Optimal Client Sampling for Federated Learning

Anonymous authors
Paper under double-blind review

Abstract

It is well understood that client-master communication can be a primary bottleneck in Federated Learning (FL). In this work, we address this issue with a novel client subsampling scheme, where we restrict the number of clients allowed to communicate their updates back to the master node. In each communication round, all participating clients compute their updates, but only the ones with "important" updates communicate back to the master. We show that importance can be measured using only the norm of the update and give a formula for optimal client participation. This formula minimizes the distance between the full update, where all clients participate, and our limited update, where the number of participating clients is restricted. In addition, we provide a simple algorithm that approximates the optimal formula for client participation, which allows for secure aggregation and stateless clients, and thus does not compromise client privacy. We show both theoretically and empirically that for Distributed `SGD` (`DSGD`) and Federated Averaging (`FedAvg`), the performance of our approach can be close to full participation and superior to the baseline where participating clients are sampled uniformly. Moreover, our approach is orthogonal to and compatible with existing methods for reducing communication overhead, such as local methods and communication compression methods.

1 Introduction

We consider the standard cross-device Federated Learning (FL) setting (Kairouz et al., 2019), where the objective is of the form

$$\min_{x\in\mathbb{R}^d}\left[f(x) \coloneqq \sum_{i=1}^{n} w_i f_i(x)\right], \tag{1}$$

where $x \in \mathbb{R}^d$ represents the parameters of a statistical model we aim to find, n is the total number of clients, each $f_i \colon \mathbb{R}^d \to \mathbb{R}$ is a continuously differentiable local loss function which depends on the data distribution $\mathcal{D}_i$ owned by client i via $f_i(x) = \mathrm{E}_{\xi\sim\mathcal{D}_i}\left[f(x,\xi)\right]$, and $w_i \geq 0$ are client weights such that $\sum_{i=1}^n w_i = 1$. We assume the classical FL setup in which a central master (server) orchestrates the training by securely aggregating updates from clients without seeing the raw data.

1.1 Motivation: Communication Bottleneck in Federated Learning

It is well understood that communication cost can be the primary bottleneck in FL. Indeed, wireless links and other end-user internet connections typically operate at lower rates than intra-datacenter or inter-datacenter links and can be potentially expensive and unreliable. Moreover, the capacity of the aggregating master and other FL system considerations imposes direct or indirect constrains on the number of clients allowed to participate in each communication round. These considerations have led to significant interest in reducing the communication bandwidth of FL systems.

Local Methods. One of the most popular strategies is to reduce the frequency of communication and put more emphasis on computation. This is usually achieved by asking the devices to perform multiple local steps before communicating their updates. A prototype method in this category is the Federated

Averaging (`FedAvg`) algorithm (McMahan et al., 2017), an adaption of local-update to parallel SGD, where each client runs some number of SGD steps locally before local updates are averaged to form the global update for the global model on the master. The original work was a heuristic, offering no theoretical guarantees, which motivated the community to try to understand the method and various existing and new variants theoretically (Stich, 2019; Lin et al., 2018; Karimireddy et al., 2019; Stich & Karimireddy, 2020; Khaled et al., 2020; Hanzely & Richtárik, 2020).

Communication Compression Methods. Another popular approach is to reduce the size of the object (typically gradients) communicated from clients to the master. This approach is referred to as gradient/communication *compression*. In this approach, instead of transmitting the full-dimensional gradient/update vector $g \in \mathbb{R}^d$, one transmits a compressed vector $\mathcal{C}(g)$, where $\mathcal{C} : \mathbb{R}^d \to \mathbb{R}^d$ is a (possibly random) operator chosen such that $\mathcal{C}(g)$ can be represented using fewer bits, for instance by using limited bit representation (quantization) or by enforcing sparsity (sparsification). A particularly popular class of quantization operators is based on random dithering (Goodall, 1951; Roberts, 1962); see Alistarh et al. (2016); Wen et al. (2017); Zhang et al. (2017); Ramezani-Kebrya et al. (2019). A new variant of random dithering developed in (Horváth et al., 2019) offers an exponential improvement on standard dithering. Sparse vectors can be obtained by random sparsification techniques that randomly mask the input vectors and preserve a constant number of coordinates (Wangni et al., 2018; Konečný & Richtárik, 2018; Stich et al., 2018; Mishchenko et al., 2019; Vogels et al., 2019). There is also a line of work (Horváth et al., 2019; Basu et al., 2019) which propose to combine sparsification and quantization to obtain a more aggressive combined effect.

1.2 Contributions

In this work, we propose a new approach to address the communication bandwidth issues appearing in FL. Our approach is based on the observation that in the situation where partial participation is desired and a budget on the number of participating clients is applied, *careful selection of the participating clients can lead to better communication complexity, and hence faster training.* In other words, we claim that in any given communication round, some clients will have "more informative" updates than others and thus the training procedure will benefit from capitalizing on this fact by ignoring some of the worthless updates.

In particular, we propose a principled *optimal client sampling scheme*, capable of identifying the most informative clients in any given communication round. Our scheme works by minimizing the variance of the stochastic gradient produced by the partial participation procedure, which then translates to a reduction in the number of communication rounds. To the best of our knowledge, this approach was not considered before. Our contributions can be summarized as follows:

- We propose a *novel adaptive partial participation strategy for reducing communication in FL*, which relies on a careful selection of the clients that are allowed to communicate their updates back to the master node in any given communication round.
- Our adaptive client sampling procedure is *optimal* in the sense that it minimizes the variance of the master update for any budget m on the number of participating clients.
- We obtain an approximation to our optimal sampling strategy which only requires aggregation, fulfilling the privacy requirements of FL. To our knowledge, our method is the first principled importance client sampling strategy that allows for both *secure aggregation* and *stateless clients*.
- Our proposal is orthogonal to and hence compatible with existing approaches to communication reduction such as communication compression and/or local updates (see Section 3.2).
- We provide convergence guarantees for our approach with `DSGD` and `FedAvg`, and show both theoretically and empirically that the performance of our approach is superior to uniform sampling and can be close to full participation.
- We show both theoretically and empirically that our approach allows for *larger learning rates* for `DSGD` and `FedAvg` algorithms than the baseline which performs uniform client sampling, which results in *better communication complexity* and hence *faster convergence*.

2 Smart Client Sampling for Reducing Communication

This section describes our proposed optimal client sampling strategy for reducing the communication bottleneck in Federated Learning. Before proceeding with our theory, we describe the problem setting and introduce the *arbitrary sampling* paradigm. In FL, each client i participating in round k computes an update vector $\mathbf{U}_i^k \in \mathbb{R}^d$. For simplicity and ease of exposition, we assume that all clients $i \in [n] \coloneqq \{1, 2, \ldots, n\}$ are available in each round[1]. In our framework, only a subset of clients communicates their updates to the master node in each communication round in order to reduce the number of transmitted bits.

In order to provide an analysis in this framework, we consider a general partial participation framework (Horváth & Richtárik, 2020), where we assume that the subset of participating clients is determined by an arbitrary random set-valued mapping $\mathbb{S}$ (i.e., a "sampling") with values in $2^{[n]}$. A sampling $\mathbb{S}$ is uniquely defined by assigning probabilities to all 2^n subsets of $[n]$. With each sampling $\mathbb{S}$ we associate a *probability matrix* $\mathbf{P} \in \mathbb{R}^{n\times n}$ defined by $\mathbf{P}_{ij} \coloneqq \text{Prob}(\{i,j\} \subseteq \mathbb{S})$. The *probability vector* associated with $\mathbb{S}$ is the vector composed of the diagonal entries of $\mathbf{P}$: $p = (p_1, \ldots, p_n) \in \mathbb{R}^n$, where $p_i \coloneqq \text{Prob}(i \in \mathbb{S})$. We say that $\mathbb{S}$ is *proper* if $p_i > 0$ for all i. It is easy to show that $b \coloneqq \mathrm{E}\left[|\mathbb{S}|\right] = \text{Trace}\left(\mathbf{P}\right) = \sum_{i=1}^n p_i$, and hence b can be seen as the expected number of clients participating in each communication round. Given parameters $p_1, \ldots, p_n \in [0,1]$, consider a random set $\mathbb{S} \subseteq [n]$ generated as follows: for each $i \in [n]$, we include i in $\mathbb{S}$ with probability p_i. This is called *independent sampling*, since the event $i \in \mathbb{S}$ is independent of $j \in \mathbb{S}$ for any $i \neq j$.

While our client sampling strategy can be adapted to essentially any underlying learning method, we give details here for `DSGD` as an illustrative example:

$$x^{k+1} = x^k - \eta^k \mathbf{G}^k, \quad \mathbf{G}^k \coloneqq \sum_{i\in S^k} \frac{w_i}{p_i^k} \mathbf{U}_i^k, \tag{2}$$

where $S^k \sim \mathbb{S}^k$ and $\mathbf{U}_i^k = g_i^k$ is an unbiased estimator of $\nabla f_i(x^k)$. The scaling factor $\frac{1}{p_i^k}$ is necessary in order to obtain an unbiased estimator of the true update, i.e., $\mathrm{E}_{S^k}\left[\mathbf{G}^k\right] = \sum_{i=1}^n w_i \mathbf{U}_i^k$.

2.1 Optimal Client Sampling

A simple observation is that the variance of our gradient estimator $\mathbf{G}^k$ can be decomposed into

$$\mathrm{E}\left[\left\|\mathbf{G}^k - \nabla f(x^k)\right\|^2\right] = \mathrm{E}\left[\left\|\mathbf{G}^k - \sum_{i=1}^n w_i \mathbf{U}_i^k\right\|^2\right] + \mathrm{E}\left[\left\|\sum_{i=1}^n w_i \mathbf{U}_i^k - \nabla f(x^k)\right\|^2\right], \tag{3}$$

where the second term on the right-hand side is independent of the sampling procedure, and the first term is zero if every client sends its update (i.e., if $p_i^k = 1$ for all i). In order to provide meaningful results, we restrict the expected number of clients to communicate in each round by bounding $b^k \coloneqq \sum_{i=1}^n p_i^k$ by some positive integer $m \leq n$. This raises the following question: *What is the sampling procedure that minimizes* (3) *for any given* m*?* We answer this question using the following technical lemma (see Appendix A for a proof):

Lemma 2.1. *Let $\zeta_1, \zeta_2, \ldots, \zeta_n$ be vectors in $\mathbb{R}^d$ and $w_1, w_2, \ldots, w_n$ be non-negative real numbers such that $\sum_{i=1}^n w_i = 1$. Define $\tilde{\zeta} \coloneqq \sum_{i=1}^n w_i \zeta_i$. Let S be a proper sampling. If $v \in \mathbb{R}^n$ is such that*

$$\mathbf{P} - pp^\top \preceq \mathbf{Diag}(p_1 v_1, p_2 v_2, \ldots, p_n v_n), \tag{4}$$

then

$$\mathrm{E}\left[\left\|\sum_{i\in S} \frac{w_i \zeta_i}{p_i} - \tilde{\zeta}\right\|^2\right] \leq \sum_{i=1}^n w_i^2 \frac{v_i}{p_i} \left\|\zeta_i\right\|^2, \tag{5}$$

where the expectation is taken over S. Whenever (4) *holds, it must be the case that $v_i \geq 1 - p_i$.*

[1]This is not a limiting factor, as all presented theory can be easily extended to the case of partial participation with an arbitrary proper sampling distribution.

Algorithm 1 Optimal Client Sampling (`OCS`).

