating, we need to bound

$$\mathbb{E}[\|\overline{w}_{t+1} - \overline{v}_{t+1}\|^2] + \mathbb{E}[\|\overline{v}_{t+1} - w^*\|^2] + 2\mathbb{E}[\langle \overline{w}_{t+1} - \overline{v}_{t+1}, \overline{v}_{t+1} - w^* \rangle]. \tag{17}$$

Let the last time of aggregation happens at step $t_0 = t + 1 - E$, when we select a subset S (associated with weights $\{\gamma_k\}_{k\in S}$) using the greedy algorithm. To bound the first term above,

$$\|\overline{w}_{t+1} - \overline{v}_{t+1}\| = \left\| \left(\overline{w}_{t_0} + \frac{1}{N}\sum_{k\in S} \gamma_k \sum_{\tau=t_0}^{t} \Delta v_\tau^k \right) - \left(\overline{w}_{t_0} + \frac{1}{N}\sum_{k\in[N]} \sum_{\tau=t_0}^{t} \Delta v_\tau^k \right) \right\| \tag{18}$$

$$= \left\| \sum_{\tau=t_0}^{t} \left(\frac{1}{N}\sum_{k\in S} \gamma_k \Delta v_\tau^k - \frac{1}{N}\sum_{k\in[N]} \Delta v_\tau^k \right) \right\| \tag{19}$$

$$\leq \sum_{\tau=t_0}^{t} \left\| \frac{1}{N}\sum_{k\in S} \gamma_k \Delta v_\tau^k - \frac{1}{N}\sum_{k\in[N]} \Delta v_\tau^k \right\| \tag{20}$$

Similar to the CRAIG paper (Mirzasoleiman et al., 2020), we assume that the subset S selected in step $t_0 = t + 1 - E$ provides an approximation of the full gradient such that

$$\left\| \frac{1}{N} \sum_{k \in S} \gamma_k \nabla F_k(v_{t_0}^k) - \frac{1}{N} \sum_{k \in [N]} \nabla F_k(v_{t_0}^k) \right\| \leq \epsilon, \tag{21}$$

For every local step $\tau \in (t_0, t]$, we use the same S to approximate the full gradient because we only communicate the local gradients every E local steps. To bound the gradient approximation at step τ using the stale S, we have

$$\begin{aligned}
\left\| \frac{1}{N} \sum_{k \in S} \gamma_k \nabla F_k(v_\tau^k) - \frac{1}{N} \sum_{k \in [N]} \nabla F_k(v_\tau^k) \right\| \leq & \left\| \frac{1}{N} \sum_{k \in S} \gamma_k \nabla F_k(v_\tau^k) - \frac{1}{N} \sum_{k \in S} \gamma_k \nabla F_k(v_{t_0}^k) \right\| + && (22) \\
& \left\| \frac{1}{N} \sum_{k \in S} \gamma_k \nabla F_k(v_{t_0}^k) - \frac{1}{N} \sum_{k \in [N]} \nabla F_k(v_{t_0}^k) \right\| + && (23) \\
& \left\| \frac{1}{N} \sum_{k \in [N]} \nabla F_k(v_\tau^k) - \frac{1}{N} \sum_{k \in [N]} \nabla F_k(v_{t_0}^k) \right\| && (24) \\
\leq & \, 2LG \sum_{\nu=t_0}^{\tau} \eta_\nu + \epsilon, && (25)
\end{aligned}$$

where the first and the third term on the right hand side are bounded using the L-smoothness of $F_k(\cdot)$ and G-bounded norm of its stochastic gradient. Hence, we can continue to bound the first term in Eq. (20) by

