Federated Learning under Importance Sampling

Elsa Rizk, Stefan Vlaski, Member, IEEE, and Ali H. Sayed, Fellow, IEEE.

Abstract—Federated learning encapsulates distributed learning strategies that are managed by a central unit. Since it relies on a selected number of agents at each iteration, and since each agent, in turn, taps into its local data, it is only natural to study optimal sampling policies for selecting agents and their data in federated learning implementations. Usually, only uniform sampling schemes are used. However, in this work, we examine the effect of importance sampling and devise schemes for sampling agents and data non-uniformly guided by a performance measure. We find that in schemes involving sampling without replacement, the performance of the resulting architecture is controlled by two factors related to data variability at each agent, and model variability across agents. We illustrate the theoretical findings with experiments on simulated and real data and show the improvement in performance that results from the proposed strategies.

Index Terms—federated learning, distributed learning, importance sampling, asynchronous SGD, non-IID data, heterogeneous agents

I. INTRODUCTION

In this work, we focus on algorithms that fall into the broad class of stochastic gradient descent (SGD). We consider a collection of \( K \) heterogeneous agents that may have different computational powers. Each agent \( k \) has locally \( N_k \) data points, which we denote by \( \{x_{k,n}\} \); the subscript \( k \) refers to the agent, while the subscript \( n \) denotes the sample index within agent \( k \)’s dataset. The goal of the agents is to find an optimizer for the aggregate risk function:

\[
\hat{w}^o \triangleq \arg\min_{w \in \mathbb{R}^M} \frac{1}{K} \sum_{k=1}^{K} P_k(w),
\]

where each \( P_k(\cdot) \) is an empirical risk defined in terms of a loss function \( Q_k(\cdot) \):

\[
P_k(w) \triangleq \frac{1}{N_k} \sum_{n=1}^{N_k} Q_k(w; x_{k,n}).
\]

Multiple strategies exist for solving such problems. They can be categorized into two main classes: a) partially decentralized strategies, which include a central process with access to all data and which controls the distribution of the data into the nodes for processing \([2]–[4]\); and b) fully decentralized strategies, which consist of multiple agents connected by a graph topology and operating locally without oversight by a central processor \([5]–[8]\). Federated learning \([9]–[21]\) offers a mid-term solution, which consists of several agents collecting and processing local data that are then aggregated at the central processor.

When implementing SGD, most strategies choose the samples according to a uniform distribution. In this work, we shall consider more general non-uniform sampling schemes, where the agents are sampled according to some distribution \( \pi_k \) and the local data at agent \( k \) are in turn sampled according to some other distribution \( \pi_n^{(k)} \). In this setting, the central processor selects the subset of agents for processing according to \( \pi_k \) and, once selected, an agent \( k \) will sample its data according to \( \pi_n^{(k)} \). The importance sampling process in this article therefore involves two layers. We use the superscript \( (k) \) to denote the sampling distribution of the data at agent \( k \). The sampling distributions \( \{\pi_k, \pi_n^{(k)}\} \) are not fixed; instead we will show how to adapt them in order to enhance performance. At the same time, we will provide a detailed convergence analysis and establish performance limits.

A. Related Work

Several works studied the convergence of the federated learning algorithm or distributed SGD under differing assumptions. These assumptions usually relate to the nature of the data (IID or non-IID), nature of the cost function (convex or non-convex), agent participation (full or partial), and operation (synchronous or asynchronous) \([10], [11], [22]–[32]\). Other works examine the convergence behavior of variations of the traditional FedAvg algorithm, such as FedProx \([10]\), hierarchical version of FedAvg \([32]\), multi-task federated learning \([11]\), and dynamic FedAvg \([11]\) – see Table 1.

By contrast, not much work has been done on selection schemes for agents and data in federated learning. Given the architecture of a federated learning solution, this is a natural and important question to consider. The existing works in this domain can be split into two categories: those seeking better accuracy, and those seeking fairness. Of the works pertaining to the first category, reference \([33]\) develops a new client selection scheme, called FedCS, where the goal of the central server is to choose as many agents as possible that can complete an iteration by a required deadline, after acquiring information about the agents’ resources. Reference \([34]\) builds on this previous work to deal with non-IID data, and allows the server to collect some of the data from the agents and participate in the training of the model. The authors of \([35]\) consider non-uniform sampling of agents and suggest approximate sampling probabilities that maximize the average inner product of the local gradient with the global gradient. References \([36], [37]\) fall under the second category; in agnostic federated learning \([36]\), the data distribution is assumed to be a mixture of the local distributions, and a minimax problem for agent selection is solved. Reference \([37]\) generalizes the previous work by
reweighting the cost function and assigning higher weights to agents with higher loss.

While there exist works that study the effect of importance sampling in distributed learning [38]-[42], all of these works apply importance sampling to the data at each agent. To our knowledge, there are no works that examine the combined effect of two hierarchical layers of sampling: one for the nodes and another for their data. By introducing a two-layer importance sampling scheme to the federated learning paradigm, we can tackle the problem of importance sampling both in relation to agents and also in relation to data.

B. Sampling and Inclusion Probabilities

Before describing the problem setting, we need to clarify the difference between two notions: (a) sampling probability and (b) inclusion probability. Consider the following illustrative example. Consider $N = 4$ balls of which we wish to choose $B = 2$ balls non-uniformly and without replacement. Let the sampling probabilities be $\pi_n = \{1/3, 1/6, 1/3, 1/6\}$. This means that, initially, balls 1 and 3 are twice as likely to be selected compared to balls 2 and 4. For the first trial, all the inclusion probabilities are equal to the sampling probabilities, i.e., $\mathbb{P}(n \text{ chosen on } 1^{st} \text{ trial}) = \pi_n$. However, since we are sampling without replacement, the inclusion probabilities for the second trial depend on the outcome of the first trial, i.e., $\mathbb{P}(n \text{ chosen on } 2^{nd} \text{ trial}|n \text{ chosen on } 1^{st} \text{ trial}) = \pi_n/(1 - \pi_n)$. Using the sampling probabilities, we can evaluate the likelihood that each ball will end up belonging to the selected set of 2 balls. In particular, the probability that ball 1 is chosen either in the first or second trial is given by:

$$
\mathbb{P}(1 \text{ chosen}) = \sum_{n=2}^{4} \mathbb{P}(1 \text{ chosen on } 1^{st} \text{ trial} \& n \text{ chosen on } 2^{nd} \text{ trial}) + \mathbb{P}(n \text{ chosen on } 1^{st} \text{ trial} \& 1 \text{ chosen on } 2^{nd} \text{ trial})
= \sum_{n=2}^{4} \pi_1 \pi_n/(1 - \pi_1) + \pi_n \pi_1/(1 - \pi_n).
$$

Thus, the sampling probability is the working probability. It is the probability used to actually choose the samples, while the inclusion probability is a descriptive probability that indicates the likelihood of a ball being included in the final selected subset. Observe that the inclusion probabilities depend on the sampling scheme, while the sampling probabilities do not.

Next, we consider a total number of $K$ agents. At each iteration $i$ of the algorithm, a subset of agents $\mathcal{L}_i$ of size $L$ is chosen randomly without replacement. We denote the probability that agent $k$ is included in the sample by $L_p_k$, i.e.,

$$
p_k = \frac{\mathbb{P}(k \in \mathcal{L}_i)}{L}. \tag{4}
$$

In addition, each sampled agent $k$ will run a mini-batch SGD by sampling $B_k$ data points $B_{k,i}$ without replacement from its local data. We denote the probability of inclusion of data point $n$ by $B_k p_n^{(k)}$, i.e.,

$$
p_n^{(k)} = \frac{\mathbb{P}(n \in B_{k,i})}{B_k}. \tag{5}
$$

We refer to $p_k$ and $p_n^{(k)}$ as the normalized inclusion probabilities. They sum to 1 over the sampling space; $p_k$ sums to 1 over all agents and $p_n^{(k)}$ over the data at each agent.

II. ALGORITHM DERIVATION

The goal of the federated learning algorithm is to approximate the centralized solution $w^*$ while dealing with the constraint of distributed data. The goal is achieved by using an unbiased estimate of the gradient of the cost function, $\frac{1}{K} \sum_{k=1}^{K} \nabla w \mathbb{E}_k P_k(w)$. As explained in [11] for the case of uniform sampling, if we assume each agent $k$ runs $E_k$ epochs
per iteration $i$ (with each epoch using $B_k$ samples in $B_{k,i,e}$), then we can construct an unbiased estimate for the true gradient by considering the following estimator:

$$\frac{1}{L} \sum_{k \in \mathcal{L}_i} \frac{1}{E_k B_k} \sum_{e=1}^{E_k} \sum_{b \in B_{k,i,e}} \nabla_{w^T} Q_k(w; x_{k,b}),$$

as opposed to the original estimator from [9], where the main difference is the scaling by the epoch size $E_k$. This correction is important for the performance of the averaged model. Since the number of epochs $E_k$ can be non-uniform across the agents, then, without correction, agents with large epoch sizes will bias the solution by driving it towards their local model and away from $w^o$.

Expression (6) is still not sufficient for our purposes in this article, since agents and data are allowed to be sampled non-uniformly without replacement. In this case, we need to adjust (6) by including the inclusion probabilities [38]. The inclusion probabilities are necessary to ensure the estimate is unbiased, as will later be seen in Lemma 1. The local estimate of the gradient at agent $k$ becomes $\frac{1}{K_p} \nabla_{w^T} P_k(w)$, with:

$$\nabla_{w^T} P_k(w) \triangleq \frac{1}{E_k B_k} \sum_{e=1}^{E_k} \sum_{b \in B_{k,i,e}} \frac{1}{N_k p_b} \nabla_{w^T} Q_k(w; x_{k,b}).$$

Motivated by (7), we can write down a stochastic gradient update at each agent $k$ at epoch $e$, and at the central processor at iteration $i$:

$$w_{k,e} = w_{k,e-1} - \frac{\mu}{K_p E_k B_k} \sum_{b \in B_{k,i,e}} \frac{1}{N_k p_b} \nabla_{w^T} Q_k(w_{k,e-1}; x_{k,b}),$$

$$w_i = \frac{1}{L} \sum_{k \in \mathcal{L}_i} w_{k,E_k},$$

where at each iteration $i$, step (8) is repeated for $e = 1, 2, \cdots, E_k$. We arrive at the Algorithm 1, which we refer to as Importance Sampling Federated Averaging (ISFedAvg).