1: **Input:** expected batch size m
2: each client i computes a local update $\mathbf{U}_i^k$ (in parallel)
3: each client i sends the norm of its update $u_i^k = w_i \left\|\mathbf{U}_i^k\right\|$ to the master (in parallel)
4: master computes optimal probabilities p_i^k using equation (7)
5: master broadcasts p_i^k to all clients
6: each client i sends its update $\frac{w_i}{p_i^k}\mathbf{U}_i^k$ to the master with probability p_i^k (in parallel)

It turns out that given probabilities $\{p_i\}$, among all samplings S satisfying $p_i = \text{Prob}(i \in S)$, the independent sampling (i.e., $p_{ij} = \text{Prob}(i, j \in S) = \text{Prob}(i \in S)\,\text{Prob}(j \in S) = p_i p_j$) minimizes the left-hand side of (5). This is due to two nice properties: a) any independent sampling admits optimal choice of v, i.e., $v_i = 1 - p_i$ for all i, and b) for independent sampling (5) holds as equality. In the context of our method, these properties can be written as

$$\mathrm{E}\left[\left\|\mathbf{G}^k - \sum_{i=1}^n w_i \mathbf{U}_i^k\right\|^2\right] = \mathrm{E}\left[\sum_{i=1}^n w_i^2 \frac{1-p_i^k}{p_i^k}\left\|\mathbf{U}_i^k\right\|^2\right]. \tag{6}$$

It now only remains to find the parameters $\{p_i^k\}$ defining the optimal independent sampling, i.e., one that minimizes (6) subject to the constraints $0 \leq p_i^k \leq 1$ and $b^k \coloneqq \sum_{i=1}^n p_i^k \leq m$. It turns out that this problem has the following closed-form solution (see Appendix B):

$$p_i^k = \begin{cases} (m+l-n)\dfrac{\left\|\tilde{U}_i^k\right\|}{\sum_{j=1}^l \left\|\tilde{U}_{(j)}^k\right\|}, & \text{if } i \notin A^k \\ 1, & \text{if } i \in A^k \end{cases}, \tag{7}$$

where $\tilde{U}_i^k \coloneqq w_i \mathbf{U}_i^k$, and $\left\|\tilde{U}_{(j)}^k\right\|$ is the j-th largest value in $\left\{\left\|\tilde{U}_i^k\right\|\right\}_{i=1}^n$, l is the largest integer for which $0 < m + l - n \leq \dfrac{\sum_{i=1}^l \left\|\tilde{U}_{(i)}^k\right\|}{\left\|\tilde{U}_{(l)}^k\right\|}$ (note that this inequality at least holds for $l = n - m + 1$), and A^k contains indices i such that $\left\|\tilde{U}_i^k\right\| \geq \left\|\tilde{U}_{(l+1)}^k\right\|$. We summarize this procedure in Algorithm 1.

Remark. Optimizing the left-hand side of (5) does not guarantee the proposed sampling to be optimal with respect to the right-hand side in the general case. For this to hold, our sampling needs to be independent, which is not a very restrictive condition, especially considering that enforcing independent sampling across clients accommodates the privacy requirements of FL. In addition, since (5) is tight, our sampling is optimal if one is allowed to communicate only norms (i.e., one float per client) as extra information. We stress that requiring optimality with respect to the right-hand side of (5) in the full general case is not practical, as it cannot be obtained without revealing, i.e., communicating, all clients' full updates to the master.

2.2 Privacy-preserving Algorithm

In the case $l = n$, the optimal probabilities $p_i^k = \dfrac{m\left\|\tilde{U}_i^k\right\|}{\sum_{j=1}^n \left\|\tilde{U}_j^k\right\|}$ can be computed easily: the master aggregates the norm of each update and then sends the sum back to the clients. However, if $l < n$, in order to compute optimal probabilities, the master would need to identify the norm of every update and perform partial sorting, which can be computationally expensive and also violates the client privacy requirements in FL. Therefore, we develop an algorithm for approximately solving this problem, which only requires to perform aggregation at the master node without compromising the privacy of any client. The construction of this algorithm is motivated by Wangni et al. (2018). We first set $\tilde{p}_i^k = \dfrac{m\left\|\tilde{U}_i^k\right\|}{\sum_{j=1}^n \left\|\tilde{U}_j^k\right\|}$ and $p_i^k = \min\{\tilde{p}_i^k, 1\}$. In the ideal situation where every $\tilde{p}_i^k$ equals the optimal solution (7), this would be sufficient. However, due to the truncation

operation, the expected mini-batch size $b^k = \sum_{i=1}^n p_i^k \leq \sum_{i=1}^n \frac{m\|\tilde{U}_i^k\|}{\sum_{j=1}^n \|\tilde{U}_j^k\|} = m$ can be strictly less than m if $\tilde{p}_i^k > 1$ holds true for at least one i. Hence, we employ an iterative procedure to fix this gap by rescaling the probabilities which are smaller than 1, as summarized in Algorithm 2. This algorithm is much easier to implement and computationally more efficient on parallel computing architectures. In addition, it only requires a secure aggregation procedure on the master, which is essential in privacy preserving FL, and thus it is compatible with existing FL software and hardware.

Extra Communication Costs. We acknowledge that Algorithm 2 brings extra communication costs, as it requires all clients to send the norms of their updates u_i^k's and probabilities p_i^k's in each round. However, since these are single floats, this only costs $\mathcal{O}(j_{\max})$ extra floats for each client. Picking $j_{\max} = \mathcal{O}(1)$, this is negligible for large models of size d.

Fairness. Based on our sampling strategy, it might be tempting to assume that the obtained solution could exhibit fairness issues. In our convergence analyses below, we show that this is not the case, as our proposed methods converge to the optimal solution. Hence, as long as the original objective has no inherent issue with fairness, our methods do not exhibit any fairness issues. Besides, our algorithm can be used in conjunction with other "more fair" objectives, e.g., Tilted ERM (Li et al., 2021), if needed.

3 Convergence Guarantees

This section provides convergence analyses for `DSGD` and `FedAvg` with our optimal client sampling scheme in both convex and non-convex settings. We compare the convergence results of our scheme with those of full participation and independent uniform sampling with sample size m. We match the forms of our convergence bounds to those of the existing bounds in the literature to make them directly comparable. We do not compare the sample complexities of these methods, as such comparisons would be difficult due to their dependence on the actual updates which are unknown in advance and do not follow a specific distribution in general. Intuitively, our method can be thought of as uniform sampling with $\tilde{m} \in [m, n]$ effective sampled clients, while only m clients are actually sampled in expectation, which indicates that it cannot be worse than uniform sampling and can be as good as full participation. The actual value of $\tilde{m}$ depends on the updates.

We use standard assumptions (Karimi et al., 2016), assuming throughout that f has a unique minimizer $x^\star$ with $f^\star = f(x^\star) > -\infty$ and f_i's are L-smooth, i.e., f_i's have L-Lipschitz continuous gradients. We first define convex functions and L-smooth functions.

Definition 3.1 (Convexity)**.** $f : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex with $\mu > 0$ if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^d. \tag{8}$$

$f : \mathbb{R}^d \to \mathbb{R}$ is convex if it satisfies (8) with $\mu = 0$.

Definition 3.2 (Smoothness)**.** $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \quad \forall x, y \in \mathbb{R}^d. \tag{9}$$

We now state standard assumptions of the gradient oracles for `DSGD` and `FedAvg`.

Assumption 3.3 (Gradient oracle for `DSGD`)**.** The stochastic gradient estimator $g_i^k = \nabla f_i(x^k) + \xi_i^k$ of the local gradient $\nabla f_i(x^k)$, for each round k and all $i = 1, \ldots, n$, satisfies

$$\mathrm{E}\left[\xi_i^k\right] = 0 \quad \text{and} \quad \mathrm{E}\left[\left\|\xi_i^k\right\|^2 | x_i^k\right] \leq M \left\|\nabla f_i(x^k)\right\|^2 + \sigma^2, \text{ for some } M \geq 0. \tag{10}$$

This further implies that $\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n g_i^k \mid x^k\right] = \nabla f(x^k)$.

Assumption 3.4 (Gradient oracle for `FedAvg`)**.** The stochastic gradient estimator $g_i(y_{i,r}^k) = \nabla f_i(y_{i,r}^k) + \xi_{i,r}^k$ of the local gradient $\nabla f_i(y_{i,r}^k)$, for each round k, each local step $r = 0, \ldots, R$ and all $i = 1, \ldots, n$, satisfies

$$\mathrm{E}\left[\xi_{i,r}^k\right] = 0 \quad \text{and} \quad \mathrm{E}\left[\left\|\xi_{i,r}^k\right\|^2 | y_{i,r}^k\right] \leq M \left\|\nabla f_i(y_{i,r}^k)\right\|^2 + \sigma^2, \text{ for some } M \geq 0, \tag{11}$$

where $y_{i,0}^k = x^k$ and $y_{i,r}^k = y_{i,r-1}^k - \eta_l g_i(y_{i,r}^k)$, for $r = 1, \cdots, R$.

Algorithm 2 Approximate Optimal Client Sampling (`AOCS`).

Input: expected batch size m, maximum number of iteration $j_{\max}$
each client i computes an update $\mathbf{U}_i^k$ (in parallel)
each client i sends the norm of its update $u_i^k = w_i \left\|\mathbf{U}_i^k\right\|$ to the master (in parallel)
master aggregates $u^k = \sum_{i=1}^n u_i^k$
master broadcasts u^k to all clients
each client i computes $p_i^k = \min\{\frac{mu_i^k}{u^k}, 1\}$ (in parallel)
for $j = 1, \cdots, j_{max}$ **do**
  each client i sends $t_i^k = (1, p_i^k)$ to the master if $p_i^k < 1$; else sends $t_i^k = (0, 0)$ (in parallel)
  master aggregates $(I^k, P^k) = \sum_{i=1}^n t_i^k$
  master computes $C^k = \frac{m-n+I^k}{P^k}$
  master broadcasts C^k to all clients
  each client i recalibrates $p_i^k = \min\{C^k p_i^k, 1\}$ if $p_i^k < 1$ (in parallel)
  if $C^k \leq 1$ **then**
    break
  end if
end for
each clients i sends its update $\frac{w_i}{p_i^k}\mathbf{U}_i^k$ to master with probability p_i^k (in parallel)

For non-convex objectives, one can construct counter-examples that would diverge for both `DSGD` and `FedAvg` if the sampling variance is not bounded. Therefore, we need to employ the following standard assumption of local gradients for bounding the sampling variance[2].

Assumption 3.5 (Similarity among local gradients)**.** The gradients of local loss functions f_i satisfy

$$\sum_{i=1}^n w_i \left\|\nabla f_i(x) - \nabla f(x)\right\|^2 \leq \rho, \quad \text{for some } \rho \geq 0. \tag{12}$$

Some works employ a more restrictive assumption which requires $\|\nabla f_i(x) - \nabla f(x)\| \leq \rho,\ \forall i$, from which Assumption 3.5 can be derived, since $\sum_{i=1}^n w_i = 1$. Therefore, Assumption 3.5 can be seen as an assumption on similarity among local gradients. Furthermore, this assumption does not require w_i's to be lower-bounded, as clients with $w_i = 0$ will never be sampled and thus can be removed from the objective.

We also define three quantities, which will appear in our convergence analyses:

$$W \coloneqq \max_{i \in [n]}\{w_i\}, \quad R_i \coloneqq f_i(x^\star) - f_i^\star, \quad r^k \coloneqq x^k - x^\star, \tag{13}$$

where $f_i^\star$ is the functional value of f_i at its optimum, R_i represents the mismatch between the local and global minimizer, and r^k captures the distance between the current point and the minimizer of f.

We are now ready to proceed with our convergence analyses. We define the improvement factor

$$\alpha^k \coloneqq \frac{\mathrm{E}\left[\left\|\sum_{i \in S^k} \frac{w_i}{p_i^k}\mathbf{U}_i^k - \sum_{i=1}^n w_i \mathbf{U}_i^k\right\|^2\right]}{\mathrm{E}\left[\left\|\sum_{i \in U^k} \frac{w_i}{p_i^U}\mathbf{U}_i^k - \sum_{i=1}^n w_i \mathbf{U}_i^k\right\|^2\right]}, \tag{14}$$

where $S^k \sim \mathbb{S}^k$ with p_i^k defined in (7) and $U^k \sim \mathbb{U}$ is an independent uniform sampling with $p_i^U = m/n$. By construction, $\alpha^k \leq 1$, as $\mathbb{S}^k$ minimizes the variance term (see Appendix B). Note that α^k can reach zero

[2]This assumption is not required for convex objectives, as one can show that the sampling variance is bounded using smoothness and convexity.

in the case where there are at most m non-zero updates. If $\alpha^k = 0$, our method performs as if all updates were communicated. In the worst-case $\alpha^k = 1$, our method performs as if we picked m updates uniformly at random, and one could not do better in theory due to the structure of the updates $\mathbf{U}_i^k$. In the following subsections, we provide convergence analyses of specific methods for solving the optimization problem (1). The proofs of the theorems are deferred to Appendices C and D. For simplicity of notation, we denote

$$\gamma^k \coloneqq \frac{m}{\alpha^k(n-m)+m} \in \left[\frac{m}{n}, 1\right], \quad k = 0, \ldots, K-1. \tag{15}$$

3.1 Distributed SGD (DSGD) with Optimal Client Sampling

We obtain analyses for DSGD (2) with optimal client sampling in both convex and non-convex settings.