$$\begin{aligned}
\|\overline{w}_{t+1} - \overline{v}_{t+1}\| &\leq \sum_{\tau=t_0}^{t} \left\| \frac{1}{N} \sum_{k \in S} \gamma_k \Delta v_\tau^k - \frac{1}{N} \sum_{k \in [N]} \Delta v_\tau^k \right\| && (26) \\
&= \sum_{\tau=t_0}^{t} \eta_\tau \left\| \frac{1}{N} \sum_{k \in S} \gamma_k \nabla F_k(v_\tau^k) - \frac{1}{N} \sum_{k \in [N]} \nabla F_k(v_\tau^k) \right\| && (27) \\
&\leq 2LG \sum_{\tau=t_0}^{t} \sum_{\nu=t_0}^{\tau} \eta_\tau \eta_\nu + E\epsilon\eta_\tau && (28) \\
&\leq LGE(E-1)\eta_{t_0}^2 + E\epsilon\eta_{t_0} && (29) \\
&= LGE(E-1) \left(1 + \frac{E-1}{t+\gamma-(E-1)}\right)^2 \eta_t^2 + E\epsilon \left(1 + \frac{E-1}{t+\gamma-(E-1)}\right) \eta_t && (30)
\end{aligned}$$

where E is the number of local steps between two communication (aggregation) rounds. Therefore, Eq. (17) can be bounded as follows:

$$\begin{aligned}
& \mathbb{E}[\|\overline{w}_{t+1} - w^*\|^2] && (31) \\
\leq & \mathbb{E}[\|\overline{w}_{t+1} - \overline{v}_{t+1}\|^2] + \mathbb{E}[\|\overline{v}_{t+1} - w^*\|^2] + 2\mathbb{E}[\langle \overline{w}_{t+1} - \overline{v}_{t+1}, \overline{v}_{t+1} - w^* \rangle] && (32) \\
\leq & \left(LGE(E-1)\eta_{t_0}^2 + E\epsilon\eta_{t_0}\right)^2 + \left[(1-\eta_t\mu)\mathbb{E}[\|\overline{w}_t - w^*\|^2] + \eta_t^2 B\right] + 2\left(LGE(E-1)\eta_{t_0}^2 + E\epsilon\eta_{t_0}\right)\mathbb{E}[\|\overline{v}_{t+1} - w^*\|] && (33) \\
\leq & (1-\eta_t\mu)\mathbb{E}[\|\overline{w}_t - w^*\|^2] + E\epsilon\rho\eta_{t_0} + \left[LGE(E-1)\rho + (LGE(E-1)\eta_{t_0} + E\epsilon)^2\right]\eta_{t_0}^2 + B\eta_t^2 && (34) \\
\leq & (1-\eta_t\mu)\mathbb{E}[\|\overline{w}_t - w^*\|^2] + \epsilon\rho E\eta_{t_0} + \left[LG\rho + (LGE\eta_{t_0} + \epsilon)^2\right]E^2\eta_{t_0}^2 + B\eta_t^2, && (35)
\end{aligned}$$

where $\eta_t = \frac{\beta}{t+\gamma}$, $\eta_{t_0} = \frac{\beta}{t-E+1+\gamma}$ and $\mathbb{E}[\|\overline{v}_{t+1} - w^*\|] \leq \rho$.

$$\mathbb{E}\left\|\overline{v}^{t+1} - w^*\right\| \leq \mathbb{E}\left\|\overline{v}^{t+1} - \sum_{i\in[N]} p_i v_i^*\right\| + \mathbb{E}\left\|\sum_{i\in[N]} p_i v_i^* - w^*\right\| \tag{36}$$

$$\leq \mathbb{E}\left\|\overline{v}^{t+1} - \sum_{i\in[N]} p_i v_i^*\right\| + M \tag{37}$$

$$\leq \sum_{i\in[N]} \mathbb{E}\|p_i(\overline{v}_i^{t+1} - v_i^*)\| + M \tag{38}$$

$$\leq \sum_{i\in[N]} \frac{p_i}{\mu}\mathbb{E}\left\|\nabla f_i(v_i^t)\right\| + M \tag{39}$$

$$\leq \frac{G}{\mu} + M \leq \rho. \tag{40}$$

The final convergence rates follows from Lemma 3 in Mirzasoleiman et al. (2020).