### III. Convergence Analysis

#### A. Modeling Conditions

To facilitate the analysis of the algorithm, we list some common assumptions on the nature of the local risk functions and their respective minimizers. Specifically, we assume convex cost functions with smooth gradients.

**Assumption 1.** The functions $P_k(\cdot)$ are $\nu-$strongly convex, and $Q_k(\cdot; x_{k,n})$ are convex, namely:

$$P_k(w_2) \geq P_k(w_1) + \nabla_{w^T} P_k(w_1)(w_2 - w_1) + \frac{\nu}{2} \|w_2 - w_1\|^2,$$

$$Q_k(w_2; x_{k,n}) \geq Q_k(w_1; x_{k,n}) + \nabla_{w^T} Q_k(w_1; x_{k,n})(w_2 - w_1).$$

**Also, the functions $Q_k(\cdot; x_{k,n})$ have $\delta-$Lipschitz gradients:**

$$\|\nabla_{w^T} Q_k(w_2; x_{k,n}) - \nabla_{w^T} Q_k(w_1; x_{k,n})\| \leq \delta \|w_2 - w_1\|.$$

#### B. Error Recursion

Iterating the local update (8) over multiple epochs and combining according to (9), we obtain the following update for the central iterate:

$$w_i = w_{i-1} - \frac{\mu}{L} \sum_{k \in \mathcal{L}_i} \sum_{e=1}^{E_k} \sum_{b \in B_{k,i,e}} \frac{1}{N_k p_b} \nabla_{w^T} Q_k(w_{k,e-1}; x_{k,b}).$$

To simplify the notation, we introduce the error terms:

$$s_i \triangleq \frac{1}{L} \sum_{l \in \mathcal{L}_i} \sum_{e=1}^{E_l} \sum_{b \in B_{l,i,e}} \frac{1}{N_l p_b} \nabla_{w^T} Q_k(w_{l,e-1}; x_{l,b}) - \nabla_{w^T} Q_k(w_{l-1}; x_{l,b}).$$

The first error term $s_i$, which we call gradient error, captures the error from approximating the true gradient by using subsets of agents and data; while, the second error term $q_i$, which we call incremental error, captures the error resulting from the incremental implementation, where at each epoch during one iteration, the gradient is calculated at the local iterate $w_{k,e-1}$. Note that this second error evaluates the loss function $w_{k,e-1}$.
at the local and global iterates. As we will show later, the incremental error will fade away, and the dominant factor will be the gradient error. Before establishing the main result in Theorem 1 on the convergence of ISFedAvg algorithm, we present preliminary results that will lead to it. Thus, to show the convergence of the algorithm, we must assure the gradient noise \(s_i\) has zero mean and bounded variance, and the incremental noise \(q_i\) has bounded variance. Furthermore, since we split the noise due to the stochastic gradient into incremental and gradient noise, we can split the analysis into that of the centralized steps and the local epochs. By proving that both the centralized and local steps converge, we show the global algorithm converges too.

Replacing the two error terms (14) and (15) into recursion (13) and subtracting \(w^o\) from both sides of the equation, we get the following error recursion:

\[
\tilde{w}_1 = \tilde{w}_{i-1} + \mu \frac{1}{K} \sum_{k=1}^{K} \nabla w^T P_k(w_{i-1}) + \mu s_i + \mu q_i.
\]  

To bound the \(\ell_2\)-norm of the error, we split it into two terms, centralized and incremental, using Jensen’s inequality with some constant \(\alpha \in (0, 1)\) to be defined later:

\[
\|\tilde{w}_i\|^2 \leq \frac{1}{\alpha} \left( \|\tilde{w}_{i-1}\|^2 + \mu \frac{1}{K} \sum_{k=1}^{K} \nabla w^T P_k(w_{i-1}) + \mu s_i \right) + \frac{1}{1-\alpha} \mu^2 \|q_i\|^2.
\]  

We start with the first term that represents the centralized solution. We need to show that it converges. To do so, we start with the gradient noise, and we establish in Lemma 1 that it remains bounded. We bound the gradient noise \(s_i\) under two constructions: sampling with replacement, and sampling without replacement.

**Lemma 1 (Estimation of first and second order moments of the gradient noise).** The gradient noise defined in (14) has zero mean:

\[
\mathbb{E}\{s_i|w_{i-1}\} = 0,
\]

with bounded variance, regardless of the sampling scheme. More specifically, sampling agents and data with replacement, results in the following bound:

\[
\mathbb{E}\{\|s_i\|^2|w_{i-1}\} \leq \beta^2 s\|\tilde{w}_{i-1}\|^2 + \sigma^2 s,
\]

where \(\tilde{w}_{i-1} = w^o - w_{i-1}\) and the constants:

\[
\beta^2 s = \frac{3\delta^2}{L} + \frac{1}{LK} \sum_{k=1}^{K} \frac{1}{p_k} \left( \beta^2 s_k + 3\delta^2 \right),
\]

\[
\sigma^2 s = \frac{1}{LK} \sum_{k=1}^{K} \frac{1}{p_k} \left( \frac{\sigma^2}{p_k} + \left(3 + \frac{6}{E_kB_k}\right) \left\|\nabla w^T P_k(w^o)\right\|^2 \right),
\]

\[
\beta^2 s_{s,k} = \frac{3\delta^2}{E_kB_k} \left(1 + \frac{1}{N_k} \sum_{n=1}^{N_k} \frac{1}{p_n(k)}\right),
\]

\[
\sigma^2 s_{s,k} = \frac{6}{E_kB_kN_k} \sum_{n=1}^{N_k} \frac{1}{p_n(k)} \left\|\nabla w^T Q_k(w^o; x_{k,n})\right\|^2.
\]

On the other hand, sampling agents and data without replacement results in the same bound but without the scaling by \(L\) in the constants \(\beta^2 s\) and \(\sigma^2 s\).

**Proof.** See Appendix B.

The term \(\sigma^2 s_{s,k}\) in the bound captures what we call data variability. It is controlled by the mini-batch size \(B_k\); as the mini-batch increases the effect of this term is reduced. The \(\|\nabla w^T P_k(w^o)\|\) term quantifies the suboptimality of the global model locally; we call its effect model variability. It is reduced when the data and agents are more heterogeneous. From Assumption 2 we can bound it uniformly.

Now that we have identified the mean and variance of the gradient noise, we can proceed to establish the important conclusion that the following centralized solution:

\[
w_i = w_{i-1} - \mu \frac{1}{L} \sum_{\ell \in E_i} \nabla w^T P_i(w_{i-1}),
\]

converges exponentially to an \(O(\mu)\)-neighbourhood of the optimizer. In this implementation, the center processor aggregates the approximate gradients of the selected agents. We will subsequently call upon this result to examine the convergence behavior of the proposed federated learning solution.

**Lemma 2 (Mean-square-error convergence of the centralized solution).** Consider the centralized recursion (24), where the cost functions satisfy Assumption 7 and where the first and second order moments of the gradient noise process satisfy the conditions in Lemma 1. Also, the samples are chosen without replacement. For step-size values satisfying \(\mu < 2\nu/(\delta^2 + \beta^2 s)\), it holds that \(\mathbb{E}\|\tilde{w}_i\|^2\) converges exponentially fast according to the recursion (25), where \(\lambda = 1 - 2\mu\nu + \mu^2(\delta^2 + \beta^2 s) \in [0, 1)\).

\[
\mathbb{E}\|\tilde{w}_i\|^2 \leq \lambda \mathbb{E}\|\tilde{w}_{i-1}\|^2 + \mu^2 \sigma^2 s
\]  

(25)

It follows from (25) that, for sufficiently small step-sizes:

\[
\mathbb{E}\|\tilde{w}_i\|^2 \leq \lambda^i \mathbb{E}\|\tilde{w}_0\|^2 + \frac{1 - \lambda^i}{1 - \lambda}\mu^2 \sigma^2 s.
\]  

(26)

**Proof.** see Appendix C.

We next bound the incremental noise \(q_i\). To do so, we introduce the local terms:

\[
q_{k,i,e} \triangleq \frac{1}{Kp_k} \left( \frac{1}{B_k} \sum_{b \in B_{k,i,e}} \frac{1}{N_kp_b(k)} \nabla w^T Q_k(w_{k,e-1}; x_{k,b}) \right) - \nabla w^T P_k(w_{k,e-1}).
\]  

(27)

We show in the next lemma that the local gradient noise \(q_{k,i,e}\) has zero mean and bounded variance. This result is useful for showing that the local SGD steps converge in the mean-square-error sense towards their local models \(w^o_k\).

**Lemma 3 (Estimation of first and second order moments of the local gradient noise).** The local gradient noise defined in (27) has zero mean:

\[
\mathbb{E}\{q_{k,i,e}|\mathcal{F}_{e-1}, \mathcal{L}_i\} = 0,
\]

(28)
and bounded variance, regardless of the sampling scheme:

\[ \mathbb{E}\left\{ \left| q_{k,i,e} \right|^2 \right| \mathcal{F}_{e-1}, \mathcal{C}_i \right\} \leq \frac{E_k}{K^2 p_k} \beta_s, k \left| \mathbf{w}_{k,e-1} \right|^2 + \frac{1}{K^2 p_k} \sigma^2_{q,k}, \]

(29)

where \( \mathcal{F}_{e-1} = \{ w_{k,0}, w_{k,1}, \ldots, w_{k,e-1} \} \) is the filtration describing all sources of randomness due to the previous iterates, \( \mathbf{w}_{k,e} = \mathbf{w}_0 - \mathbf{w}_{k,e} \), and the constants are as defined in (22) and:

\[ \sigma^2_{q,k} = \frac{3}{B_k K^2} \sum_{n=1}^{N_k} \frac{1}{P_n} \left\| \nabla w^T Q(w^e_k; x_{k,n}) \right\|^2. \]

(30)

Proof. see Appendix D.