Theorem 3.6. *Let f_i be L-smooth and convex for $i = 1, \ldots, n$. Let f be μ-strongly convex. Suppose that Assumption 3.3 holds. Choose $\eta^k \in \left(0, \frac{\gamma^k}{(1+WM)L}\right]$. Define*

$$\beta_1 \coloneqq \sum_{i=1}^n w_i^2(2L(1+M)R_i + \sigma^2 \quad \text{and} \quad \beta_2 \coloneqq 2L\sum_{i=1}^n w_i^2 R_i. \tag{16}$$

The iterates of DSGD *with optimal client sampling* (7) *satisfy*

$$\mathrm{E}\left[\left\|r^{k+1}\right\|^2\right] \leq (1-\mu\eta^k)\mathrm{E}\left[\left\|r^k\right\|^2\right] + (\eta^k)^2\left(\frac{\beta_1}{\gamma^k} - \beta_2\right). \tag{17}$$

Interpretation. We first look at the best and worst case scenarios. In the best case scenario, we have $\gamma^k = 1$ for all k's. This implies that there is no loss of speed comparing to the method with full participation. It is indeed confirmed by our theory as our obtained recursion recovers the best-known rate of DSGD in the full participation regime (Gower et al., 2019). Similarly, in the worst case, we have $\gamma^k = m/n$ for all k's, which corresponds to uniform sampling with sample size m, and our recursion recovers the best-known rate for DSGD in this regime. This is expected as (14) implies that every update $\mathbf{U}_i^k$ is equivalent, and thus it is theoretically impossible to obtain a better rate than that of uniform sampling in the worst case scenario. In the general scenario, our obtained recursion sits somewhere between full and uniform partial participation, where the actual position is determined by γ^k's which capture the distribution of updates (here gradients) on the clients. For instance, with a larger number of γ^k's tending to 1, we are closer to the full participation regime. Similarly, with more γ^k's tending to m/n, we are closer to the rate of uniform partial participation.

Theorem 3.7. *Let f_i be L-smooth for $i = 1, \ldots, n$. Suppose that Assumptions 3.3 and 3.5 hold. Let η^k be the step size and define*

$$\beta^k \coloneqq \frac{L}{2\gamma^k}\left((1+M-\gamma^k)W\rho + \sum_{i=1}^n w_i^2\sigma^2\right). \tag{18}$$

The iterates of DSGD *with optimal client sampling* (7) *satisfy*

$$\mathrm{E}\left[f(x^{k+1})\right] \leq \mathrm{E}\left[f(x^k)\right] - \eta^k\left(1 - \frac{(1+M)L}{2\gamma^k}\eta^k\right)\mathrm{E}\left[\left\|\nabla f(x^k)\right\|^2\right] + (\eta^k)^2\beta^k. \tag{19}$$

Interpretation. The iterate (19) recovers the standard form of the convergence result of DSGD for one recursion step in the non-convex setting. Similar to the previous results, this convergence bound sits between the best-known rate of full participation and uniform sampling (Bottou et al., 2018).

3.2 Federated Averaging (FedAvg) with Optimal Client Sampling

Pseudo-code that adapts the standard FedAvg algorithm to our framework is provided in Algorithm 3. We obtain analyses for FedAvg with optimal client sampling in both convex and non-convex settings.

Algorithm 3 `FedAvg` with Optimal Client Sampling.

Input: initial global model x^1, global and local step-sizes η_g^k, η_l^k
for each round $k = 1, \dots, K$ **do**
  master broadcasts x^k to all clients $i \in [n]$
  for each client $i \in [n]$ (in parallel) **do**
    initialize local model $y_{i,0}^k \leftarrow x^k$
    for $r = 1, \dots, R$ **do**
      compute mini-batch gradient $g_i(y_{i,r-1}^k)$
      update $y_{i,r}^k \leftarrow y_{i,r-1}^k - \eta_l^k g_i(y_{i,r-1}^k)$
    end for
    compute $\mathbf{U}_i^k \coloneqq \Delta y_i^k = x^k - y_{i,R}^k$
    compute p_i^k using Algorithm 1 or 2
    send $\frac{w_i}{p_i^k}\Delta y_i^k$ to master with probability p_i^k
  end for
  master computes $\Delta x^k = \sum_{i \in S^k} \frac{w_i}{p_i^k}\Delta y_i^k$
  master updates global model $x^{k+1} \leftarrow x^k - \eta_g^k \Delta x^k$
end for

Theorem 3.8. *Let f_i be L-smooth and μ-strongly convex for $i = 1, \dots, n$. Suppose that Assumption 3.4 holds. Let $\eta^k \coloneqq R\eta_l^k\eta_g^k$ be the effective step-size and $\eta_g^k \geq \sqrt{\frac{\gamma^k}{\sum_i w_i^2}}$. Choose $\eta^k \in \left(0, \frac{1}{8}\min\left\{\frac{1}{L(2+M/R)}, \frac{\gamma^k}{(1+W(1+M/R))L}\right\}\right]$,*

$$\beta_1^k \coloneqq \frac{2\sigma^2}{\gamma^k R}\sum_{i=1}^n w_i^2 + 4L\left(\frac{M}{R} + 1 - \gamma^k\right)\sum_{i=1}^n w_i^2 R_i \quad \text{and} \quad \beta_2 \coloneqq 72L^2\left(1 + \frac{M}{R}\right)\sum_{i=1}^n w_i R_i. \tag{20}$$

The iterates of `FedAvg` *($R \geq 2$) with optimal client sampling* (7) *satisfy*

$$\frac{3}{8}\mathrm{E}\left[(f(x^k) - f^\star)\right] \leq \frac{1}{\eta^k}\left(1 - \frac{\mu\eta^k}{2}\right)\mathrm{E}\left[\left\|r^k\right\|^2\right] - \frac{1}{\eta^k}\mathrm{E}\left[\left\|r^{k+1}\right\|^2\right] + \eta^k\beta_1^k + (\eta^k)^2\beta_2. \tag{21}$$

Theorem 3.9. *Let f_i be L-smooth for all $i = 1, \dots, n$. Suppose that Assumptions 3.4 and 3.5 hold. Let $\eta^k \coloneqq R\eta_l^k\eta_g^k$ be the effective step-size and $\eta_g^k \geq \sqrt{\frac{5\gamma^k}{4\sum_i w_i^2}}$. Choose $\eta^k \in \left(0, \frac{1}{8L(2+M/R)}\right]$. The iterates of* `FedAvg` *($R \geq 2$) with optimal client sampling* (7) *satisfy*

$$\mathrm{E}\left[f(x^{k+1})\right] \leq \mathrm{E}\left[f(x^k)\right] - \frac{3\eta^k}{8}\left(1 - \frac{10\eta^k L}{3}\right)\mathrm{E}\left[\left\|\nabla f(x^k)\right\|^2\right] + \frac{\eta^k\rho}{8} + (\eta^k)^2\left(\frac{\rho}{4} + \frac{\sigma^2}{R\gamma^k}\sum_{i=1}^n w_i^2\right)L. \tag{22}$$

Interpretation. The convergence guarantees from Theorems 3.8 and 3.9 sit somewhere between those for full and uniform partial participation. The actual position is again determined by the distribution of the updates which are linked to γ^k's. In the edge cases, i.e., $\gamma^k = 1$ (best case) or $\gamma^k = m/n$ (worst case), we recover the state-of-the-art complexity guarantees provided in Karimireddy et al. (2019) in both regimes. Note that our results are slightly more general, as Karimireddy et al. (2019) assumes $M = 0$ and $w_i = 1/n$.

4 Related Work

4.1 Importance Client Sampling in Federated Learning

Several recent works have studied efficient importance client sampling methods in FL. Unfortunately, none of these methods is principled, as they rely on heuristics, historical losses, or partial information. Furthermore,

they violate at least one of the core privacy requirements of FL (secure aggregation and stateless clients). For example, Cho et al. (2020) biases client selection towards clients with higher local loss, Lai et al. (2021) guides client selection based on the system and statistical utility of clients, and Ribero & Vikalo (2020) models the progression of the model's weights by an Ornstein-Uhlenbeck process based on partial information.

In contrast, our proposed method is the first *principled optimal client sampling strategy* in the sense that it minimizes the variance of the master update and is *compatible with the core privacy requirements of FL.*

4.2 Importance Sampling in Stochastic Optimization

Importance sampling methods for optimization have been studied extensively in the last few years in several contexts, including convex optimization and deep learning. `LASVM` developed in Bordes et al. (2005) is an online algorithm that uses importance sampling to train kernelized support vector machines. The first importance sampling for randomized coordinate descent methods was proposed in a seminal paper (Nesterov, 2012). It was showed in Richtárik & Takáč (2014) that the proposed sampling is optimal. Later, several extensions and improvements followed (Shalev-Shwartz & Zhang, 2014; Lin et al., 2014; Fercoq & Richtárik, 2015; Qu et al., 2015; Allen-Zhu et al., 2016; Stich et al., 2017). Another branch of work studies sample complexity. In Needell et al. (2014); Zhao & Zhang (2015), the authors make a connection with the variance of the gradient estimates of `SGD` and show that the optimal sampling distribution is proportional to the per-sample gradient norm. However, obtaining this distribution is as expensive as computing the full gradient in terms of computation, and thus it is not practical. For simpler problems, one can sample proportionally to the norms of the inputs, which can be linked to the Lipschitz constants of the per-sample loss function for linear and logistic regression. For instance, it was shown in Horváth & Richtárik (2019) that static optimal sampling can be constructed even for mini-batches and the probability is proportional to these Lipschitz constants under the assumption that these constants of the per-sample loss function are known. Unfortunately, importance measures such as smoothness of the gradient are often hard to compute/estimate for more complicated models such as those arising in deep learning, where most of the importance sampling schemes are based on heuristics. For instance, a manually designed sampling scheme was proposed in Bengio et al. (2009). It was inspired by the perceived way that human children learn; in practice, they provide the network with examples of increasing difficulty in an arbitrary manner. In a diametrically opposite approach, it is common for deep embedding learning to sample hard examples because of the plethora of easy non-informative ones (Schroff et al., 2015; Simo-Serra et al., 2015). Other approaches use a history of losses for previously seen samples to create the sampling distribution and sample either proportionally to the loss or based on the loss ranking (Schaul et al., 2015; Loshchilov & Hutter, 2015). Katharopoulos & Fleuret (2018) proposes to sample based on the gradient norm of a small uniformly sampled subset of samples.

Although our proposed optimal sampling method adapts and extends the importance sampling results from Horváth & Richtárik (2019) to the distributed setting of FL, it does not suffer from any of the limitations discussed above, since the motivation of our work is to *reduce communication* rather than computation. In particular, our method allows for any budge $m < n$ on the number of participating clients, which generalizes the theoretical results from Zhao & Zhang (2015) which only applies to the case $m = 1$.

5 Experiments

5.1 Setup

We empirically evaluate our optimal client sampling method on standard federated datasets from LEAF (Caldas et al., 2018). We compare our method with 1) the baseline where participating clients are sampled uniformly from available clients in each round and 2) full participation where all available clients participate. We chose not to compare with other client sampling methods, as such comparisons would be unfair. This is because they violate the privacy requirements of FL: our method is the only importance client sampling strategy that is deployable to real-world FL systems (see Section 4.1). We simulate the cross-device FL distributed setting and train our models using TensorFlow Federated (TFF). We conclude our evaluations

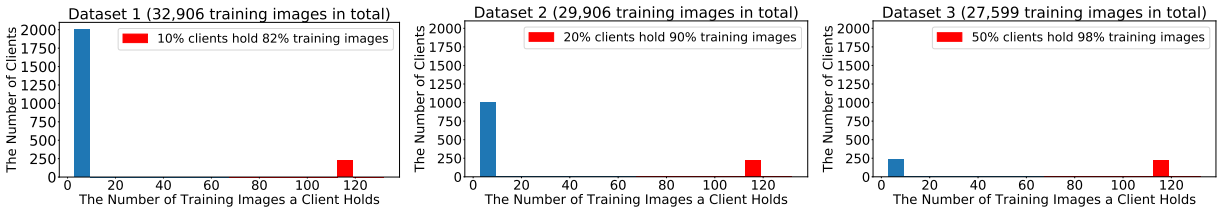

Figure 1: Distributions of the three modified Federated EMNIST training sets.

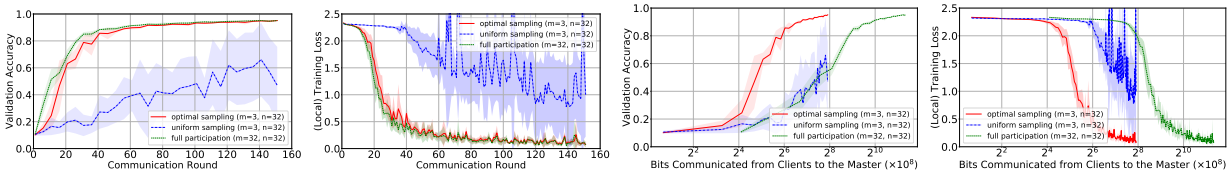

Figure 2: (FEMNIST Dataset 1) Validation accuracy and (local) training loss as a function of the number of communication rounds and the number of bits communicated from clients to the master.