B. Additional Experiments

B.1. Synthetic IID Dataset

The average test accuracies and the variances of test accuracies on synthetic IID dataset are presented in Figure 4 and 5. The test performance is observed for different choices of number of local epochs τ as well as the number of clients K participating in each global round. For all choices of (τ, K), both variations of `DivFL`, i.e., `DivFL` (m=1) and `DivFL` (m=10) achieve higher average and lower variance of accuracies. The relative gains for `DivFL` w.r.t Random sampling and Power-of-Choice approaches vary, however, depending on the choice of (τ, K). For example, for small values of K, τ, the relative gains of `DivFL` are the highest. The relative gains start diminishing as both (τ, K) increase. This is explained by the fact that as K increases, the similarity between all client selection approaches start increasing. Likewise, as the number of local iterations τ increases, the submodular metrics become increasingly suboptimal leading to diminished gains. It is, however, important to note that the two variants of `DivFL` do not have a significant difference in the performance except for the case $\tau = 1, K = 5$. This is highly encouraging since this shows that in a practical setting, `DivFL` can be implemented with submodular metric exchanged only periodically between clients and server, thereby reducing computational and communication overheads.

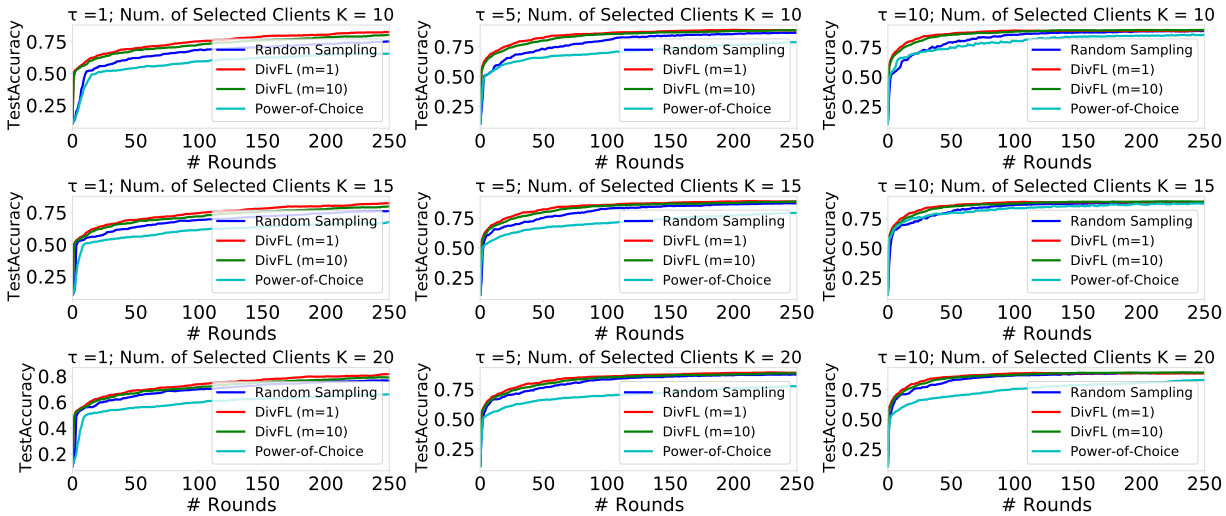

Figure 4. Average Test Accuracy on Synthetic IID

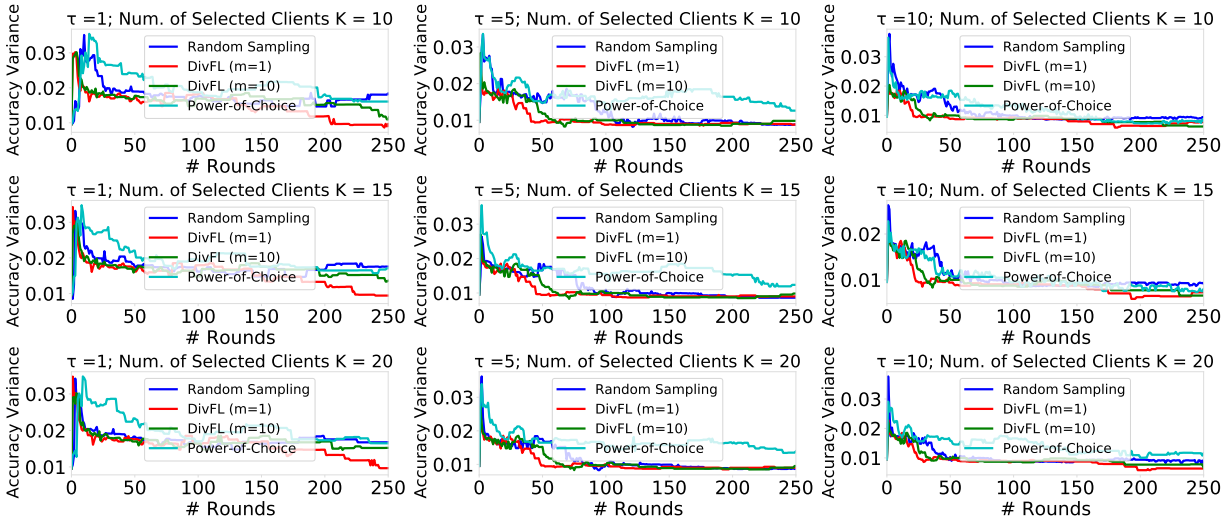

Figure 5. Variance of Test Accuracy on Synthetic IID

B.2. Synthetic non-IID Dataset