Now that we have showed that the local gradient noise of the incremental step has bounded variance, we can study the mean square deviation of the local SGD.

**Lemma 4 (Mean-square-error convergence of the local incremental step).** For every agent \( k \), consider the local stochastic gradient recursion (3) where the cost function is subject to Assumption 2 and where the first and second order moments of the gradient noise process satisfy the conditions in Lemma 3. For step-size values satisfying:

\[ \mu < \frac{2\nu}{\sigma^2 + \frac{E_k}{K^2 p_k} \beta^2_s, k}, \]

(31)

it holds that \( \mathbb{E}\left\{ \left| \mathbf{w}_{k,e} \right|^2 \right| \mathcal{F}_{e-1} \right\} \) converges exponentially fast according to the recursion:

\[ \mathbb{E}\left\{ \left| \mathbf{w}_{k,e} \right|^2 \right| \mathcal{F}_{e-1} \right\} \leq \lambda_k \mathbb{E}\left\{ \left| \mathbf{w}_{k,e-1} \right|^2 \right| \mathcal{F}_{e-1} \right\} + \mu^2 \sigma^2_{q,k}, \]

(32)

where:

\[ \lambda_k = 1 - 2\nu \mu + \mu^2 \left( \sigma^2 + \frac{E_k}{K^2 p_k} \beta^2_s, k \right) \in [0, 1). \]

(33)

It follows from (32) that, for sufficiently small step-sizes:

\[ \mathbb{E}\left\{ \left| \mathbf{w}_{k,e} \right|^2 \right| \mathcal{F}_{e-1} \right\} \leq \lambda_k^e \mathbb{E}\left\{ \left| \mathbf{w}_{k,0} \right|^2 \right| \mathcal{F}_{e-1} \right\} + \frac{1 - \lambda_k^e}{1 - \lambda_k} \mu^2 \sigma^2_{q,k}. \]

(34)

Proof. see Appendix E.

We can finally bound the incremental noise in the following lemma.

**Lemma 5 (Estimation of the second order moments of the incremental noise).** The incremental noise defined in (13) has bounded variance:

\[ \mathbb{E}\left\{ \left| q_i \right|^2 \right\} \leq O(\mu) \mathbb{E}\left\{ \left| \mathbf{w}_{t-1} \right|^2 \right\} + O(\mu) \xi^2 + O(\mu^2) \frac{1}{K} \sum_{k=1}^{K} \sigma^2_{q,k}, \]

(35)

where the \( O(\cdot) \) terms depend on epoch sizes, local convergence rates, total number of data samples, number of agents, Lipschitz constant, and data and agent normalized inclusion probabilities. Thus, \( \mathbb{E}\left\{ \left| q_i \right|^2 \right\} = O(\mu) \).

Proof. see Appendix F.

We observe an average data variability term across agents \( \sigma^2_{q,k} \), and a model variability term \( \xi^2 \). However, the effect of the latter dominates since it is multiplied by an \( O(\mu) \) term as opposed to \( O(\mu^2) \).

### C. Main Theorem

Now that we have bounded each term of (17) and using Lemma 2 and 5 we find:

\[ \mathbb{E}\left\{ \left| \mathbf{w}_{\mathcal{I}} \right|^2 \right\} \]

\[ \leq \frac{1}{\alpha} \left( \frac{\lambda}{\mu} \mathbb{E}\left\{ \left| \mathbf{w}_{t-1} \right|^2 \right\} + \mu^2 \sigma^2_s \right) + O(\mu^3) \frac{1}{1 - \alpha} \mathbb{E}\left\{ \left| \mathbf{w}_{t-1} \right|^2 \right\} + \xi^2 \]

\[ + O(\mu^4) \frac{1}{1 - \alpha} K \sum_{k=1}^{K} \sigma^2_{q,k}, \]

(36)

where:

\[ \lambda' = \frac{\lambda}{\alpha} + O(\mu^3) \frac{1}{1 - \alpha}, \]

\[ = \frac{1 - 2\mu \nu + \mu^2 (\sigma^2 + \beta^2_s, k)}{\alpha} + O(\mu^3) \frac{1}{1 - \alpha}, \]

\[ = O\left( \frac{1}{\alpha} \right) + O\left( \frac{\mu}{\alpha} \right) + O\left( \frac{\mu^3}{\alpha} \right) + O\left( \frac{\mu^4}{1 - \alpha} \right). \]

(37)

 Applying this bound recursively we obtain:

\[ \mathbb{E}\left\{ \left| \mathbf{w}_{\mathcal{I}} \right|^2 \right\} \leq \left( \lambda' \right)^i \mathbb{E}\left\{ \left| \mathbf{w}_{0} \right|^2 \right\} + \frac{1 - \left( \lambda' \right)^i}{1 - \lambda'} \left( \frac{\mu^2 \sigma^2_s}{\alpha} + O(\mu^3) \frac{1}{1 - \alpha} \xi^2 \right) + O(\mu^4) \frac{1}{1 - \alpha} K \sum_{k=1}^{K} \sigma^2_{q,k}, \]

(38)

and then taking the limit \( i \to \infty \):

\[ \lim_{i \to \infty} \mathbb{E}\left\{ \left| \mathbf{w}_{\mathcal{I}} \right|^2 \right\} \]

\[ \leq \frac{1}{1 - \lambda'} \left( \frac{\mu^2 \sigma^2_s}{\alpha} + O(\mu^3) \frac{1}{1 - \alpha} \xi^2 + O(\mu^4) \frac{1}{1 - \alpha} K \sum_{k=1}^{K} \sigma^2_{q,k} \right). \]

(39)

for \( \lambda' < 1 \), which for \( \alpha = \sqrt{\lambda} \) is achieved when:

\[ \mu < \min \left\{ \frac{2\nu}{\sigma^2 + \frac{E_k}{K^2 p_k} \beta^2_s, k}, \frac{2\nu}{\sigma^2 + \frac{E_k}{K^2 p_k} \beta^2_s, k} \right\}, \]

(40)

\[ O(\mu^3) < (1 - \sqrt{\lambda})^2. \]

(41)

Thus, since \( \alpha = O(1), 1 - \alpha = O(\mu) \) and \( 1 - \lambda' = O(\mu) \):

\[ \lim_{i \to \infty} \mathbb{E}\left\{ \left| \mathbf{w}_{\mathcal{I}} \right|^2 \right\} \leq O(\mu \sigma^2_s + O(\mu) \xi^2 + O(\mu^2) \frac{1}{K} \sum_{k=1}^{K} \sigma^2_{q,k}). \]

(42)

The result is summarized in the theorem.

### Theorem 1 (Mean-square-error convergence of federated learning under importance sampling).

Consider the iterates \( \mathbf{w}_t \) generated by the importance sampling federated averaging algorithm. For sufficiently small step-size \( \mu \), it holds that the mean-square-error converges exponentially fast:

\[ \mathbb{E}\left\{ \left| \mathbf{w}_{\mathcal{I}} \right|^2 \right\} \leq O \left( (\lambda')^i \right) + O(\mu) \left( \sigma^2_s + \xi^2 \right) + O(\mu^2) \frac{1}{K} \sum_{k=1}^{K} \sigma^2_{q,k}, \]

(43)

where \( \lambda' = 1 - O(\mu) + O(\mu^2) \in [0, 1) \).
IV. IMPORTANCE SAMPLING

Due to the heterogeneity of nodes which arise from their data and computational capabilities, it is important to guide the algorithm based on the potential contribution that each agent can have on the overall performance. By allowing asynchronicity, i.e., different epoch sizes among agents, we can take advantage of the varying computational capabilities. From [38], we know that the choice of samples at each iteration affects the solution. Therefore, instead of choosing the samples uniformly, we consider importance sampling where samples are chosen according to some distribution to be determined. A similar scheme can be enforced on the participating agents.

In what follows, we show that using importance sampling enhances the overall performance.

A. Agent Level: Importance Sampling of Data

Every agent $k$ at each epoch must select a mini-batch of data based on the normalized inclusion probabilities $p_n^{(k)}$. To find the optimal probabilities, we minimize the bound on the variance of the local gradient noise $\sigma^2_{s,k}$. We solve the problem for both cases when sampling is done with replacement and without replacement. The results are the same for both sampling schemes.

**Lemma 6 (Optimal local data inclusion probabilities).** The optimal local data normalized inclusion probabilities are given by:

$$p_n^{(k), o} \triangleq \frac{\|\nabla_w^T Q_k(w^{o}; x_{k,n})\|}{\sum_{m=1}^{N_k} \|\nabla_w^T Q_k(w^{o}; x_{k,m})\|}.$$  

(44)

**Proof.** By introducing a Lagrange multiplier $\lambda$, the optimization problem can be reformulated as:

$$\min_{p_n^{(k), o}, \lambda} \sum_{k=1}^{N_k} p_n^{(k)} \frac{\|\nabla_w^T Q_k(w^{o}; x_{k,n})\|^2}{\lambda} + \lambda \left( \sum_{n=1}^{N_k} p_n^{(k)} - 1 \right).$$

(45)

Then, taking the derivative with respect to $p_n^{(k), o}$ and setting it to zero we get (46). Next, substituting $p_n^{(k), o}$ into the condition, we find (47).

$$p_n^{(k), o} = \frac{\|\nabla_w^T Q_k(w^{o}; x_{k,n})\|}{\sqrt{\lambda}}$$  

(46)

$$\sqrt{\lambda} = \sum_{m=1}^{N_k} \|\nabla_w^T Q_k(w^{o}; x_{k,m})\|$$  

(47)

As seen in Lemma [6], more weight is given to a data point that has a greater gradient norm, thus increasing its chances of being sampled and resulting in a faster convergence rate. In addition, we observe that the more homogeneous the data is the more uniform the inclusion probability is.