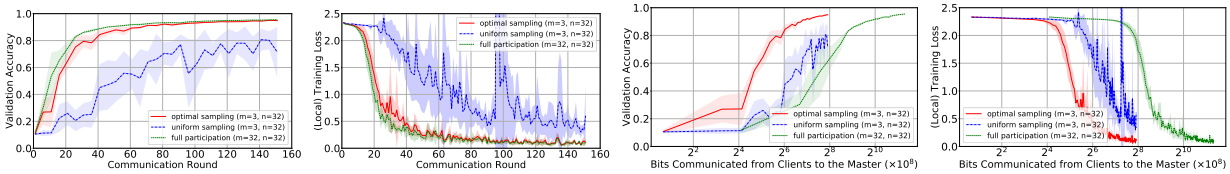

Figure 3: (FEMNIST Dataset 2) Validation accuracy and (local) training loss as a function of the number of communication rounds and the number of bits communicated from clients to the master.

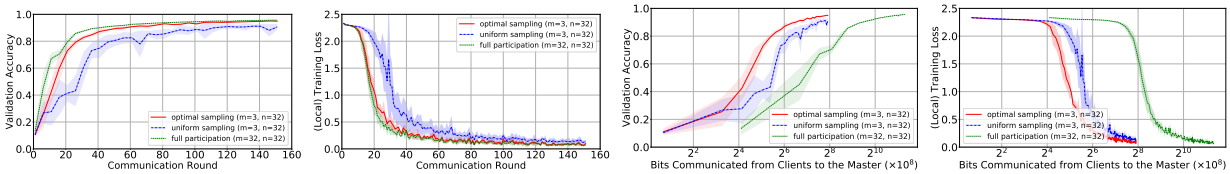

Figure 4: (FEMNIST Dataset 3) Validation accuracy and (local) training loss as a function of the number of communication rounds and the number of bits communicated from clients to the master.

using `FedAvg` with Algorithm 2[3], as it supports stateless clients and secure aggregation. We extend the TFF implementation of `FedAvg` to fit our framework. For all three methods, we report validation accuracy and (local) training loss as a function of the number of communication rounds and the number of bits communicated from clients to the master[4] . Each figure displays the mean performance with standard error over 5 independent runs. For a fair comparison, we use the same random seed for all three methods in a single run and vary random seeds across different runs. Detailed experimental settings and extra results can be found in Appendices E.1 and E.2. Our code together with datasets is included in the supplementary material.

[3]We compared the results of Algorithms 1 and 2 for all experiments as a subroutine. Their results are identical, so we only show results for Algorithm 2 and argue that the performance loss caused by its approximation is negligible.

[4]The communication from the master to clients is not considered as a bottleneck and thus not included in the results. This is a standard consideration for distributed systems, as one-to-many communication primitives (i.e., from the master to clients) are several orders of magnitude faster than many-to-one communication primitives (i.e., from clients to the master). This gap is further exacerbated in FL due to the large number of clients and slow client connections.

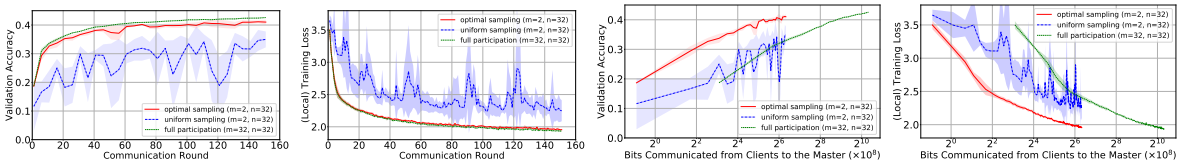

Figure 5: (Shakespeare Dataset) Validation accuracy and (local) training loss as a function of the number of communication rounds and the number of bits communicated from clients to the master.

5.1.1 Federated EMNIST Dataset

We first evaluate our method on the Federated EMNIST (FEMNIST) image dataset for image classification. Since it is a well-balanced dataset with data of similar quality on each client, we modify its training set by removing some images from some clients, in order to better simulate the conditions in which our proposed method brings significant theoretical improvements. As a result, we produce three unbalanced training sets as summarized in Figure 1[5]. We use the same CNN model as the one used in McMahan et al. (2017). For validation, we use the unchanged EMNIST validation set, which consists of $40,832$ images. In each communication round, $n = 32$ clients are sampled uniformly from the client pool, each of which then performs several SGD steps on its local training images for 1 epoch with batch size 20. For partial participation, the expected number of clients allowed to communicate their updates back to the master is set to $m = 3$. We use vanilla SGD optimizers with constant step sizes for both clients and the master, with $\eta_g = 1$ and η_l tuned on a holdout set. For full participation and optimal sampling, it turns out that $\eta_l = 2^{-3}$ is the optimal local step size for all three datasets. For uniform sampling, the optimal is $\eta_l = 2^{-5}$ for Dataset 1 and $\eta_l = 2^{-4}$ for Datasets 2 and 3. We set $j_{\max} = 4$ and include the extra communication costs in our results. The main results are shown in Figures 2, 3 and 4.

5.1.2 Shakespeare Dataset

We also evaluate our method on the (unchanged) Shakespeare text dataset for next character prediction. The vocabulary set for this task consists of 86 unique characters. The dataset contains 715 clients, each corresponding to a character in Shakespeare's plays. We divide the text into batches such that each batch contains 8 example sequences of length 5. We use a two-layer GRU model. We set $n = 32$, $m = 2$, $j_{max} = 4$ and run several SGD steps for 1 epoch on each client's local dataset in every communication round. We use vanilla SGD optimizers with constant step sizes, with $\eta_g = 1$ and η_l tuned on a holdout set. For full participation and optimal sampling, it turns out that the optimal is $\eta_l = 2^{-2}$. For uniform sampling, the optimal is $\eta_l = 2^{-3}$. The main result is shown in Figure 5.

5.2 Discussions

As predicted by our theory, the performance of `FedAvg` with our proposed optimal client sampling strategy is in between that with full and uniform partial participation. For all datasets, the optimal sampling strategy performs slightly worse than but is still competitive with the full participation strategy in terms of the number of communication rounds: it almost reached the performance of full participation while only less than 10% of the available clients communicate their updates back to the master. Note that the uniform sampling strategy performs significantly worse, which indicates that a careful choice of sampling probabilities can go a long way towards closing the gap between the performance of naive uniform sampling and full participation.

More importantly, and this was the main motivation of our work, our optimal sampling strategy is significantly better than both the uniform sampling and full participation strategies when we compare validation accuracy as a function of the number of bits communicated from clients to the master. For instance, on FEMNIST Dataset 1 (Figure 2), while our optimal sampling approach reached around 85% validation accuracy after $2^6 \times 10^8$ communicated bits, neither the full nor the uniform sampling strategies are able to exceed 40%

[5]The aim of creating various unbalanced datasets is to show that optimal sampling has more performance gains over uniform sampling on more unbalanced datasets, since α^k's (defined in Equation (14)) are more likely to be close to zero in this case.

validation accuracy within the same communication budget. Indeed, to reach the same 85% validation accuracy, full participation approach needs to communicate more than $2^9 \times 10^8$ bits, i.e., $8\times$ more, and uniform sampling approach needs to communicate about the same number of bits as full participation or even more. The results for FEMNIST Datasets 2 and 3 and for the Shakespeare dataset are of a similar qualitative nature, showing that these conclusions are robust across the datasets considered.

It is also worth noting that the empirical results from Sections 5.1.1 and 5.1.2 confirm that our optimal sampling strategy allows for larger step sizes than uniform sampling, as the hyperparameter search returns larger step sizes η_l for optimal sampling than for uniform sampling.

6 Conclusion and Future Work

In this work, we have proposed a principled optimal client sampling clients strategies to address the communication bottleneck issue of Federated Learning. Our optimal client sampling can be computed by a closed-form formula using only the norms of the updates. Furthermore, our proposed method is the first principled importance client sampling strategy that is compatible with stateless clients and secure aggregation. We have obtained convergence guarantees for our method with `DSGD` and `FedAvg`, and have performed empirical evaluations of our method on the standard federated datasets from the LEAF database. The empirical results show that our method is superior to uniform sampling and close to full participation, which corroborates our theoretical analysis.

Some of the directions for future work are as follows: 1) extending our optimal client sampling strategy to take into account the constraints of local clients (e.g., computational speed and network bandwidth); and 2) combining our approach with communication compression methods to further reduce the sizes of communicated updates.

References

Dan Alistarh, Jerry Li, Ryota Tomioka, and Milan Vojnovic. QSGD: Randomized quantization for communication-optimal stochastic gradient descent. *arXiv preprint arXiv:1610.02132*, 2016.

Zeyuan Allen-Zhu, Zheng Qu, Peter Richtárik, and Yang Yuan. Even faster accelerated coordinate descent using non-uniform sampling. In *International Conference on Machine Learning*, pp. 1110–1119, 2016.

Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations. In *Advances in Neural Information Processing Systems*, pp. 14668–14679, 2019.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Antoine Bordes, Seyda Ertekin, Jason Weston, and Léon Bottou. Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6(Sep):1579–1619, 2005.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243*, 2020.

Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

WM Goodall. Television by pulse code modulation. *Bell System Technical Journal*, 30(1):33–49, 1951.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. *Proceedings of the 36th International Conference on Machine Learning, Long Beach, California*, 2019.

Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv:2002.05516*, 2020.

Samuel Horváth and Peter Richtárik. Nonconvex variance reduced optimization with arbitrary sampling. *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Samuel Horváth and Peter Richtárik. A better alternative to error feedback for communication-efficient distributed learning. *arXiv preprint arXiv:2006.11077*, 2020.

Samuel Horváth, Chen-Yu Ho, Ľudovit Horváth, Atal Narayan Sahu, Marco Canini, and Peter Richtárik. Natural compression for distributed deep learning. *arXiv preprint arXiv:1905.10988*, 2019.

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

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 795–811. Springer, 2016.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for on-device federated learning. *arXiv preprint arXiv:1910.06378*, 2019.

Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. *arXiv preprint arXiv:1803.00942*, 2018.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local SGD on identical and heterogeneous data. In *The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*, 2020.

Jakub Konečný and Peter Richtárik. Randomized distributed mean estimation: Accuracy vs. communication. *Frontiers in Applied Mathematics and Statistics*, 4:62, 2018.

Fan Lai, Xiangfeng Zhu, Harsha V Madhyastha, and Mosharaf Chowdhury. Oort: Efficient federated learning via guided participant selection. In *15th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 21)*, pp. 19–35, 2021.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=K5YasWXZT3O`.

Qihang Lin, Zhaosong Lu, and Lin Xiao. An accelerated proximal coordinate gradient method. In *Advances in Neural Information Processing Systems*, pp. 3059–3067, 2014.

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

Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015.

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

Konstantin Mishchenko, Filip Hanzely, and Peter Richtárik. 99% of parallel optimization is inevitably a waste of time. *arXiv preprint arXiv:1901.09437*, 2019.

Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in neural information processing systems*, pp. 1017–1025, 2014.

Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

Zheng Qu, Peter Richtárik, and Tong Zhang. Quartz: Randomized dual coordinate ascent with arbitrary sampling. In *Advances in Neural Information Processing Systems 28*, pp. 865–873, 2015.

Ali Ramezani-Kebrya, Fartash Faghri, and Daniel M Roy. NUQSGD: Improved communication efficiency for data-parallel SGD via nonuniform quantization. *arXiv preprint arXiv:1908.06077*, 2019.

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

Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

Lawrence Roberts. Picture coding using pseudo-random noise. *IRE Transactions on Information Theory*, 8 (2):145–154, 1962.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. In *International conference on machine learning*, pp. 64–72, 2014.

Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 118–126, 2015.

Sebastian U Stich. Local SGD converges fast and communicates little. *ICLR 2019 - International Conference on Learning Representations*, 2019.

Sebastian U Stich and Sai Praneeth Karimireddy. The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *ICLR 2020 - International Conference on Learning Representations*, 2020.

Sebastian U Stich, Anant Raj, and Martin Jaggi. Safe adaptive importance sampling. In *Advances in Neural Information Processing Systems*, pp. 4381–4391, 2017.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified SGD with memory. In *Advances in Neural Information Processing Systems*, pp. 4447–4458, 2018.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. PowerSGD: Practical low-rank gradient compression for distributed optimization. In *Advances in Neural Information Processing Systems*, pp. 14236–14245, 2019.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pp. 1299–1309, 2018.

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. In *Advances in Neural Information Processing Systems*, pp. 1509–1519, 2017.

Hantian Zhang, Jerry Li, Kaan Kara, Dan Alistarh, Ji Liu, and Ce Zhang. Zipml: Training linear models with end-to-end low precision, and a little bit of deep learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 4035–4043. JMLR. org, 2017.

Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *international conference on machine learning*, pp. 1–9, 2015.