Figure 6 and 7 show the mean and variances of accuracies on synthetic non-IID dataset. One can observe that the `DivFL` approach achieves an improved solution with the highest mean accuracies for different choices of τ, K. The power-of-choice approach has the lowest variance of accuracies between clients. However, this comes at a cost of the overall accuracy which falls under random sampling, especially, when $\tau > 1$. The `DivFL` approach provides the best tradeoff between mean accuracy and variance of accuracies. The impact of a large m is notable in the mean accuracy for `DivFL` approach only when $\tau > 1$. In such cases, the `DivFL` approach with $m = 10$ achieves about the same mean accuracy as random sampling, although with a lower variance compared to random sampling. Further, K does not impact the final mean and variance of accuracy for `DivFL` in relation to the baselines.

B.3. FEMNIST Dataset

Figures 8 and 9 show the mean and variances of accuracies on FEMNIST dataset for different choices of K. The `DivFL` approach has a comparable final mean test accuracy (86%) and variance of accuracies 0.01 for both $K = 10$ and $K = 20$. On the other hand, the accuracies of the baselines for $K = 10$ drops significantly lower (especially for random sampling) than `DivFL`. Random sampling achieve a test accuracy of only 60% while the power-of-choice approach reaches a test accuracy of 67%. When $K = 20$, however, random sampling achieves a final test accuracy of 78.7% and the power-of-choice approach marginally outperforms `DivFL` at 89.7%. This is an important advantage for `DivFL` approach where a small value of K is able to achieve a solution that's comparable to large values of K for baselines for the same number of communication rounds. But this also leads to a significant reduction in overhead in computational resources at clients as only fewer clients need to participate in training.