B. Cloud Level: Importance Sampling of Agents

At each iteration, the cloud must select a subset of agents to participate. The agents are selected in accordance with the normalized inclusion probabilities $p_k$. To find the optimal probabilities, we minimize the bound on the variance of the gradient noise $\sigma^2_s$. The following result holds for sampling with and without replacement, since the gradient noise only differ by a multiplicative factor.

**Lemma 7 (Optimal agent inclusion probabilities for sampling with replacement).** The optimal agent normalized inclusion probabilities are given by (48), where $\alpha_k = \left(3 + \frac{6}{E_k}B_k\right)$:

$$p_k^o \triangleq \frac{\sqrt{\sigma^2_{s,k} + \alpha_k \|\nabla_w^T P_k(w^{o})\|^2}}{\sum_{\ell=1}^{K} \sqrt{\sigma^2_{s,\ell} + \alpha_\ell \|\nabla_w^T P_\ell(w^{o})\|^2}}.$$  

(48)

**Proof.** The proof follows similarly to that of Lemma [6].

We observe that the normalized inclusion probabilities will be closer to a uniform distribution the more the data and model variability terms are similar across agents.

C. Practical Issues

In the previous subsections, we focused on finding the optimal inclusion probabilities for the agents and data. However, several practical issues arise. The first is that all probabilities are calculated based on the optimal model $w^{o}$, which we do not have access to. To overcome this issue, we estimate the probabilities at each iteration by calculating them according to the current model $w_{i-1}$. Thus,

$$\hat{p}_n^{(k), o} = \frac{\|\nabla_w^T Q_k(w_{i-1}; x_{k,n})\|}{\sum_{m=1}^{N_k} \|\nabla_w^T Q_k(w_{i-1}; x_{k,m})\|},$$

$$\hat{p}_k^o = \frac{\sqrt{\sigma^2_{s,k} + \alpha_k \|\nabla_w^T P_k(w_{i-1})\|^2}}{\sum_{\ell=1}^{K} \sqrt{\sigma^2_{s,\ell} + \alpha_\ell \|\nabla_w^T P_\ell(w_{i-1})\|^2}}.$$  

(49)

In addition, since calculating the true gradient of the local loss function is costly, we replace it with the mini-batch approximation when calculating $\hat{p}_k^o$:

$$\hat{p}_k^o = \frac{\sqrt{\sigma^2_{s,k} + \alpha_k \|\nabla_w^T \hat{P}_k(w_{i-1})\|^2}}{\sum_{\ell=1}^{K} \sqrt{\sigma^2_{s,\ell} + \alpha_\ell \|\nabla_w^T \hat{P}_\ell(w_{i-1})\|^2}}.$$  

(50)

Furthermore, every agent has access to all of its data and consequently to all of the gradients. However, the cloud does not have access to the gradients of all agents, and in turn cannot calculate the denominator of $p_k$. Instead, we propose the following solution: at iteration 0, all probabilities are set to $p_k = \frac{1}{K}$; then, during the $i^{th}$ iteration, after the participating agents $\ell \in L_i$ send the cloud their stochastic gradients $\nabla_w^T \hat{P}_i(w_{i-1})$, the probabilities are updated as follows:

$$\hat{p}_k^o = \frac{\sqrt{\sigma^2_{s,k} + \alpha_k \|\nabla_w^T \hat{P}_k(w_{i-1})\|^2}}{\sum_{\ell \in L_i} \sqrt{\sigma^2_{s,\ell} + \alpha_\ell \|\nabla_w^T \hat{P}_\ell(w_{i-1})\|^2}} \left(1 - \sum_{\ell \in L_i} \hat{p}_k^o\right),$$

(51)

where the multiplicative factor follows from ensuring all the probabilities $\hat{p}_k^o$ sum to 1. Similarly for the local probabilities,
since we are implementing mini-batch SGD, we only update the
probabilities of the data points that were sampled:
\[ \tilde{p}^{(k),o}_n = \frac{\| \nabla_w Q_k(w_{i-1}; x_{k,n}) \|^2}{\sum_{b \in B_{k,i},e} \| \nabla_w Q_k(w_{i-1}; x_b) \|^2} \left( 1 - \sum_{b \in B_{k,i,e}} \tilde{p}^{(k),o}_b \right) \]  \hspace{1cm} (53)

Finally, the last problem arises when sampling without replacement. We have found the optimal inclusion probabilities
and not the optimal sampling probabilities, and moving from the
former to the latter is not trivial. Thus, we rely on the
literature under sampling without replacement with unequal
probabilities. Multiple sampling schemes exist such that the
sampling probabilities do not need to be calculated explicitly.
In general, there are multiple non-uniform sampling with-
out probabilities. Multiple sampling schemes exist such that the
inclusion probabilities \( \Pi_k = \sum_{k=1}^{K_0} \Pi_{k,n} \) for the agents and data, respectively. More
explicitly, we first calculate the \( \text{progressive totals of the} \) inclusion probabilities \( \Pi_k = \sum_{k=1}^{K} \Pi_{k,n} \), for \( k = 1, 2, \ldots, K \), and we set \( \Pi_0 = 0 \). Then, we select uniformly at random a
\( \text{uniform variate} \) \( d \in [0, 1] \). Then, we select the \( L \) agents that
satisfy \( \Pi_{k-1} \leq q + \ell < \Pi_k \), for some \( \ell = 0, 1, \ldots, L - 1 \).

V. EXPERIMENTAL SECTION

To validate the theoretical results, we devise two experi-
ments. The first consists of simulated data with quadratic risk
functions, and the second consists of a real dataset with logistic
risk functions.

A. Regression

We first validate the theory on a regression problem. We
consider \( K = 300 \) agents, for which we generate \( N_k = 100 \)
data points for each agent \( k \) as follow: Let \( u_{k,n} \) denote
an independent streaming sequence of two-dimensional random
vectors with zero mean and covariance matrix \( R_{u_k} = \mathbb{E} u_{k,n} u_{k,n}^T \). Let \( d_k(n) \) denote a streaming sequence of random
variables that have zero mean and variance \( \sigma^2_{d_k} = \mathbb{E} d_k^2(n) \). Let
\( r_{d_k,u_k} = \mathbb{E} d_k(n) u_{k,n} \) be the cross-variance vector. The data \( \{d_k(n), u_{k,n}\} \) are related by the following linear regression model:
\[ d_k(n) = u_{k,n} w^* + v_k(n), \]  \hspace{1cm} (54)
for some randomly generated parameter vector \( w^* \) and where \( v_k(n) \) is a zero mean white noise process with variance \( \sigma^2_{v_k} = \mathbb{E} v_k^2(n) \), independent of \( u_{k,n} \). The local risk is given by:
\[ P_k(w) = \frac{1}{N_k} \sum_{n=1}^{N_k} \| d_k(n) - u_{k,n} w^* \|^2 + \rho \| w \|^2. \]  \hspace{1cm} (55)
We set \( \rho = 0.001 \), while the batch sizes \( B_k \) and the epoch
sizes \( E_k \) are chosen uniformly at random from the range \([1, 10]\)
and \([1, 5]\), respectively. During each iteration, there are \( L = 6 \)
active agents. To test the performance of the algorithm, we
calculate at each iteration the mean-square-deviation (MSD)
of the parameter vector \( w \), with respect to the true model \( w^* \):
\[ \text{MSD}_i = \| w_i - w^* \|^2. \]  \hspace{1cm} (56)

The optimization problem has the closed form expression:
\[ w^* = (\tilde{R}_u + \rho I)^{-1} \tilde{R}_u w^* + (\tilde{R}_u + \rho I)^{-1} \tilde{r}_{uv}, \]  \hspace{1cm} (57)
where:
\[ \tilde{R}_u = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N_k} \sum_{n=1}^{N_k} u_{k,n}^T u_{k,n}, \]  \hspace{1cm} (58)
\[ \tilde{r}_{uv} = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N_k} \sum_{n=1}^{N_k} v_k(n) u_{k,n}. \]  \hspace{1cm} (59)

We run four tests: we first run the standard FedAvg algo-
rithm where the mini-batches are chosen uniformly with
replacement. We then run Algorithm [1] once with the
optimal probabilities \( p_n^{(k)} \) and \( p_k \) in (44)–(48), once with the
approximate probabilities (49)–(50), and once with (52)–(53).
We implement the sampling scheme from (44). We set the
step-size \( \mu = 0.01 \). Each test is repeated 100 times, and the
resulting MSD is averaged. We get the curves as shown in
Figure 1. We see that the importance sampling scheme does
better than the standard sampling algorithm. This comes as no
surprise, since the probabilities were chosen to minimize the
bound on the MSD. The importance sampling scheme (green
curve) improved the MSD bound by 23.1 dB compared to the
standard federated learning scheme (blue curve). Furthermore,
we observe that approximate probabilities do not degrade
the performance of the algorithm. Our proposed approximate
solution (52)–(53) (purple curve) performs just as well as using
the true probabilities. In fact, we observe that the approximate
probabilities converge to the true ones, \( \| p_n^{(k)} - p_k \| = 1.22 e - 2 \)
and \( \frac{1}{K} \sum_{k=1}^{K} \| p_n^{(k),o} - p_n^{(k),o} \| = 1.54 e - 2 \). Similarly, the
approximate probabilities (50)–(53) do not degrade the overall
performance. They, in fact, outperform the other solutions and
converge faster (red curve). This is not surprising, since, at
each iteration, we are attributing higher probabilities to agents
and data points that have greater gradients. We are increasing
their chances of being selected and thus taking steeper steps
towards the true model.