A Proof of Lemma 2.1

Proof. Our proof technique can be seen as an extended version of that in (Horváth & Richtárik, 2019). Let $1_{i\in S}=1$ if $i\in S$ and $1_{i\in S}=0$ otherwise. Likewise, let $1_{i,j\in S}=1$ if $i,j\in S$ and $1_{i,j\in S}=0$ otherwise. Note that $\mathrm{E}\left[1_{i\in S}\right]=p_i$ and $\mathrm{E}\left[1_{i,j\in S}\right]=p_{ij}$. Next, let us compute the mean of $X\coloneqq\sum_{i\in S}\frac{w_i\zeta_i}{p_i}$:

$$\mathrm{E}\left[X\right]=\mathrm{E}\left[\sum_{i\in S}\frac{w_i\zeta_i}{p_i}\right]=\mathrm{E}\left[\sum_{i=1}^n\frac{w_i\zeta_i}{p_i}1_{i\in S}\right]=\sum_{i=1}^n\frac{w_i\zeta_i}{p_i}\mathrm{E}\left[1_{i\in S}\right]=\sum_{i=1}^n w_i\zeta_i=\tilde{\zeta}.$$

Let $\boldsymbol{A}=[a_1,\ldots,a_n]\in\mathbb{R}^{d\times n}$, where $a_i=\frac{w_i\zeta_i}{p_i}$, and let e be the vector of all ones in $\mathbb{R}^n$. We now write the variance of X in a form which will be convenient to establish a bound:

$$\begin{aligned}
\mathrm{E}\left[\left\|X-\mathrm{E}\left[X\right]\right\|^2\right]&=\mathrm{E}\left[\left\|X\right\|^2\right]-\left\|\mathrm{E}\left[X\right]\right\|^2\\
&=\mathrm{E}\left[\left\|\sum_{i\in S}\frac{w_i\zeta_i}{p_i}\right\|^2\right]-\left\|\tilde{\zeta}\right\|^2\\
&=\mathrm{E}\left[\sum_{i,j}\frac{w_i\zeta_i^\top}{p_i}\frac{w_j\zeta_j}{p_j}1_{i,j\in S}\right]-\left\|\tilde{\zeta}\right\|^2\\
&=\sum_{i,j}p_{ij}\frac{w_i\zeta_i^\top}{p_i}\frac{w_j\zeta_j}{p_j}-\sum_{i,j}w_iw_j\zeta_i^\top\zeta_j\\
&=\sum_{i,j}(p_{ij}-p_ip_j)a_i^\top a_j\\
&=e^\top((\boldsymbol{P}-pp^\top)\circ\boldsymbol{A}^\top\boldsymbol{A})e.
\end{aligned}\tag{23}$$

Since, by assumption, we have $\boldsymbol{P}-pp^\top\preceq\mathbf{Diag}(p\circ v)$, we can further bound

$$e^\top((\boldsymbol{P}-pp^\top)\circ\boldsymbol{A}^\top\boldsymbol{A})e\leq e^\top(\mathbf{Diag}(p\circ v)\circ\boldsymbol{A}^\top\boldsymbol{A})e=\sum_{i=1}^n p_iv_i\left\|a_i\right\|^2.$$

To obtain (5), it remains to combine this with (23). The inequality $v_i\ \geq 1-p_i$ follows by comparing the diagonal elements of the two matrices in (4). Consider now the independent sampling. Clearly,

$$\boldsymbol{P}-pp^\top=\begin{bmatrix}p_1(1-p_1)&0&\ldots&0\\0&p_2(1-p_2)&\ldots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\ldots&p_n(1-p_n)\end{bmatrix}=\mathbf{Diag}(p_1v_1,\ldots,p_nv_n),$$

which implies $v_i=1-p_i$. □

B The Improvement Factor for Optimal Client Sampling

By Lemma 2.1, the independent sampling (which operates by independently flipping a coin and with probability p_i includes element i into S) is optimal. In addition, for independent sampling, (5) holds as equality. Thus, letting $\tilde{U}_i^k=w_i\mathbf{U}_i^k$, we have

$$\tilde{\alpha}_{S^k}\coloneqq\mathrm{E}\left[\left\|\sum_{i\in S^k}\frac{w_i}{p_i^k}\mathbf{U}_i^k-\sum_{i=1}^n w_i\mathbf{U}_i^k\right\|^2\right]=\mathrm{E}\left[\left\|\sum_{i\in S^k}\frac{1}{p_i^k}\tilde{U}_i^k-\sum_{i=1}^n\tilde{U}_i^k\right\|^2\right]=\mathrm{E}\left[\sum_{i=1}^n\frac{1-p_i^k}{p_i^k}\left\|\tilde{U}_i^k\right\|^2\right].\tag{24}$$

The optimal probabilities are obtained by optimizing (24) subject to the constraints $0 \leq p_i^k \leq 1$ and $m \geq b^k = \sum_{i=1}^n p_i^k$ using KKT conditions. Using a similar argument in (Horváth & Richtárik, 2019) (Lemma 2) gives the following solution

$$p_i^k = \begin{cases} (m+l-n)\frac{\left\|\tilde{U}_i^k\right\|}{\sum_{j=1}^l \left\|\tilde{U}_{(j)}^k\right\|}, & \text{if } i \notin A^k, \\ 1, & \text{if } i \in A^k, \end{cases} \tag{25}$$

where $\left\|\tilde{U}_{(j)}^k\right\|$ is the j-th largest value among the values $\left\|\tilde{U}_1^k\right\|, \left\|\tilde{U}_2^k\right\|, \ldots, \left\|\tilde{U}_n^k\right\|$, l is the largest integer for which $0 < m+l-n \leq \frac{\sum_{i=1}^l \left\|\tilde{U}_{(i)}^k\right\|}{\left\|\tilde{U}_{(l)}^k\right\|}$ (note that this inequality at least holds for $l = n-m+1$), and A^k contains indices i such that $\left\|\tilde{U}_i^k\right\| \geq \left\|\tilde{U}_{(l+1)}^k\right\|$.

Plugging the optimal probabilities obtained in (25) into (24) gives

$$\tilde{\alpha}_{S^k}^\star = \mathrm{E}\left[\sum_{i=1}^n \frac{1}{p_i^k}\left\|\tilde{U}_i^k\right\|^2 - \sum_{i=1}^n \left\|\tilde{U}_i^k\right\|^2\right] = \mathrm{E}\left[\frac{1}{m-(n-l)}\left(\sum_{i=1}^l \left\|\tilde{U}_{(i)}^k\right\|\right)^2 - \sum_{i=1}^l \left\|\tilde{U}_{(i)}^k\right\|^2\right].$$

With $m\left\|\tilde{U}_{(n)}^k\right\| \leq \sum_{i=1}^n \left\|\tilde{U}_i^k\right\|$, we have

$$\begin{aligned}\tilde{\alpha}_{S^k}^\star = \mathrm{E}\left[\frac{1}{m}\left(\sum_{i=1}^n \left\|\tilde{U}_i^k\right\|\right)^2 - \sum_{i=1}^n \left\|\tilde{U}_i^k\right\|^2\right] &= \mathrm{E}\left[\frac{1}{m}\left(\sum_{i=1}^n \left\|\tilde{U}_i^k\right\|\right)^2 \left(1 - m\frac{\sum_{i=1}^n \left\|\tilde{U}_i^k\right\|^2}{\left(\sum_{i=1}^n \left\|\tilde{U}_i^k\right\|\right)^2}\right)\right] \\ &\leq \frac{n-m}{nm}\mathrm{E}\left[\left(\sum_{i=1}^n \left\|\tilde{U}_i^k\right\|\right)^2\right].\end{aligned}$$

For independent uniform sampling $U^k \sim \mathbb{U}$ ($p_i^U = \frac{m}{n}$ for all i), we have

$$\tilde{\alpha}_{U^k} \coloneqq \mathrm{E}\left[\left\|\sum_{i\in U^k}\frac{w_i}{p_i^U}\mathbf{U}_i^k - \sum_{i=1}^n w_i\mathbf{U}_i^k\right\|^2\right] = \mathrm{E}\left[\sum_{i=1}^n \frac{1-\frac{m}{n}}{\frac{m}{n}}\left\|\tilde{U}_i^k\right\|^2\right] = \frac{n-m}{m}\mathrm{E}\left[\sum_{i=1}^n \left\|\tilde{U}_i^k\right\|^2\right].$$

Putting them together gives the improvement factor:

$$\alpha^k \coloneqq \frac{\tilde{\alpha}_{S^k}^\star}{\tilde{\alpha}_{U^k}} = \frac{\mathrm{E}\left[\left\|\sum_{i\in S^k}\frac{w_i}{p_i^k}\mathbf{U}_i^k - \sum_{i=1}^n w_i\mathbf{U}_i^k\right\|^2\right]}{\mathrm{E}\left[\left\|\sum_{i\in U^k}\frac{w_i}{p_i^U}\mathbf{U}_i^k - \sum_{i=1}^n w_i\mathbf{U}_i^k\right\|^2\right]} \leq \frac{\mathrm{E}\left[\left(\sum_{i=1}^n \left\|\tilde{U}_i^k\right\|\right)^2\right]}{n\mathrm{E}\left[\sum_{i=1}^n \left\|\tilde{U}_i^k\right\|^2\right]} \leq 1,$$

The upper bound is attained when all $\left\|\tilde{U}_i^k\right\|$ are identical. Note that the lower bound 0 can also be attained in the case where the number of non-zero updates is at most m. These considerations are discussed in the main paper.

C DSGD with Optimal Client Sampling

C.1 Proof of Theorem 3.6

Proof. L-smoothness of f_i and the assumption on the gradient imply that the inequality

$$\mathrm{E}\left[\left\|g_i^k\right\|^2\right] \leq 2L(1+M)(f_i(x^k) - f_i(x^\star) + R_i) + \sigma^2$$

holds for all $k \geq 0$. We first take expectations over x^{k+1} conditioned on x^k and over the sampling S^k:

$$\begin{aligned}
\mathrm{E}\left[\left\|r^{k+1}\right\|^2\right] &= \left\|r^k\right\|^2 - 2\eta^k \mathrm{E}\left[\left\langle \sum_{i\in S^k} \frac{w_i}{p_i^k} g_i^k, r^k \right\rangle\right] + (\eta^k)^2 \mathrm{E}\left[\left\| \sum_{i\in S^k} \frac{w_i}{p_i^k} g_i^k \right\|^2\right] \\
&= \left\|r^k\right\|^2 - 2\eta^k \left\langle \nabla f(x^k), r^k \right\rangle + (\eta^k)^2 \left(\mathrm{E}\left[\left\| \sum_{i\in S^k} \frac{w_i}{p_i^k} g_i^k - \sum_{i=1}^n w_i g_i^k \right\|^2\right] + \mathrm{E}\left[\left\| \sum_{i=1}^n w_i g_i^k \right\|^2\right] \right) \\
&\leq (1-\mu\eta^k)\left\|r^k\right\|^2 - 2\eta^k \left(f(x^k) - f^\star\right) + (\eta^k)^2 \left(\mathrm{E}\left[\left\| \sum_{i\in S^k} \frac{w_i}{p_i^k} g_i^k - \sum_{i=1}^n w_i g_i^k \right\|^2\right] + \mathrm{E}\left[\left\| \sum_{i=1}^n w_i g_i^k \right\|^2\right] \right),
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{E}\left[\left\| \sum_{i\in S^k} \frac{w_i}{p_i^k} g_i^k - \sum_{i=1}^n w_i g_i^k \right\|^2\right] &= \alpha^k \frac{n-m}{m} \mathrm{E}\left[\sum_{i=1}^n w_i^2 \left\|g_i^k\right\|^2\right] \\
&= \alpha^k \frac{n-m}{m} \mathrm{E}\left[\sum_{i=1}^n w_i^2 \left(\left\|g_i^k - \nabla f_i(x^k)\right\|^2 + \left\|\nabla f_i(x^k)\right\|^2\right)\right] \\
&= \alpha^k \frac{n-m}{m} \mathrm{E}\left[\sum_{i=1}^n w_i^2 \left(\left\|\xi_i^k\right\|^2 + \left\|\nabla f_i(x^k)\right\|^2\right)\right] \\
&\leq \alpha^k \frac{n-m}{m} \sum_{i=1}^n w_i^2 \left(2L(1+M)(f_i(x^k) - f_i(x^\star) + R_i) + \sigma^2\right) \\
&\leq \alpha^k \frac{n-m}{m} \left(2WL(1+M)(f(x^k) - f^\star) + \sum_{i=1}^n w_i^2 (2L(1+M)R_i + \sigma^2)\right),
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left[\left\| \sum_{i=1}^n w_i g_i^k \right\|^2\right] &= \mathrm{E}\left[\left\| \sum_{i=1}^n w_i g_i^k - \nabla f(x^k) \right\|^2\right] + \left\|\nabla f(x^k)\right\|^2 \\
&= \sum_{i=1}^n \mathrm{E}\left[\left\| w_i g_i^k - w_i \nabla f_i(x^k) \right\|^2\right] + \left\|\nabla f(x^k)\right\|^2 \\
&= \sum_{i=1}^n w_i^2 \mathrm{E}\left[\left\|\xi_i^k\right\|^2\right] + \left\|\nabla f(x^k)\right\|^2 \\
&\leq \sum_{i=1}^n w_i^2 (2LM(f_i(x^k) - f_i^\star) + \sigma^2) + 2L(f(x^k) - f^\star) \\
&= 2L\left(1 + WM\right)\left(f(x^k) - f^\star\right) + \sum_{i=1}^n w_i^2 (2LMR_i + \sigma^2).
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
\mathrm{E}\left[\left\|r^{k+1}\right\|^2\right] &\leq (1-\mu\eta^k)\left\|r^k\right\|^2 - 2\eta^k\left(f(x^k)-f^\star\right) \\
&\quad + (\eta^k)^2\left(2L\left(1+WM\right)\left(f(x^k)-f^\star\right) + \sum_{i=1}^n w_i^2(2LMR_i+\sigma^2)\right) \\
&\quad + (\eta^k)^2\alpha^k\frac{n-m}{m}\left(2WL(1+M)(f(x^k)-f^\star) + \sum_{i=1}^n w_i^2(2L(1+M)R_i+\sigma^2)\right) \\
&\leq (1-\mu\eta^k)\left\|r^k\right\|^2 - 2\eta^k\left(1-\eta^k\frac{(\alpha^k(n-m)+m)(1+WM)L}{m}\right)\left(f(x^k)-f^\star\right) \\
&\quad + (\eta^k)^2\frac{\alpha^k(n-m)+m}{m}\left(\sum_{i=1}^n w_i^2(2L(1+M)R_i+\sigma^2)\right) - (\eta^k)^2 2L\sum_{i=1}^n w_i^2 R_i.
\end{aligned}$$