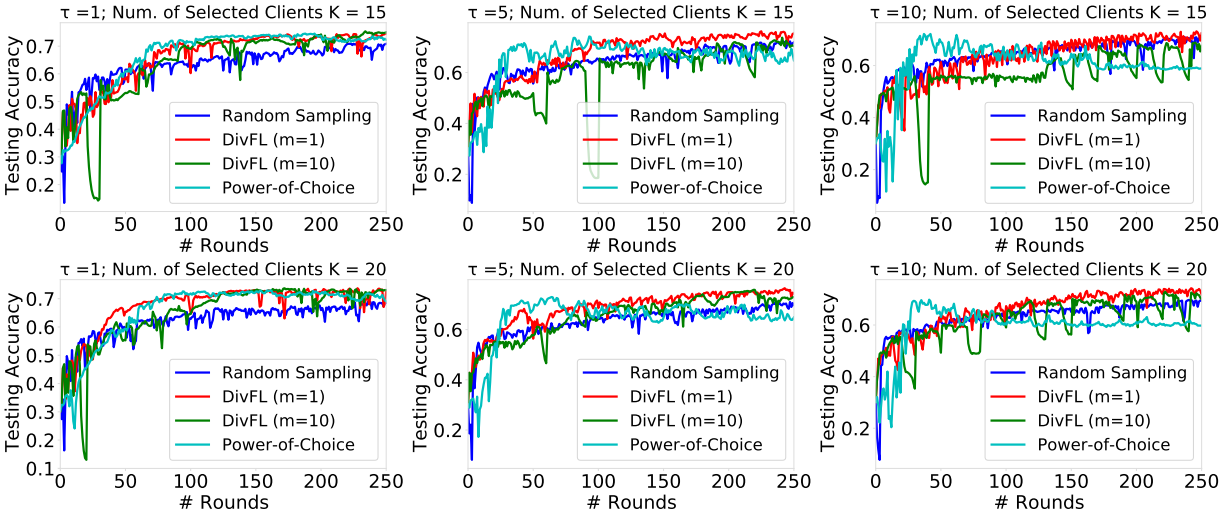

Figure 6. Average Test Accuracy on Synthetic non-IID

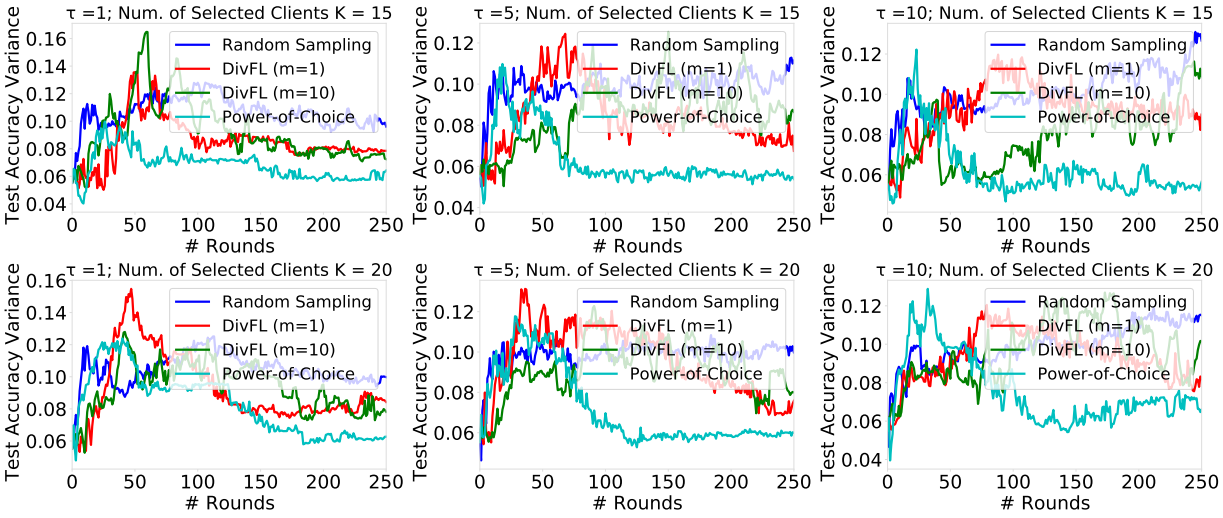

Figure 7. Variance of Test Accuracy on Synthetic non-IID