Fig. 1. MSD plots of the regression problem: blue curve is the standard mini-
batch implementation, green curve is the importance sampling implementation
with the true probabilities, red curve is the importance sampling implementa-
tion with approximate probabilities (49)–(50), purple curve is the importance
sampling implementation with approximate probabilities (52)–(53).
B. Classification

We next study the theory in a classification context. We consider the ijcnn1 dataset \[45\]. The dataset consists of 35000 training samples and 91701 testing samples of \( M = 22 \) attributes. We distribute the data randomly in a non-IID fashion to \( K = 100 \) agents. Each agent receives a random number \( N_k \) of data points, where \( N_k \) ranges from 79 to 688. We run the two algorithms FedAvg and ISFedAvg. We set \( \mu = 0.25, \rho = 0.0001, L = 10, B_k = 1, \) and \( E_k = 1. \) We plot the testing error in Figure 2. We observe that importance sampling improves the testing error from 22.45\% to 18.46\%. This is because importance sampling is more sample efficient.

\[ \text{Fig. 2. Testing error plots of the classification problem.} \]

VI. CONCLUSION

This work incorporates two levels of importance sampling into the operation of federated learning: one for selecting agents and another for selecting data batches at the agents. Optimal dynamical choices for the sampling probabilities are derived, and a detailed convergence analysis is performed. We also provided approximate expressions for the optimal sampling policies and illustrate the theoretical findings and the performance enhancement by means of simulations.

APPENDIX A

RESULT ON THE VARIANCE OF THE MINI-BATCH ESTIMATE

We introduce the following auxiliary result that is a generalization of \[46\]. Let \( \{ S = x_n \in \mathbb{R}^M \}_{n=1}^{N} \) denote a set of \( N \) independent random variables, each with mean \( \mathbb{E}[x_n] = \bar{x} \) and variance \( \sigma_n^2 = \mathbb{E}[\|x_n - \mathbb{E}[x_n]\|^2] \). We consider the problem of estimating the expected value of the sample mean:

\[ \bar{x} \triangleq \mathbb{E}\left( \frac{1}{N} \sum_{n=1}^{N} x_n \right) \quad (60) \]

We consider two estimators for \( \bar{x} \), both constructed by considering a mini-batch of samples, where \( x_b^r \) is constructed by sampling from \( S \) with replacement, and \( x_b^{nr} \) is sampled from \( S \) without replacement. Let \( p_n \) be the normalized inclusion probability of \( x_n \). We then define the two estimators \[61\] and \[62\], and we would like to quantify the efficacy of these estimators in estimating \( \bar{x} \).

\[ \hat{x}^r \triangleq \frac{1}{B} \sum_{b=1}^{B} \frac{1}{N_p} x_b^r \quad (61) \]

\[ \hat{x}^{nr} \triangleq \frac{1}{B} \sum_{b=1}^{B} \frac{1}{N_p} x_b^{nr} \quad (62) \]

Lemma 8 (Variance of the mini-batch mean with and without replacement). Both estimators are unbiased and it holds that:

\[ \mathbb{E}[\hat{x}^r] = \mathbb{E}[\hat{x}^{nr}] = \bar{x}, \quad (63) \]

\[ \mathbb{E}[\|\hat{x}^r - \bar{x}\|^2] = \frac{1}{B} \sum_{n=1}^{N} \mathbb{E}\left( \frac{1}{N_p}^2 \sigma_n^2 + \left\| \frac{1}{N_p} \bar{x}_n - \bar{x} \right\|^2 \right), \quad (64) \]

\[ \mathbb{E}[\|\hat{x}^{nr} - \bar{x}\|^2] = \frac{1}{B} \sum_{n=1}^{N} \mathbb{E}\left( \frac{1}{N_p^2} \sigma_n^2 + \left\| \frac{1}{N_p} \bar{x}_n - \bar{x} \right\|^2 \right) \]

\[ + \frac{1}{B^2} \sum_{n_1 \neq n_2} \mathbb{P}(\bar{x}_{n_1} = \bar{x}, \bar{x}_{n_2} = \bar{x}) \left( \frac{1}{N_p} \bar{x}_{n_1} - \bar{x} \right) \left( \frac{1}{N_p} \bar{x}_{n_2} - \bar{x} \right), \quad (65) \]

where the notation \( \bar{x} \) evaluates to one if \( x_n \in B^{nr} \) and is zero otherwise.

Proof. We begin with the with-replacement setting. The randomness of the samples introduces some intricacies that need to be accounted for in the notation. For the mean, we have:

\[ \mathbb{E}[\hat{x}^r] = \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}\left( \frac{1}{N_p} \bar{x}_b^r \right) = \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}\left( \frac{1}{N_p} \bar{x}_b \mid S \right) \]

\[ = \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}\left( \sum_{n=1}^{N} \frac{1}{N_p} \bar{x}_n \right) = \frac{1}{B} \sum_{b=1}^{B} \bar{x} = \bar{x}. \quad (66) \]

For the variance we find:

\[ \mathbb{E}[\|\hat{x}^r - \bar{x}\|^2] = \mathbb{E}\left[ \left\| \frac{1}{B} \sum_{b=1}^{B} \frac{1}{N_p} x_b^r - \bar{x} \right\|^2 \right], \]

\[ = \mathbb{E}\left[ \frac{1}{B} \sum_{b=1}^{B} \left( \frac{1}{N_p} x_b^r - \bar{x} \right) \right]^2, \]

\[ = \frac{1}{B^2} \sum_{b=1}^{B} \mathbb{E}\left[ \frac{1}{N_p} \bar{x}_b - \bar{x} \right]^2 \]

\[ + \frac{1}{B^2} \sum_{b_1 \neq b_2} \mathbb{E}\left\{ \left( \frac{1}{N_p} \bar{x}_{b_1} - \bar{x} \right) \left( \frac{1}{N_p} \bar{x}_{b_2} - \bar{x} \right) \right\}, \quad (a) \]

\[ = \frac{1}{B^2} \sum_{b=1}^{B} \mathbb{E}\left[ \frac{1}{N_p} \bar{x}_b^r - \bar{x} \right]^2 \]

\[ + \frac{1}{B^2} \sum_{b_1 \neq b_2} \mathbb{E}\left\{ \left( \frac{1}{N_p} \bar{x}_{b_1} - \bar{x} \right) \mathbb{E}\left\{ \frac{1}{N_p} \bar{x}_{b_2} - \bar{x} \right\} \right\}, \quad (b) \]

\[ \leq \frac{1}{B^2} \sum_{b=1}^{B} \mathbb{E}\left[ \frac{1}{N_p} \bar{x}_b^r - \bar{x} \right]^2, \quad (67) \]
where \((a)\) is a result of the fact that the elements of \(S\) are independent and \(x_b^i\) is sampled from \(S\) independently, and hence \(x_{b_1}\) and \(x_{b_2}\) are independent. Step \((b)\) then follows from:

\[
E\left(\frac{1}{N_pB}x_b\right) = E\left(\frac{1}{N} \sum_{n=1}^{N} x_n\right) = \bar{x}.
\] (68)

Then,

\[
E\|\bar{x}^r - \bar{x}\|^2 = \frac{1}{B^2} \sum_{b=1}^{B} E\left\|\frac{1}{N_pB}x_b^r - \bar{x}\right\|^2 ,
\]

\[
= \frac{1}{B^2} \sum_{b=1}^{B} E\left\{\left\|\frac{1}{N_pB}x_b - \bar{x}\right\|^2 \mid S\right\} ,
\]

\[
= \frac{1}{B^2} \sum_{b=1}^{B} \sum_{n=1}^{N} p_n E\left\|\frac{1}{N_p} x_n - \bar{x}\right\|^2 ,
\]

\[
= \frac{1}{B^2} \sum_{b=1}^{B} \sum_{n=1}^{N} p_n E\left\{\left\|\frac{1}{N_p} x_n - \bar{x}\right\|^2 + \left\|\frac{1}{N_p} \bar{x} - \bar{x}\right\|^2 \right\} ,
\]

\[
= \frac{1}{B^2} \sum_{b=1}^{B} \sum_{n=1}^{N} p_n E\left\{\left\|\frac{1}{N_p} x_n - \bar{x}\right\|^2 + \frac{1}{N_p} \bar{x} - \bar{x}\right\} ,
\]

\[
= \frac{1}{B} \sum_{n=1}^{N} p_n \left\|\frac{1}{N_p} x_n - \bar{x}\right\|^2 ,
\]

\[
= \frac{1}{B} \sum_{n=1}^{N} p_n \left\{\frac{1}{N_p^2} \sigma_n^2 + \left\|\frac{1}{N_p} \bar{x} - \bar{x}\right\|^2 \right\} ,
\] (69)

We now proceed to study the efficiency of the without replacement mini-batch mean. The fact that the \(x_b\) are sampled from \(S\) without replacement causes pairs \(x_{b_1}, x_{b_2}\) to no longer be independent. We denote the set of points sampled from \(S\) without replacement by \(B^{ur}\) and introduce the activation function by:

\[
I_n \triangleq \begin{cases} 
1, & \text{if } x_n \in B^{ur} , \\
0, & \text{otherwise}.
\end{cases}
\] (70)

Then, the estimator \(\hat{x}^{ur}\) can be written equivalently as:

\[
\hat{x}^{ur} = \frac{1}{B} \sum_{n=1}^{N} I_n \frac{1}{N_p} x_n .
\] (71)

For the mean, we have:

\[
E\|\hat{x}^{ur} - \bar{x}\|^2 = \frac{1}{B} \sum_{n=1}^{N} E\left\{\frac{1}{N_p} x_n\right\} = \frac{1}{B} \sum_{n=1}^{N} E[I_n] \times E\left\{\frac{1}{N_p} x_n\right\}
\]

\[
= \frac{1}{B} \sum_{n=1}^{N} B p_n \times \frac{1}{N_p} \bar{x} = \frac{1}{N} \sum_{n=1}^{N} \bar{x} = \bar{x}.
\] (72)

For the variance, we have:

\[
E\|\hat{x}^{ur} - \bar{x}\|^2 = E\left\|\frac{1}{B} \sum_{n=1}^{N} I_n \left(\frac{1}{N_p} x_n - \bar{x}\right)\right\|^2 ,
\]

\[
= \frac{1}{B^2} \sum_{n=1}^{N} E\left\{I_n \left(\frac{1}{N_p} x_n - \bar{x}\right)\right\}^2 ,
\]

\[
+ \frac{1}{B^2} \sum_{n_1 \neq n_2} E\left\{I_{n_1} \left(\frac{1}{N_p} x_{n_1} - \bar{x}\right) I_{n_2} \left(\frac{1}{N_p} x_{n_2} - \bar{x}\right)\right\} .
\] (73)