Now choose any $0 < \eta^k \leq \frac{m}{(\alpha^k(n-m)+m)(1+WM)L}$ and define

$$\beta_1 \coloneqq \sum_{i=1}^n w_i^2(2L(1+M)R_i+\sigma^2), \quad \beta_2 \coloneqq 2L\sum_{i=1}^n w_i^2 R_i, \quad \gamma^k \coloneqq \frac{m}{\alpha^k(n-m)+m} \in \left[\frac{m}{n}, 1\right].$$

Taking full expectation yields the desired result:

$$\mathrm{E}\left[\left\|r^{k+1}\right\|^2\right] \leq (1-\mu\eta^k)\mathrm{E}\left[\left\|r^k\right\|^2\right] + (\eta^k)^2\left(\frac{\beta_1}{\gamma^k}-\beta_2\right).$$

□

C.2 Proof of Theorem 3.7

Proof. Using equation (2), we have

$$\begin{aligned}
f(x^{k+1}) &= f(x^k - \eta^k\mathbf{G}^k) \\
&= f(x^k) - \eta^k\left\langle\mathbf{G}^k, \nabla f(x^k)\right\rangle + \frac{(\eta^k)^2}{2}\left\langle\mathbf{G}^k, \nabla^2 f(z^k)\mathbf{G}^k\right\rangle, \quad \text{for some } z^k \in \mathbb{R}^d.
\end{aligned}$$

Since all f_i's are L-smooth, f is also L-smooth. Therefore, we have $-L\boldsymbol{I} \preceq \nabla^2 f(x) \preceq L\boldsymbol{I}$ for all $x \in \mathbb{R}^d$. Combining this with the fact that $\mathbf{G}^k$ is an unbiased estimator of $\nabla f(x^k)$, we have

$$\mathrm{E}\left[f(x^{k+1})\right] \leq f(x^k) - \eta^k\left\|\nabla f(x^k)\right\|^2 + \frac{(\eta^k)^2 L}{2}\mathrm{E}\left[\left\|\mathbf{G}^k\right\|^2\right], \tag{26}$$

where the expectations are conditioned on x^k. In Appendix C.1, we already obtained the upper bound for the last term in equation (26):

$$\begin{aligned}
\mathrm{E}\left[\left\|\mathbf{G}^k\right\|^2\right] &\leq \left((1+M)\alpha^k\frac{n-m}{m}+M\right)\sum_{i=1}^n w_i^2\left\|\nabla f_i(x^k)\right\|^2 + \left(\alpha^k\frac{n-m}{m}+1\right)\sum_{i=1}^n w_i^2\sigma^2 + \left\|\nabla f(x^k)\right\|^2 \\
&= \left(\frac{1+M}{\gamma^k}-1\right)\sum_{i=1}^n w_i^2\left\|\nabla f_i(x^k)\right\|^2 + \frac{1}{\gamma^k}\sum_{i=1}^n w_i^2\sigma^2 + \left\|\nabla f(x^k)\right\|^2.
\end{aligned}$$

By Assumption 3.5, we further bound

$$\begin{aligned}
\sum_{i=1}^n w_i^2\left\|\nabla f_i(x^k)\right\|^2 &\leq W\sum_{i=1}^n w_i\left\|\nabla f_i(x^k)\right\|^2 \\
&\leq W\left(\sum_{i=1}^n w_i\left\|\nabla f_i(x^k)-\nabla f(x^k)\right\|^2 + \left\|\nabla f(x^k)\right\|^2\right) \\
&\leq W\rho + \left\|\nabla f(x^k)\right\|^2.
\end{aligned}$$

Combining the inequalities above and taking full expectation yields equation (19). □

D FedAvg with Optimal Client Sampling

Lemma D.1 ((Karimireddy et al., 2019)). *For any L-smooth and μ-strongly convex function $h : \mathbb{R}^d \to \mathbb{R}$ and any $x, y, z \in \mathbb{R}^d$, the following inequality holds*

$$\langle \nabla h(x), z - y \rangle \geq h(z) - h(y) + \frac{\mu}{4} \|y - z\|^2 - L \|z - x\|^2 . \tag{27}$$

Proof. For any given x, y, and z, the two inequalities below follows by the smoothness and strong convexity of the function h:

$$\langle \nabla h(x), z - x \rangle \geq h(z) - h(x) - \frac{L}{2} \|z - x\|^2 ,$$
$$\langle \nabla h(x), x - y \rangle \geq h(x) - h(y) + \frac{\mu}{2} \|y - x\|^2 .$$

Further, applying the relaxed triangle inequality gives

$$\frac{\mu}{2} \|y - x\|^2 \geq \frac{\mu}{4} \|y - z\|^2 - \frac{\mu}{2} \|x - z\|^2 .$$

Combining all these inequalities together we have

$$\langle \nabla h(x), z - y \rangle \geq h(z) - h(y) + \frac{\mu}{4} \|y - z\|^2 - \frac{L + \mu}{2} \|z - x\|^2 .$$

The lemma follows by $L \geq \mu$. □

D.1 Proof of Theorem 3.8

Proof. The master update during round k can be written as (superscript k is dropped from here onward)

$$\eta_g \Delta x = \frac{\eta}{R} \sum_{i \in S, r} \frac{w_i}{p_i} g_i(y_{i,r-1}) \quad \text{and} \quad \mathrm{E}\left[\eta_g \Delta x\right] = \frac{\eta}{R} \sum_{i,r} w_i \mathrm{E}\left[\nabla f_i(y_{i,r-1})\right].$$

Summations are always over $i \in [n]$ and $r \in [R]$ unless stated otherwise. Taking expectations over x conditioned on the results prior to round k and over the sampling S gives

$$\mathrm{E}\left[\|x - \eta_g \Delta x - x^\star\|^2\right] = \|x - x^\star\|^2 \underbrace{- \frac{2\eta}{R} \sum_{i,r} \langle w_i \nabla f_i(y_{i,r-1}), x - x^\star \rangle}_{\mathcal{A}_1} + \underbrace{\frac{\eta^2}{R^2} \mathrm{E}\left[\left\| \sum_{i \in S, r} \frac{w_i}{p_i} g_i(y_{i,r-1}) \right\|^2\right]}_{\mathcal{A}_2} .$$

Applying Lemma D.1 with $h = w_i f_i$, $x = y_{i,r-1}$, $y = x^\star$ and $z = x$ gives

$$\begin{aligned}
\mathcal{A}_1 &\leq -\frac{2\eta}{R} \sum_{i,r} \left(w_i f_i(x) - w_i f_i(x^\star) + w_i \frac{\mu}{4} \|x - x^\star\|^2 - w_i L \|x - y_{i,r-1}\|^2 \right) \\
&\leq -2\eta \left(f(x) - f^\star + \frac{\mu}{4} \|x - x^\star\|^2 \right) + 2L\eta\mathcal{E},
\end{aligned}$$

where $\mathcal{E}$ is the drift caused by the local updates on the clients:

$$\mathcal{E} \coloneqq \frac{1}{R} \sum_{i,r} w_i \mathrm{E}\left[\|x - y_{i,r-1}\|^2\right]. \tag{28}$$

Bounding $\mathcal{A}_2$, we obtain

$$\begin{aligned}
\frac{1}{\eta^2}\mathcal{A}_2 &= \mathrm{E}\left[\left\|\sum_{i\in S}\frac{w_i}{p_i}\frac{1}{R}\sum_r g_i(y_{i,r-1}) - \sum_i w_i\frac{1}{R}\sum_r g_i(y_{i,r-1})\right\|^2\right] + \mathrm{E}\left[\left\|\sum_i w_i\frac{1}{R}\sum_r g_i(y_{i,r-1})\right\|^2\right] \\
&\leq \alpha\frac{n-m}{m}\sum_i w_i^2\mathrm{E}\left[\left\|\frac{1}{R}\sum_r g_i(y_{i,r-1})\right\|^2\right] + \mathrm{E}\left[\left\|\sum_i w_i\frac{1}{R}\sum_r g_i(y_{i,r-1})\right\|^2\right] \\
&= \alpha\frac{n-m}{m}\sum_i w_i^2\left(\mathrm{E}\left[\left\|\frac{1}{R}\sum_r \xi_{i,r-1}\right\|^2\right] + \mathrm{E}\left[\left\|\frac{1}{R}\sum_r \nabla f_i(y_{i,r-1})\right\|^2\right]\right) \\
&\quad + \mathrm{E}\left[\left\|\sum_i w_i\frac{1}{R}\sum_r \xi_{i,r-1}\right\|^2\right] + \mathrm{E}\left[\left\|\sum_i w_i\frac{1}{R}\sum_r \nabla f_i(y_{i,r-1})\right\|^2\right].
\end{aligned}$$

Using independence, zero mean and bounded second moment of the random variables $\xi_{i,r}$, we obtain

$$\begin{aligned}
\frac{1}{\eta^2}\mathcal{A}_2 &\leq \alpha\frac{n-m}{m}\sum_i w_i^2\left(\frac{1}{R^2}\sum_r \mathrm{E}\left[\|\xi_{i,r-1}\|^2\right] + \mathrm{E}\left[\left\|\frac{1}{R}\sum_r \nabla f_i(y_{i,r-1})\right\|^2\right]\right) \\
&\quad + \sum_i w_i^2\frac{1}{R^2}\sum_r \mathrm{E}\left[\|\xi_{i,r-1}\|^2\right] + \mathrm{E}\left[\left\|\sum_i w_i\frac{1}{R}\sum_r \nabla f_i(y_{i,r-1})\right\|^2\right] \\
&\leq \alpha\frac{n-m}{m}\sum_i w_i^2\left(\left(\frac{M}{R^2}+\frac{1}{R}\right)\sum_r \mathrm{E}\left[\|\nabla f_i(y_{i,r-1})\|^2\right] + \frac{\sigma^2}{R}\right) \\
&\quad + \sum_i w_i^2\left(\frac{M}{R^2}\sum_r \mathrm{E}\left[\|\nabla f_i(y_{i,r-1})\|^2\right] + \frac{\sigma^2}{R}\right) + \mathrm{E}\left[\left\|\sum_i w_i\frac{1}{R}\sum_r \nabla f_i(y_{i,r-1})\right\|^2\right] \\
&= \frac{\sigma^2}{R\gamma}\sum_i w_i^2 + \left(\frac{M}{R} + \left(\frac{M}{R}+1\right)\alpha\frac{n-m}{m}\right)\sum_i w_i^2\frac{1}{R}\sum_r \mathrm{E}\left[\|\nabla f_i(y_{i,r-1}) - \nabla f_i(x) + \nabla f_i(x)\|^2\right] \\
&\quad + \mathrm{E}\left[\left\|\sum_i w_i\frac{1}{R}\sum_r (\nabla f_i(y_{i,r-1}) - \nabla f_i(x)) + \nabla f(x)\right\|^2\right] \\
&\leq \frac{\sigma^2}{R\gamma}\sum_i w_i^2 + \left(\frac{M}{R} + \left(\frac{M}{R}+1\right)\alpha\frac{n-m}{m}\right)\sum_i w_i^2\left(\frac{2}{R}\sum_r \mathrm{E}\left[\|\nabla f_i(y_{i,r-1}) - \nabla f_i(x)\|^2\right] + 2\mathrm{E}\left[\|\nabla f_i(x)\|^2\right]\right) \\
&\quad + 2\mathrm{E}\left[\left\|\sum_i w_i\frac{1}{R}\sum_r (\nabla f_i(y_{i,r-1}) - \nabla f_i(x))\right\|^2\right] + 2\mathrm{E}\left[\|\nabla f(x)\|^2\right].
\end{aligned}$$