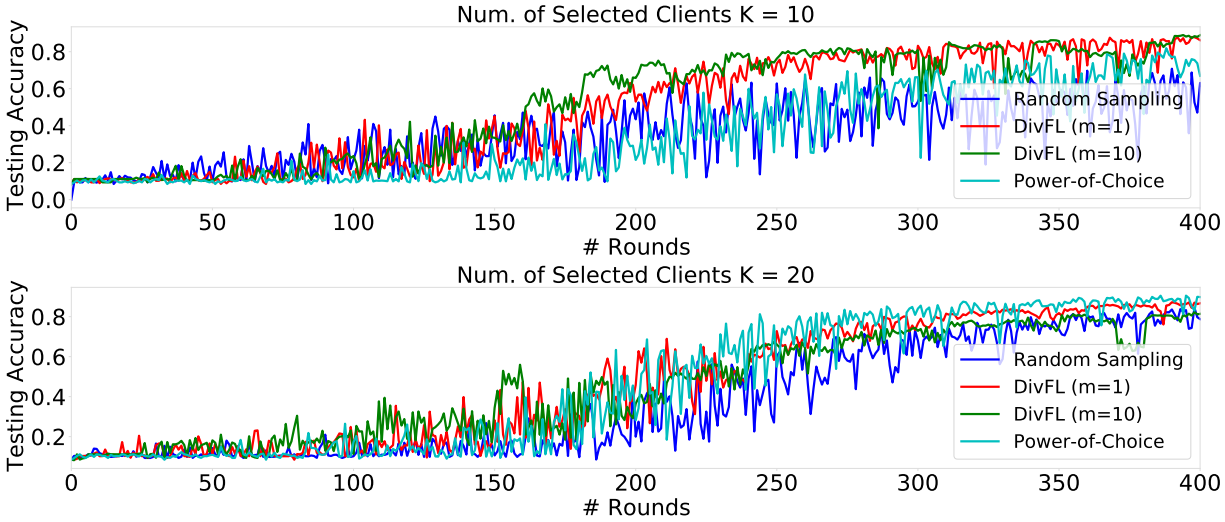

Figure 8. Average Test Accuracy on FEMNIST Dataset

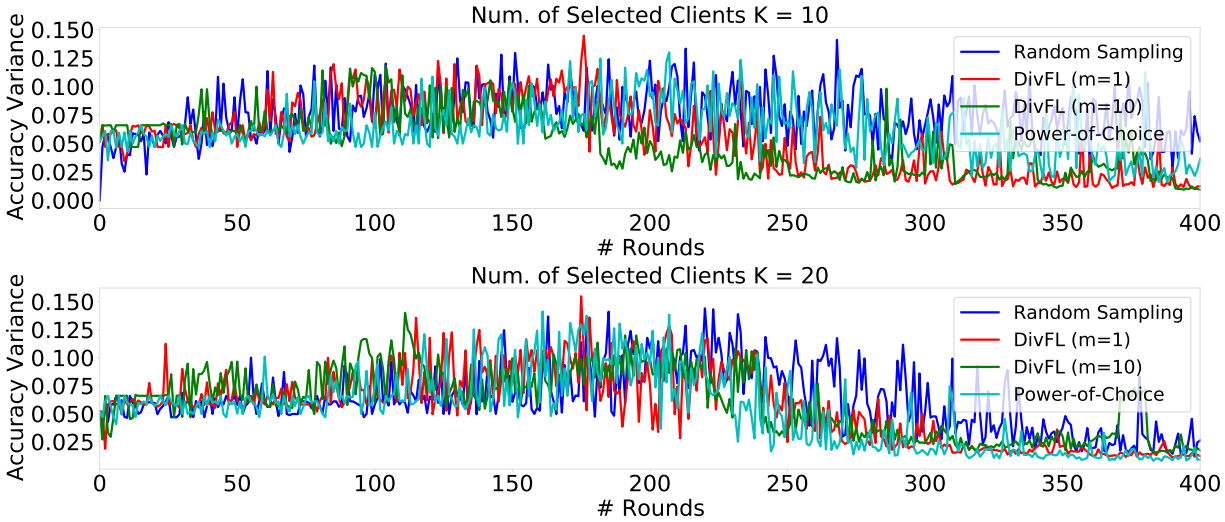

Figure 9. Variance of Test Accuracy on FEMNIST Dataset