We begin with:

\[
E\left\{I_n \left(\frac{1}{N_p} x_n - \bar{x}\right)\right\}^2
\]

\[
= E\left\{I_n \left(\frac{1}{N_p} x_n - \bar{x}\right)\right\} \times P(I_n = 1)
\]

\[
+ E\left\{I_n \left(\frac{1}{N_p} x_n - \bar{x}\right)\right\} \times P(I_n = 0)
\]

\[
= B p_n \left\{E\left\{\frac{1}{N_p} x_n - \bar{x}\right\}^2 + \frac{1}{N_p} \bar{x} - \bar{x}\right\} ,
\]

\[
= B p_n \left\{\frac{1}{N_p^2} \sigma_n^2 + \left\|\frac{1}{N_p} \bar{x} - \bar{x}\right\|^2 \right\} .
\] (74)

For the cross-term we have:

\[
E\left\{I_{n_1} \left(\frac{1}{N_p} x_{n_1} - \bar{x}\right) I_{n_2} \left(\frac{1}{N_p} x_{n_2} - \bar{x}\right)\right\}
\]

\[
= E\left\{I_{n_1} \left(\frac{1}{N_p} x_{n_1} - \bar{x}\right) \left(\frac{1}{N_p} x_{n_2} - \bar{x}\right)\right\} \times P(I_{n_1} = 1, I_{n_2} = 1)
\]

\[
= P(I_{n_2} = 1, I_{n_1} = 1) \left\{\left(\frac{1}{N_p} \bar{x}_{n_1} - \bar{x}\right) \left(\frac{1}{N_p} \bar{x}_{n_2} - \bar{x}\right)\right\} ,
\]

\[
= P(I_{n_2} = 1, I_{n_1} = 1) \left\{\left(\frac{1}{N_p} \bar{x}_{n_1} - \bar{x}\right) \left(\frac{1}{N_p} \bar{x}_{n_2} - \bar{x}\right)\right\} .
\] (75)

We then get the desired result.

\(\square\)

We note the following bound on the variance of sampling without replacement estimator. Using Jensen’s inequality, we can get rid of the cross-term, and deduce from (26) the following inequality:

\[
E\|\hat{x}^{ur} - \bar{x}\|^2 \leq \sum_{n=1}^{N} p_n \left\{\frac{1}{N_p^2} \sigma_n^2 + \left\|\frac{1}{N_p} \bar{x} - \bar{x}\right\|^2 \right\} .
\] (76)
\textbf{APPENDIX B}
\textbf{PROOF OF LEMMA 11}

\textit{Proof.} We start with the sampling with-replacement construction. We have $K$ agents from which we sample $L$. Thus, $N$ and $B$ in Lemma 8 are $K$ and $L$, respectively:

$$ x_k = \nabla w^T P_k(w_{l-1}), $$

$$ x_k = \nabla w^T P_k(w_{l-1}), $$

$$ \bar{x} = \frac{1}{K} \sum_{k = 1}^{K} \nabla w^T P_k(w_{l-1}). $$

(77), (78), (79)

Then $\sigma^2_k$, which quantifies the second order moment of the local gradient noise, becomes:

$$ \sigma^2_k = \mathbb{E}\left\{ \left\| \nabla w^T P_k(w_{l-1}) - \nabla w^T P_k(w_{l-1}) \right\|^2 \middle| w_{l-1} \right\}, $$

$$ = \frac{1}{E_k B_k^2} \sum_{e = 1}^{E_k} \sum_{b \in B_{k,i,e}} \mathbb{E}\left\{ \left\| \frac{1}{N_k p_b} \nabla w^T Q_k(w_{l-1}; x_{k,b}) - \nabla w^T P_k(w_{l-1}) \right\|^2 \middle| w_{l-1} \right\}, $$

$$ = \frac{1}{E_k B_k^2} \sum_{b \in B_{k,i,e}} \mathbb{E}\left\{ \left\| \frac{1}{N_k p_b} \nabla w^T Q_k(w_{l-1}; x_{k,b}) - \nabla w^T P_k(w_{l-1}) \right\|^2 \middle| w_{l-1} \right\}, $$

(a)

$$ \leq \frac{3}{E_k B_k^2} \sum_{b \in B_{k,i,e}} \mathbb{E}\left\{ \left\| \frac{1}{N_k p_b} \nabla w^T Q_k(w_{l-1}; x_{k,b}) \right\|^2 \middle| w_{l-1} \right\}, $$

$$ + \mathbb{E}\left\{ \left\| \frac{1}{N_k p_b} \nabla w^T Q_k(w_o; x_{k,b}) - \nabla w^T P_k(w_o) \right\|^2 \middle| w_{l-1} \right\}, $$

$$ = \frac{3}{E_k B_k^2} \sum_{b \in B_{k,i,e}} \left\{ \sum_{n = 1}^{N_k p_b} \left\| \frac{1}{N_k p_b} \nabla w^T Q_k(w_{l-1}; x_{k,b}) \right\|^2 \right\}, $$

(b)

$$ + \sum_{n = 1}^{N_k p_b} \left\| \frac{1}{N_k p_b} \nabla w^T Q_k(w_o; x_{k,b}) - \nabla w^T P_k(w_o) \right\|^2 \right\}, $$

(82)

Then, putting things together, we get:

\[ \mathbb{E}\{\|s_i\|^2 | w_{l-1}\} \leq \frac{K}{E_k B_k^2} \sum_{b \in B_{k,i,e}} \left\{ \sum_{n = 1}^{N_k p_b} \left\| \frac{1}{N_k p_b} \nabla w^T Q_k(w_{l-1}; x_{k,b}) \right\|^2 \right\}, \]

where (a) and (c) follow from using Jensen’s inequality, and (b) follows from using the $\delta$-Lipschitz property of the gradients. Thus, using Lemma 8, we bound the stochastic noise variance as follows:

$$ \mathbb{E}\{\|s_i\|^2 | w_{l-1}\} = \frac{1}{E_k B_k^2} \sum_{b \in B_{k,i,e}} \left\{ \sum_{n = 1}^{N_k p_b} \left\| \frac{1}{N_k p_b} \nabla w^T Q_k(w_{l-1}; x_{k,b}) \right\|^2 \right\}, $$

(81)
\[
\begin{align*}
\mathbb{E}\left\{ \|s_i\|^2 | w_{i-1} \right\} & \leq \frac{1}{L} \sum_{k=1}^{K} p_k \left\{ \frac{\beta_{s,k}^2}{K^2 p_k} \|w_{i-1}\|^2 + \frac{1}{K^2 p_k} \sigma_{s,k}^2 \right. \\
& \quad + \frac{6}{K^2 p_k} E_k B_k \| \nabla_w^T P_k(w) \|^2 + \frac{3}{K^2 p_k} \| \nabla_w^T P_k(w^o) \|^2 \\
& \quad + 3 \delta^2 \left( 1 + \frac{1}{K^2 p_k} \right) \|w_{i-1}\|^2 \left\} \\
& = \frac{1}{L} \sum_{k=1}^{K} p_k \left\{ \frac{\beta_{s,k}^2}{K^2 p_k} \|w_{i-1}\|^2 + \left( 3 + \frac{6}{E_k B_k} \right) \| \nabla_w^T P_k(w^o) \|^2 \right\} \\
& = \frac{1}{L K^2} \sum_{k=1}^{K} \frac{1}{p_k} \left\{ \sigma_{s,k}^2 + \left( 3 + \frac{6}{E_k B_k} \right) \| \nabla_w^T P_k(w^o) \|^2 \right\} \\
& = \frac{3 \delta^2}{L} \left( 1 + \frac{1}{L K^2} \right) \sum_{k=1}^{K} \frac{1}{p_k} \left\{ \beta_{s,k}^2 + \frac{3 \delta^2}{K^2 p_k} \right\} \|w_{i-1}\|^2 \\
& = \beta_{s,k}^2 \|w_{i-1}\|^2 + \sigma_{s,k}^2. \quad (83)
\end{align*}
\]

Next, we move to the sampling without replacement construction. The variance \( \sigma_k^2 \) becomes:

\[
\sigma_k^2 = \mathbb{E} \left\{ \frac{1}{E_k B_k} \sum_{k=1}^{E_k} \sum_{n=1}^{N_k} \left[ 1 + \frac{1}{N_k p_n^{(k)}} \nabla_w^T Q_k(w_{i-1}; x_n) - \nabla_w^T P_k(w_{i-1}) \right]^2 \right\} \\
= \frac{1}{E_k B_k^2} \sum_{k=1}^{E_k} \mathbb{E} \left\{ \left[ 1 + \frac{1}{N_k p_n^{(k)}} \nabla_w^T Q_k(w_{i-1}; x_n) - \nabla_w^T P_k(w_{i-1}) \right]^2 \right\} \\
= \frac{1}{E_k B_k^2} \sum_{k=1}^{E_k} \sum_{n=1}^{N_k} \mathbb{E} \left\{ \left[ 1 + \frac{1}{N_k p_n^{(k)}} \nabla_w^T Q_k(w_{i-1}; x_n) - \nabla_w^T P_k(w_{i-1}) \right]^2 \right\} \\
+ \frac{1}{E_k B_k^2} \sum_{n \neq n_1} \mathbb{E} \left\{ \left[ 1 + \frac{1}{N_k p_n^{(k)}} \nabla_w^T Q_k(w_{i-1}; x_{n_1}) - \nabla_w^T P_k(w_{i-1}) \right]^2 \right\} \\
+ \frac{1}{E_k B_k^2} \sum_{n \neq n_2} \mathbb{E} \left\{ \left[ 1 + \frac{1}{N_k p_n^{(k)}} \nabla_w^T Q_k(w_{i-1}; x_{n_2}) - \nabla_w^T P_k(w_{i-1}) \right]^2 \right\} \\
= \beta_{s,k}^2 \|\bar{w}_{i-1}\|^2 + \sigma_{s,k}^2 + \frac{6}{E_k B_k} \| \nabla_w^T P_k(w^o) \|^2. \quad (84)
\]