Combining the smoothness of f_i's, the definition of $\mathcal{E}$, and Jensen's inequality with definition $\gamma \coloneqq \frac{m}{\alpha(n-m)+m}$, we obtain

$$\begin{aligned}
\frac{1}{\eta^2}\mathcal{A}_2 &\leq \frac{\sigma^2}{R\gamma}\sum_i w_i^2 + 2\left(\frac{M}{R} + \left(\frac{M}{R}+1\right)\alpha\frac{n-m}{m}\right)\left(WL^2\mathcal{E} + 2WL(f(x)-f^\star) + 2L\sum_i w_i^2 R_i\right) \\
&\quad + 2L^2\mathcal{E} + 4L(f(x) - f(x^\star)) \\
&= \frac{\sigma^2}{R\gamma}\sum_i w_i^2 + 2L^2\left((1-W) + \frac{W}{\gamma}\left(\frac{M}{R}+1\right)\right)\mathcal{E} + 4L\left(\frac{1}{\gamma}\left(\frac{M}{R}+1\right) - 1\right)\sum_i w_i^2 R_i \\
&\quad + 4L\left((1-W) + \frac{W}{\gamma}\left(\frac{M}{R}+1\right)\right)(f(x) - f^\star).
\end{aligned}$$

Putting these bounds on $\mathcal{A}_1$ and $\mathcal{A}_2$ together and using the fact that $1 - W \leq 1/\gamma$ yields

$$\begin{aligned}\mathrm{E}\left[\left\|x - \eta_g \Delta x - x^\star\right\|^2\right] \leq \left(1 - \frac{\mu\eta}{2}\right)\left\|x - x^\star\right\|^2 - 2\eta\left(1 - 2L\frac{\eta}{\gamma}\left(W\left(\frac{M}{R}+1\right)+1\right)\right)(f(x) - f^\star) \\ + \eta^2\left(\frac{\sigma^2}{R\gamma}\sum_i w_i^2 + 4L\left(\frac{1}{\gamma}\left(\frac{M}{R}+1\right)-1\right)\sum_i w_i^2 R_i\right) \\ + \left(1 + \eta L\left((1-W) + \frac{W}{\gamma}\left(\frac{M}{R}+1\right)\right)\right) 2L\eta\mathcal{E}.\end{aligned}$$

Let $\eta \leq \frac{\gamma}{8(1+W(1+M/R))L}$, then

$$\frac{3}{4} \leq 1 - 2L\frac{\eta}{\gamma}\left(W\left(\frac{M}{R}+1\right)+1\right),$$

which in turn yields

$$\begin{aligned}\mathrm{E}\left[\left\|x - \eta_g \Delta x - x^\star\right\|^2\right] \leq \left(1 - \frac{\mu\eta}{2}\right)\left\|x - x^\star\right\|^2 - \frac{3\eta}{2}(f(x) - f^\star) \\ + \eta^2\left(\frac{\sigma^2}{R\gamma}\sum_i w_i^2 + 4L\left(\frac{1}{\gamma}\left(\frac{M}{R}+1\right)-1\right)\sum_i w_i^2 R_i\right) \\ + \left(1 + \eta L\left((1-W) + \frac{W}{\gamma}\left(\frac{M}{R}+1\right)\right)\right) 2L\eta\mathcal{E}. \qquad (29)\end{aligned}$$

Next, we need to bound the drift $\mathcal{E}$. For $R \geq 2$, we have

$$\begin{aligned}\mathrm{E}\left[\left\|y_{i,r} - x\right\|^2\right] &= \mathrm{E}\left[\left\|y_{i,r-1} - x - \eta_l g_i(y_{i,r-1})\right\|^2\right] \\ &\leq \mathrm{E}\left[\left\|y_{i,r-1} - x - \eta_l \nabla f_i(y_{i,r-1})\right\|^2\right] + \eta_l^2(M\left\|\nabla f_i(y_{i,r-1})\right\|^2 + \sigma^2) \\ &\leq \left(1 + \frac{1}{R-1}\right)\mathrm{E}\left[\left\|y_{i,r-1} - x\right\|^2\right] + (R+M)\eta_l^2\left\|\nabla f_i(y_{i,r-1})\right\|^2 + \eta_l^2\sigma^2 \\ &= \left(1 + \frac{1}{R-1}\right)\mathrm{E}\left[\left\|y_{i,r-1} - x\right\|^2\right] + \left(1 + \frac{M}{R}\right)\frac{\eta^2}{R\eta_g^2}\left\|\nabla f_i(y_{i,r-1})\right\|^2 + \frac{\eta^2\sigma^2}{R^2\eta_g^2} \\ &\leq \left(1 + \frac{1}{R-1}\right)\mathrm{E}\left[\left\|y_{i,r-1} - x\right\|^2\right] + \left(1 + \frac{M}{R}\right)\frac{2\eta^2}{R\eta_g^2}\left\|\nabla f_i(y_{i,r-1}) - \nabla f_i(x)\right\|^2 \\ &\quad + \left(1 + \frac{M}{R}\right)\frac{2\eta^2}{R\eta_g^2}\left\|\nabla f_i(x)\right\|^2 + \frac{\eta^2\sigma^2}{R^2\eta_g^2} \\ &\leq \left(1 + \frac{1}{R-1} + \left(1 + \frac{M}{R}\right)\frac{2\eta^2L^2}{R\eta_g^2}\right)\mathrm{E}\left[\left\|y_{i,r-1} - x\right\|^2\right] + \left(1 + \frac{M}{R}\right)\frac{2\eta^2}{R\eta_g^2}\left\|\nabla f_i(x)\right\|^2 + \frac{\eta^2\sigma^2}{R^2\eta_g^2}.\end{aligned}$$

If we further restrict $\eta \leq \frac{1}{8L(2+M/R)}$, then for any $\eta_g \geq 1$, we have

$$\left(1 + \frac{M}{R}\right)\frac{2\eta^2L^2}{R\eta_g^2} \leq \frac{2L^2}{R\eta_g^2}\frac{1}{64L^2} \leq \frac{1}{32R} \leq \frac{1}{32(R-1)},$$

and therefore,

$$\begin{aligned}
\mathrm{E}\left[\left\|y_{i,r}-x\right\|^2\right] &\leq \left(1+\frac{33}{32(R-1)}\right)\mathrm{E}\left[\left\|y_{i,r-1}-x\right\|^2\right]+\left(1+\frac{M}{R}\right)\frac{2\eta^2}{R\eta_g^2}\left\|\nabla f_i(x)\right\|^2+\frac{\eta^2\sigma^2}{R^2\eta_g^2}\\
&\leq \sum_{\tau=0}^{r-1}\left(1+\frac{33}{32(R-1)}\right)^{\tau}\left(\left(1+\frac{M}{R}\right)\frac{2\eta^2}{R\eta_g^2}\left\|\nabla f_i(x)\right\|^2+\frac{\eta^2\sigma^2}{R^2\eta_g^2}\right)\\
&\leq 8R\left(\left(1+\frac{M}{R}\right)\frac{2\eta^2}{R\eta_g^2}\left\|\nabla f_i(x)\right\|^2+\frac{\eta^2\sigma^2}{R^2\eta_g^2}\right)\\
&= 16\left(1+\frac{M}{R}\right)\eta^2\left\|\nabla f_i(x)\right\|^2+\frac{8\eta^2\sigma^2}{R\eta_g^2}.
\end{aligned}$$

Hence, the drift is bounded by

$$\begin{aligned}
\mathcal{E} &\leq 16\left(1+\frac{M}{R}\right)\eta^2\sum_i w_i\left\|\nabla f_i(x)\right\|^2+\frac{8\eta^2\sigma^2}{R\eta_g^2}\\
&\leq 32\left(1+\frac{M}{R}\right)\eta^2L\sum_i w_i(f_i(x)-f_i^\star)+\frac{8\eta^2\sigma^2}{R\eta_g^2}\\
&= 32\left(1+\frac{M}{R}\right)\eta^2L(f(x)-f^\star)+32\left(1+\frac{M}{R}\right)\eta^2L\sum_i w_iR_i+\frac{8\eta^2\sigma^2}{R\eta_g^2}\\
&\leq 4\eta(f(x)-f^\star)+32\left(1+\frac{M}{R}\right)\eta^2L\sum_i w_iR_i+\frac{8\eta^2\sigma^2}{R\eta_g^2}.
\end{aligned}$$

Due to the upper bound on the step size $\eta\leq\frac{1}{8L(2+M/R)}$, we have the inequalities

$$1+\eta L\left((1-W)+\frac{W}{\gamma}\left(\frac{M}{R}+1\right)\right)\leq\frac{9}{8}\quad\text{and}\quad 8\eta L\leq 1. \tag{30}$$

Plugging these to (29), we obtain

$$\begin{aligned}
\mathrm{E}\left[\left\|x-\eta_g\Delta x-x^\star\right\|^2\right] &\leq \left(1-\frac{\mu\eta}{2}\right)\left\|x-x^\star\right\|^2-\frac{3}{8}\eta(f(x)-f^\star)\\
&\quad+\eta^2\left(\frac{\sigma^2}{\gamma R}\left(\frac{\gamma}{\eta_g^2}+\sum_i w_i^2\right)+4L\left(\frac{M}{R}+1-\gamma\right)\sum_i w_i^2R_i\right)\\
&\quad+\eta^3 72L^2\left(1+\frac{M}{R}\right)\sum_i w_iR_i.
\end{aligned}$$

Rearranging the terms in the last inequality, taking full expectation and including superscripts lead to

$$\begin{aligned}
\frac{3}{8}\mathrm{E}\left[(f(x^k)-f^\star)\right] &\leq \frac{1}{\eta^k}\left(1-\frac{\mu\eta^k}{2}\right)\mathrm{E}\left[\left\|x^k-x^\star\right\|^2\right]-\frac{1}{\eta^k}\mathrm{E}\left[\left\|x^{k+1}-x^\star\right\|^2\right]\\
&\quad+\eta^k\left(\frac{\sigma^2}{\gamma^kR}\left(\frac{\gamma^k}{\eta_g^2}+\sum_i w_i^2\right)+4L\left(\frac{M}{R}+1-\gamma^k\right)\sum_i w_i^2R_i\right)\\
&\quad+(\eta^k)^2 72L^2\left(1+\frac{M}{R}\right)\sum_i w_iR_i.
\end{aligned}$$

Plugging the assumption $\eta_g^k\geq\sqrt{\frac{\gamma^k}{\sum_i w_i^2}}$ into the RHS of the above inequality completes the proof.

□

D.2 Proof of Theorem 3.9

Proof. We drop superscript k and write the master update during round k as:

$$\eta_g \Delta x = \frac{\eta}{R} \sum_{i \in S, r} \frac{w_i}{p_i} g_i(y_{i,r-1}) \coloneqq \eta \tilde{\Delta}.$$

Summations are always over $i \in [n]$ and $r \in [R]$ unless stated otherwise. Taking expectations conditioned on x and using a similar argument as in the proof in Appendix C.2, we have

$$\begin{aligned}
\mathrm{E}\left[f(x - \eta_g \Delta x)\right] &\leq f(x) - \eta \left\langle \nabla f(x), \mathrm{E}\left[\tilde{\Delta}\right] \right\rangle + \frac{\eta^2 L}{2} \mathrm{E}\left[\left\|\tilde{\Delta}\right\|^2\right] \\
&= f(x) - \eta \left\|\nabla f(x)\right\|^2 + \eta \left\langle \nabla f(x), \nabla f(x) - \mathrm{E}\left[\tilde{\Delta}\right] \right\rangle + \frac{\eta^2 L}{2} \mathrm{E}\left[\left\|\tilde{\Delta}\right\|^2\right] \\
&\leq f(x) - \frac{\eta}{2} \left\|\nabla f(x)\right\|^2 + \frac{\eta}{2} \mathrm{E}\left[\left\|\nabla f(x) - \mathrm{E}_S\left[\tilde{\Delta}\right]\right\|^2\right] + \frac{\eta^2 L}{2} \mathrm{E}\left[\left\|\tilde{\Delta}\right\|^2\right],
\end{aligned}$$

where the last inequality follows since $\langle a, b \rangle \leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2, \ \forall a, b \in \mathbb{R}^d$. Since f_i's are L-smooth, by the (relaxed) triangular inequality, we have

$$\begin{aligned}
\frac{\eta}{2} \mathrm{E}\left[\left\|\nabla f(x) - \mathrm{E}\left[\tilde{\Delta}\right]\right\|^2\right] &= \frac{\eta}{2} \mathrm{E}\left[\left\|\frac{1}{R} \sum_{i,r} w_i \left(\nabla f_i(x) - \nabla f_i(y_{i,r-1})\right)\right\|^2\right] \\
&\leq \frac{\eta L^2}{2R} \sum_{i,r} w_i \mathrm{E}\left[\left\|x - y_{i,r-1}\right\|^2\right] = \frac{\eta L^2}{2} \mathcal{E},
\end{aligned}$$

where $\mathcal{E}$ is the drift caused by the local updates on the clients as defined in (28).