Starting with the first term, we use Jensen’s inequality in (a) and (c) and the Lipschitz condition in (b) to get:

\[
\begin{align*}
\frac{1}{E_k B_k^2} \sum_{n=1}^{N_k} \mathbb{E} \left\{ 1 + \frac{1}{N_k p_n^{(k)}} \nabla_w^T Q_k(w_{i-1}; x_n) - \nabla_w^T P_k(w_{i-1}) \right\}^2 & = \mathbb{E} \left\{ 1 + \frac{1}{N_k p_n^{(k)}} \nabla_w^T Q_k(w_{i-1}; x_n) - \nabla_w^T P_k(w_{i-1}) \right\}^2 \\
& \leq \mathbb{E} \left\{ 1 + \frac{1}{N_k p_n^{(k)}} \nabla_w^T Q_k(w_{i-1}; x_n) - \nabla_w^T P_k(w_{i-1}) \right\}^2 \\
& = \frac{1}{E_k B_k^2} \sum_{n=1}^{N_k} \mathbb{E} \left\{ 1 + \frac{1}{N_k p_n^{(k)}} \nabla_w^T Q_k(w_{i-1}; x_n) - \nabla_w^T P_k(w_{i-1}) \right\}^2 \\
& \leq \beta_{s,k}^2 \|\bar{w}_{i-1}\|^2 + \sigma_{s,k}^2 + \frac{6}{E_k B_k} \| \nabla_w^T P_k(w^o) \|^2. \quad (86)
\end{align*}
\]
Next, to bound the second order moment of the gradient noise, we use \((76)\):

\[
\mathbb{E} \left\{ \|s_i\|^2 | w_{i-1} \right\} \leq \sum_{k=1}^{K} p_k \left( \frac{1}{K^2 p_k^2} \sigma_k^2 \right) + \left\| \frac{1}{K p_k^2} \nabla_{w^T} P_k(w_{i-1}) - \frac{1}{K} \sum_{k=1}^{K} \nabla_{w^T} P_i(w_{i-1}) \right\|^2. \tag{87}
\]

The first term is of the same form as for sampling with replacement, and thus can be bounded similarly:

\[
\mathbb{E} \left\{ \|s_i\|^2 | w_{i-1} \right\} \leq \sum_{k=1}^{K} p_k \left\{ \frac{\beta_{i,k}^2}{K p_k^2} \|\tilde{w}_{i-1}\|^2 + \frac{1}{K^2 p_k^2} \sigma_k^2 \right\} + \frac{6}{K^2 p_k^2} \mathbb{E}_k B_k \|\nabla_{w^T} P_k(w^o)\|^2 + 3 \delta \left( 1 + \frac{1}{K^2 p_k^2} \right) \|\tilde{w}_{i-1}\|^2 \right\}
= \beta_{i}^2 \|\tilde{w}_{i-1}\|^2 + \sigma_{i}^2. \tag{88}
\]

**APPENDIX C**

**PROOF OF LEMMA 2**

*Proof.* We first note the following result by using \(\frac{1}{K} \sum_{k=1}^{K} \nabla_{w^T} P_k(w^o) = 0\):

\[
\left\| \tilde{w}_{i-1} + \frac{1}{K} \sum_{k=1}^{K} \nabla_{w^T} P_k(w_{i-1}) \right\|^2 = \|\tilde{w}_{i-1}\|^2 + \mu^2 \left( \frac{1}{K} \sum_{k=1}^{K} \nabla_{w^T} P_k(w^o) - \nabla_{w^T} P_k(w_{i-1}) \right)^2 + 2 \mu \tilde{w}_{i-1}^T \frac{1}{K} \sum_{k=1}^{K} \nabla_{w^T} P_k(w_{i-1}),
\]

\[(a) \leq \|\tilde{w}_{i-1}\|^2 + \mu^2 \left( \frac{1}{K} \sum_{k=1}^{K} \nabla_{w^T} P_k(w^o) - \nabla_{w^T} P_k(w_{i-1}) \right)^2 + 2 \mu \tilde{w}_{i-1}^T \frac{1}{K} \sum_{k=1}^{K} \nabla_{w^T} P_k(w_{i-1}),
\]

\[(b) \leq (1 + \mu^2 \delta^2) \|\tilde{w}_{i-1}\|^2 + 2 \mu \tilde{w}_{i-1}^T \frac{1}{K} \sum_{k=1}^{K} \nabla_{w^T} P_k(w_{i-1}),
\]

\[(c) \leq (1 + \mu^2 \delta^2) \|\tilde{w}_{i-1}\|^2 + 2 \mu \frac{1}{K} \sum_{k=1}^{K} \left( P_k(w^o) - P_k(w_{i-1}) - \frac{\mu}{2} \|\tilde{w}_{i-1}\|^2 \right),
\]

\[(d) \leq (1 + \mu^2 \delta^2) \|\tilde{w}_{i-1}\|^2 - 2 \mu \frac{1}{K} \sum_{k=1}^{K} \nu \|\tilde{w}_{i-1}\|^2,
\]

\[
= (1 - 2\mu \nu + \mu^2 \delta^2) \|\tilde{w}_{i-1}\|^2, \tag{89}
\]

where \((a)\) follows from Jensen’s inequality, \((b)\) from the Lipschitz condition, and \((c)\) and \((d)\) from strong convexity condition.

Returning to the main expression:

\[
\tilde{w}_{i-1} + \mu \frac{1}{K} \sum_{k=1}^{K} \nabla_{w^T} P_k(w_{i-1}) + \mu s_i,
\]

and taking conditional expectations, we obtain:

\[
\mathbb{E} \left\{ \left( \tilde{w}_{i-1} + \mu \frac{1}{K} \sum_{k=1}^{K} \nabla_{w^T} P_k(w_{i-1}) + \mu s_i \right)^2 | w_{i-1} \right\}
= (1 - 2\mu \nu + \mu^2 \delta^2) \mathbb{E} \left\{ \|\tilde{w}_{i-1}\|^2 \right\} + \mu^2 \mathbb{E} \left\{ \|s_i\|^2 \right\}.
\]

**APPENDIX D**

**PROOF OF LEMMA 3**

*Proof.* To show the mean is zero, it is enough to calculate the mean of the approximate gradient. We start with the sampling with replacement scheme where the samples are chosen independently from each other:

\[
\mathbb{E} \left\{ \frac{1}{B_k} \sum_{b \in B, i, e} \frac{1}{N_{k} P_b} \nabla_{w^T} Q_k(w_{k,e-1}; x_{k,b}, \mathcal{F}_{e-1}, \mathcal{L}_i) \right\}
= \frac{1}{B_k} \sum_{b \in B, i, e} \mathbb{E} \left\{ \frac{1}{N_{k} P_b} \nabla_{w^T} Q_k(w_{k,e-1}; x_{k,b}, \mathcal{F}_{e-1}, \mathcal{L}_i) \right\},
\]

\[
= \frac{1}{B_k} \sum_{b \in B, i, e} \frac{1}{N_{k} P_b} \nabla_{w^T} Q_k(w_{k,e-1}; x_{k,b}, \mathcal{F}_{e-1}, \mathcal{L}_i),
\]

\[
= \frac{1}{N_{k} P_b} \nabla_{w^T} Q_k(w_{k,e-1}; x_{k,b}),
\]

\[
= \frac{1}{N_{k}} \sum_{n=1}^{N_{k}} \nabla_{w^T} Q_k(w_{k,e-1}; x_{k,n}). \tag{93}
\]
As for the sampling without replacement scheme, since the samples are now dependent, we introduce the indicator function \( \mathbb{I}_n \) and the derivation goes as follows:

\[
\mathbb{E} \left\{ \frac{1}{B_k} \sum_{b \in B_{k, i, e}} \frac{1}{N_k p_b(k)} \nabla w^\top Q_k(w_{k, e-1}; x_{k,b}) \mid F_{e-1}, \mathcal{L}_i \right\} = \mathbb{E} \left\{ \frac{1}{B_k} \sum_{n=1}^{N_k} \mathbb{I}_n \frac{1}{N_k p_n(k)} \nabla w^\top Q_k(w_{k, e-1}; x_{k,n}) \mid F_{e-1}, \mathcal{L}_i \right\}
\]

\[
= \frac{1}{B_k} \mathbb{E} \left\{ \sum_{n=1}^{N_k} \mathbb{P}(\mathbb{I}_n = 1) \frac{1}{N_k p_n(k)} \nabla w^\top Q_k(w_{k, e-1}; x_{k,n}) \right\}.
\]

\[
= \frac{1}{N_k} \sum_{n=1}^{N_k} \nabla w^\top Q_k(w_{k, e-1}; x_{k,n}). 
\]

(94)

Next, to bound the second order moment, we start with an intermediate step and bound the second order moment of the individual gradient noise of one sample. The derivation below holds regardless of the sampling scheme. By adding and subtracting \( \frac{1}{N_k p_n(k)} \nabla w^\top Q_k(w_{k, e-1}; x_{k,n}) \) and \( \nabla w^\top P_k(w_{k, e-1}) \) and using Jensen’s inequality and Lipschitz condition, we get:

\[
\left\{ \frac{1}{N_k p_n(k)} \nabla w^\top Q_k(w_{k, e-1}; x_{k,n}) - \nabla w^\top P_k(w_{k, e-1}) \right\}^2
\]

\[
\leq 3 \left\{ \frac{1}{N_k p_n(k)} \nabla w^\top Q_k(w_{k, e-1}; x_{k,n}) - \nabla w^\top P_k(w_{k, e-1}) \right\}^2
\]

\[
+ 3 \left\{ \frac{1}{N_k p_n(k)} \nabla w^\top Q_k(w_{k, e-1}; x_{k,n}) \right\}^2 + 3\delta^2 \left\{ \bar{w}_{k,e-1} \right\}^2.
\]