In Appendix D.1, we already obtained the upper bound for $\frac{1}{\eta^2}\mathcal{A}_2 = \mathrm{E}\left[\left\|\tilde{\Delta}\right\|^2\right]$:

$$\mathrm{E}\left[\left\|\tilde{\Delta}\right\|^2\right] \leq \frac{\sigma^2}{R\gamma} \sum_i w_i^2 + 2W \left(\frac{M}{R} + \left(\frac{M}{R} + 1\right) \alpha \frac{n - m}{m}\right) \left(L^2 \mathcal{E} + \sum_i w_i \left\|\nabla f_i(x)\right\|^2\right) + 2L^2 \mathcal{E} + 2\left\|\nabla f(x)\right\|^2.$$

Together with Assumption 3.5 that

$$\sum_i w_i \left\|\nabla f_i(x)\right\|^2 - \left\|\nabla f(x)\right\|^2 \leq \sum_i w_i \left\|\nabla f_i(x) - \nabla f(x)\right\|^2 \leq \rho,$$

we have

$$\mathrm{E}\left[\left\|\tilde{\Delta}\right\|^2\right] \leq \frac{\sigma^2}{R\gamma} \sum_i w_i^2 + \frac{2W}{\gamma} \left(\frac{M}{R} + 1 - \gamma\right) \left(L^2 \mathcal{E} + \left\|\nabla f(x)\right\|^2 + \rho\right) + 2L^2 \mathcal{E} + 2\left\|\nabla f(x)\right\|^2.$$

Combining the above inequalities gives

$$\begin{aligned}
\mathrm{E}\left[f(x - \eta_g \Delta x)\right] \leq f(x) &+ \eta^2 \frac{\sigma^2 L}{2R\gamma} \sum_i w_i^2 + \eta L^2 \left(\eta L \left((1 - W) + \frac{W}{\gamma}\left(1 + \frac{M}{R}\right)\right) + \frac{1}{2}\right) \mathcal{E} \\
&+ \eta \left(\eta L \left((1 - W) + \frac{W}{\gamma}\left(1 + \frac{M}{R}\right)\right) - \frac{1}{2}\right) \left\|\nabla f(x)\right\|^2 \\
&+ \eta \left(\eta L \left((1 - W) + \frac{W}{\gamma}\left(1 + \frac{M}{R}\right)\right) - \eta L\right) \rho.
\end{aligned}$$

Now, applying inequality (30) gives

$$\mathrm{E}\left[f(x - \eta_g \Delta x)\right] \leq f(x) + \frac{\eta^2 \sigma^2 L}{2R\gamma} \sum_i w_i^2 + \frac{5\eta L^2}{8} \mathcal{E} - \frac{3\eta}{8} \left\|\nabla f(x)\right\|^2 + \frac{\eta}{8}(1 - 8\eta L)\rho.$$

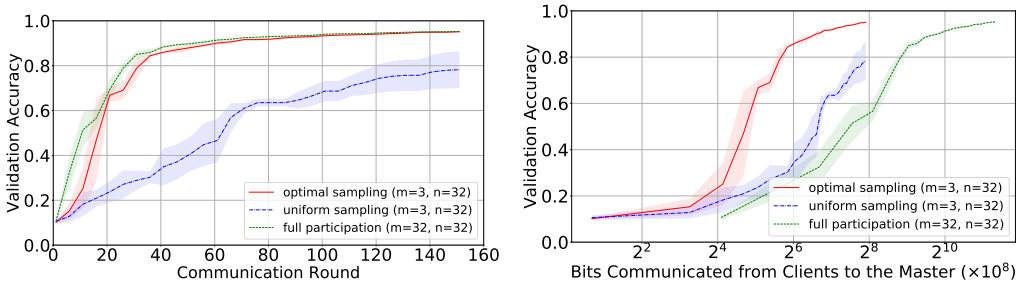

Figure 6: (FEMNIST Dataset 1) current best validation accuracy as a function of the number of communication rounds and the number of bits communicated from clients to the master.

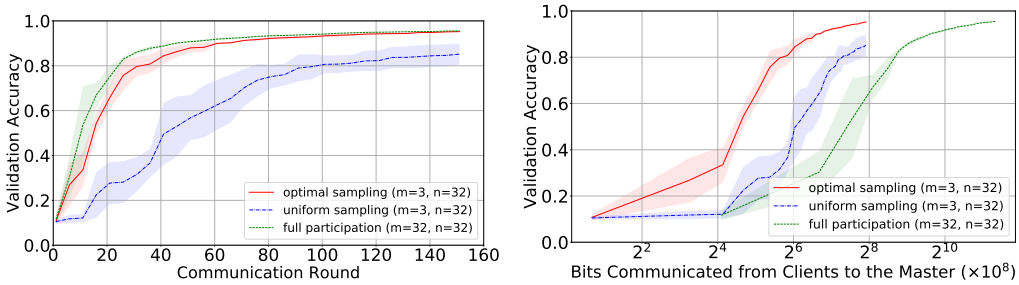

Figure 7: (FEMNIST Dataset 2) current best validation accuracy as a function of the number of communication rounds and the number of bits communicated from clients to the master.

In Appendix D.1, we also obtained the upper bound for the drift $\mathcal{E}$:

$$\begin{aligned}\mathcal{E} &\leq 16\left(1+\frac{M}{R}\right)\eta^2\sum_i w_i\left\|\nabla f_i(x)\right\|^2+\frac{8\eta^2\sigma^2}{R\eta_g^2}\\ &\leq 16\left(1+\frac{M}{R}\right)\eta^2(\left\|\nabla f(x)\right\|^2+\rho)+\frac{8\eta^2\sigma^2}{R\eta_g^2}.\end{aligned}$$

Since $8\eta L \leq 8\eta L(1+{}^{M}/_{R}) \leq 1$, we have

$$\begin{aligned}\frac{5\eta L^2}{8}\mathcal{E} &\leq 10\eta^3L^2\left(1+\frac{M}{R}\right)(\left\|\nabla f(x)\right\|^2+\rho)+\frac{5\eta^3L^2\sigma^2}{R\eta_g^2}\\ &\leq \frac{5\eta^2L}{4}(\left\|\nabla f(x)\right\|^2+\rho)+\frac{5\eta^2L\sigma^2}{8R\eta_g^2}.\end{aligned}$$

This further simplifies the iterate to

$$\mathrm{E}\left[f(x-\eta_g\Delta x)\right] \leq f(x)-\frac{3}{8}\eta\left(1-\frac{10}{3}\eta L\right)\left\|\nabla f(x)\right\|^2+\frac{1}{8}\eta\left(1+2\eta L\right)\rho+\frac{\eta^2\sigma^2L}{2R\gamma}\left(\frac{5\gamma}{4\eta_g^2}+\sum_i w_i^2\right).$$

Applying the assumption that $\eta_g \geq \sqrt{\frac{5\gamma}{4\sum_i w_i^2}}$ and taking full expectations completes the proof:

$$\mathrm{E}\left[f(x-\eta_g\Delta x)\right] \leq \mathrm{E}\left[f(x)\right]-\frac{3}{8}\eta\left(1-\frac{10}{3}\eta L\right)\mathrm{E}\left[\left\|\nabla f(x)\right\|^2\right]+\eta\frac{\rho}{8}+\eta^2\left(\frac{\rho}{4}+\frac{\sigma^2}{R\gamma}\sum_{i=1}^{n}w_i^2\right)L.$$

□

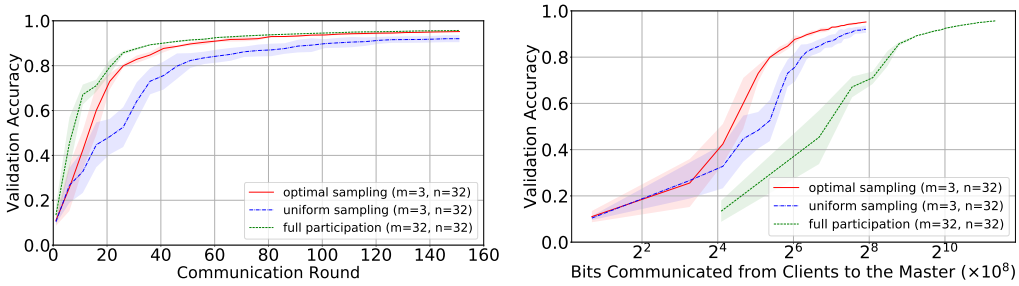

Figure 8: (FEMNIST Dataset 3) current best validation accuracy as a function of the number of communication rounds and the number of bits communicated from clients to the master.

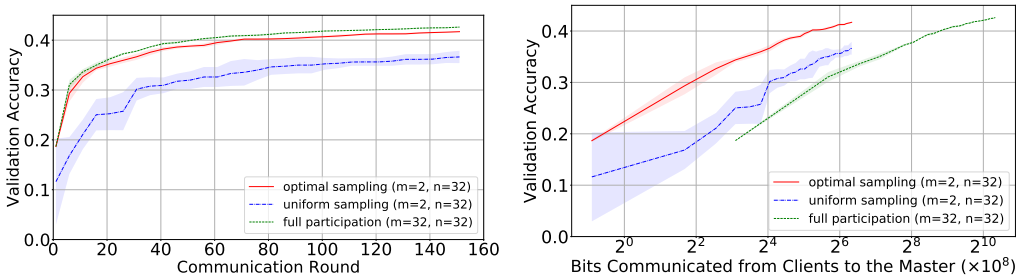

Figure 9: (Shakespeare Dataset) current best validation accuracy as a function of the number of communication rounds and the number of bits communicated from clients to the master.

E Experimental Details

E.1 Federated EMNIST Dataset

We detail the hyper-parameters used in the experiments on the FEMNIST datasets. For each experiment, we run 151 communication rounds, reporting (local) training loss every round and validation accuracy every 5 rounds. In each round, $n = 32$ clients are sampled from the client pool, each of which then performs SGD for 1 epoch on its local training images with batch size 20. For partial participation, the expected number of clients allowed to communicate their updates back to the master is set to $m = 3$. We use vanilla SGD and constant step sizes for all experiments, where we set $\eta_g = 1$ and tune η_l from the set of value $\{2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, 2^{-5}\}$. If the optimal step size hits a boundary value, then we try one more step size by extending that boundary and repeat this until the optimal step size is not a boundary value. For full participation and optimal sampling, it turns out that $\eta_l = 2^{-3}$ is the optimal local step size for all three datasets. For uniform sampling, the optimal is $\eta_l = 2^{-5}$ for Dataset 1 and $\eta_l = 2^{-4}$ for Datasets 2 and 3. For the extra communications in Algorithm 2, we set $j_{max} = 4$.

We also present some additional figures of the experiment results. Figures 6, 7 and 8 show the current best validation accuracy as a function of the number of communication rounds and the number of bits communicated from clients to the master on Datasets 1, 2 and 3, respectively.

E.2 Shakespeare Dataset

We detail the hyper-parameters used in the experiments on the Shakespeare dataset. For each experiment, we run 151 communication rounds, reporting (local) training loss every round and validation accuracy every 5 rounds. In each round, $n = 32$ clients are sampled from the client pool, each of which then performs SGD for 1 epoch on its local training data with batch size 8 (each batch contains 8 example sequences of length 5). For partial participation, the expected number of clients allowed to communicate their updates back to the master is set to $m = 2$. We use vanilla SGD and constant step sizes for all experiments, where we set $\eta_g = 1$ and tune η_l from the set of value $\{2^{-1}, 2^{-2}, 2^{-3}, 2^{-4}, 2^{-5}\}$. If the optimal step size hits a boundary value,

then we try one more step size by extending that boundary and repeat this until the optimal step size is not a boundary value. For full participation and optimal sampling, it turns out that $\eta_l = 2^{-2}$ is the optimal local step size. For uniform sampling, the optimal is $\eta_l = 2^{-3}$. For the extra communications in Algorithm 2, we set $j_{max} = 4$.

We also present an additional figure of the experiment result. Figure 9 shows the current best validation accuracy as a function of the number of communication rounds and the number of bits communicated from clients to the master.