(95)

Then, taking the conditional expectation and using the Lipschitz property, we get:

\[
\mathbb{E} \left\{ \left\{ \frac{1}{N_k p_n(k)} \nabla w^\top Q_k(w_{k, e-1}; x_{k,n}) - \nabla w^\top P_k(w_{k, e-1}) \right\}^2 \mid F_{e-1}, \mathcal{L}_i \right\}
\]

\[
\leq \sum_{n=1}^{N_k} \frac{3}{N_k p_n(k)^2} \left( \left\{ \nabla w^\top Q_k(w_{k, e-1}; x_{k,n}) \right\}^2 + \left\{ \nabla w^\top P_k(w_{k, e-1}) \right\}^2 \right)
\]

\[
\leq \sum_{n=1}^{N_k} \frac{3}{N_k} \left( \delta^2 \left\{ \bar{w}_{k,e-1} \right\}^2 + \left\{ \nabla w^\top Q_k(w_{k, e-1}; x_{k,n}) \right\}^2 \right)
\]

(96)

Now going back to calculating the second order moment of the local incremental gradient noise, we first start with the sampling with replacement. Using the fact that the samples are independent we get:

\[
\mathbb{E} \left\{ \left\{ q_{k,i,e} \right\}^2 \mid F_{e-1}, \mathcal{L}_i \right\}
\]

\[
= \frac{1}{K^2 p_k^2 B_k^2} \sum_{b \in B_{k, i, e}} \mathbb{E} \left\{ \left\{ \frac{1}{N_k p_b(k)} \nabla w^\top Q_k(w_{k, e-1}; x_{k,b}) - \nabla w^\top P_k(w_{k, e-1}) \right\}^2 \mid F_{e-1}, \mathcal{L}_i \right\}
\]

\[
\leq \frac{3\delta^2}{K^2 p_k^2 B_k^2} \left( 1 + \frac{1}{N_k} \sum_{n=1}^{N_k} \frac{1}{p_n(k)} \right) \left\{ \bar{w}_{k,e-1} \right\}^2
\]

\[
+ \frac{3}{K^2 p_k^2 B_k N_k^2} \sum_{n=1}^{N_k} \frac{1}{p_n(k)} \left\{ \nabla w^\top Q_k(w_{k, e-1}; x_{k,n}) \right\}^2.
\]

(97)

As for the sampling without replacement, we also introduce the indicator function and write out the square of sums. The cross-terms disappear since each term has zero mean. The derivation then follows similarly to that of the sampling with replacement. More formally:

\[
\mathbb{E} \left\{ \left\{ q_{k,i,e} \right\}^2 \mid F_{e-1}, \mathcal{L}_i \right\}
\]

\[
= \frac{1}{K^2 p_k^2 B_k^2} \sum_{b \in B_{k, i, e}} \mathbb{E} \left\{ \left\{ \frac{1}{N_k p_b(k)} \nabla w^\top Q_k(w_{k, e-1}; x_{k,b}) - \nabla w^\top P_k(w_{k, e-1}) \right\}^2 \mid F_{e-1}, \mathcal{L}_i \right\}
\]

\[
\leq \frac{3\delta^2}{K^2 p_k^2 B_k^2} \left( 1 + \frac{1}{N_k} \sum_{n=1}^{N_k} \frac{1}{p_n(k)} \right) \left\{ \bar{w}_{k,e-1} \right\}^2
\]

\[
+ \frac{3}{K^2 p_k^2 B_k N_k^2} \sum_{n=1}^{N_k} \frac{1}{p_n(k)} \left\{ \nabla w^\top Q_k(w_{k, e-1}; x_{k,n}) \right\}^2.
\]

(98)
Returning to (99), squaring both sides, conditioning on the filtration \( \mathcal{F}_{e-1} \), and taking expectations we obtain:

\[
\mathbb{E}\left\{ \|\bar{w}_{k,e}\|^2 | \mathcal{F}_{e-1}\right\}
\]

\[
\leq \left( 1 - 2\nu \mu + \mu^2 \left( \delta^2 + \frac{E_k}{K^2p_k^2}\right) \right) \left\| \bar{w}_{k,e-1} \right\|^2 + \frac{1}{K^2p_k^2} \sigma_{q,k}^2,
\]

(101)

where the cross term in (a) is zero because of the zero mean property of the local incremental gradient noise. Taking expectations on both sides again removes the condition on the filtration and leads to the desired result. By further iterating recursion (32) we obtain:

\[
\mathbb{E}\|\bar{w}_{k,e}\|^2 \leq \lambda_k^2 \mathbb{E}\|\bar{w}_{k,0}\|^2 + \frac{1 - \lambda_k}{1 - \lambda_k} \mu^2 \sigma_{q,k}^2.
\]

(102)

**Appendix F**

**Proof of Lemma [4]**

**Proof.** First, using Jensen’s inequality (a) and Lipschitz continuity (b), we obtain:

\[
\|q_i\|^2 \leq \frac{1}{L} \sum_{\ell \in L_i} \frac{K^2p_i^2}{E_i B_i} \mathbb{E}\left\{ \frac{1}{N_\ell^2} \left( \frac{p_i}{p_b} \right)^2 \right\} \sum_{e=1}^{E_i} \sum_{b \in B_{i,e}} \|\nabla_{w^\ell} Q_\ell(w_{\ell,e-1}; x_{\ell,b}) - \nabla_{w^\ell} Q_\ell(w_{\ell-1}; x_{\ell,b})\|^2,
\]

(103)

Next, we focus on \( \|w_{\ell-1} - w_{\ell,e-1}\|^2 \), and by applying Jensen’s inequality in (a) and (b) and Lipschitz condition in (c) we obtain:

\[
\|w_{\ell-1} - w_{\ell,e-1}\|^2
\]

\[
= \mu^2 \left[ \frac{1}{E_i B_i} \sum_{f=0}^{c-2} \sum_{b \in B_{i,f}} \left( \frac{1}{N_f^2} \left( \frac{p_f}{p_b} \right)^2 \right) \|\nabla_{w^\ell} Q_\ell(w_{\ell,f}; x_{\ell,b})\|^2 + \frac{e(1 - \lambda_k)}{1 - \lambda_k} \|w_{\ell-1}\|^2 + \|w_{\ell-1}\|^2 \right] + \frac{1}{K^2p_k^2} \sigma_{q,k}^2
\]

(104)

Then, taking the expectation given the previous filtration \( \mathcal{F}_{e-2} \) and the participating agents \( \mathcal{L}_i \), we see that:

\[
\mathbb{E}\left\{ \|w_{\ell-1} - w_{\ell,e-1}\|^2 | \mathcal{F}_{e-2}, \mathcal{L}_i \right\}
\]

\[
\leq \frac{2\mu^2 \delta_{\ell}^2}{E_i B_i} \sum_{f=0}^{c-2} \left\{ \frac{1}{N_f^2} \left( \frac{p_f}{p_b} \right)^2 \right\} \|\nabla_{w^\ell} Q_\ell(w_{\ell-1}; x_{\ell,f})\|^2 + \frac{2\mu^2}{E_i B_i} \sum_{f=0}^{c-2} \sum_{n=1}^{N_f} \frac{P(I_n = 1)}{N_f^2} \frac{1}{p_n^2} \|\nabla_{w^\ell} Q_\ell(w_{\ell-1}; x_{\ell,f})\|^2,
\]

(105)

where we used Lemma [4] in (a), and in (b) we added and subtracted \( w_0 \) and used Jensen’s inequality. Then, summing
over \( e \) results in:

\[
\frac{1}{E} \sum_{e=1}^{E} \mathbb{E} \left\{ \left\| \mathbf{w}_{t-1} - \mathbf{w}_{t,e-1} \right\|^2 \middle| \mathcal{L}_1 \right\}
\leq \frac{2\mu^2 \sigma^2}{E} \sum_{e=1}^{E} \frac{1}{N^2 \mathbb{P}(t)} \left( \frac{E(t+1)(1-\lambda_t) - 1 + \lambda^2 E_{t+1}}{(1-\lambda_t)^2} \right)
\times \left( \mathbb{E} \left\{ \left\| \mathbf{w}_{t-1} \right\|^2 \middle| \mathcal{L}_1 \right\} + \mathbb{E} \left\{ \left\| \mathbf{w}_o - \mathbf{w}_o^2 \right\|^2 \middle| \mathcal{L}_1 \right\} \right)
+ \frac{E(t+1)(1-\lambda_t)^2 - 2E(t)(1-\lambda_t) + 2\lambda_t - 2\lambda^2 E_{t+1}}{(1-\lambda_t)^3}
\times \frac{\mu^2}{K^2 \mathbb{P}^2(t)} \sigma_{\eta,t}^2
+ \frac{E(t)(E_t - 1)}{3E^2} \sigma_{\eta,t}^2.
\]  

(107)

Taking the expectation of (103) given the choice of the agents and plugging the above expression, we get:

\[
\mathbb{E} \left\{ \left\| \mathbf{q}_i \right\|^2 \middle| \mathcal{L}_1 \right\}
\leq \frac{\sigma^2}{L} \sum_{t \in \mathcal{L}_1} \frac{1}{K^2 \mathbb{P}(t)} \left( a^2 \mu^2 \mathbb{E} \left\{ \left\| \mathbf{w}_{t-1} \right\|^2 \middle| \mathcal{L}_1 \right\} \right)
+ a^2 \mu^2 \mathbb{E} \left\{ \left\| \mathbf{w}_o - \mathbf{w}_o^2 \right\|^2 \middle| \mathcal{L}_1 \right\}
+ (b\mu^2 + c\mu^2) \sigma_{\eta,t}^2
\leq \frac{\sigma^2}{L} \sum_{t \in \mathcal{L}_1} \frac{1}{K^2 \mathbb{P}(t)} \left( b E_{t+1} \right)
+ a^2 \mu^2 \mathbb{E} \left\{ \left\| \mathbf{w}_{t-1} \right\|^2 \middle| \mathcal{L}_1 \right\}
+ (b\mu^2 + c\mu^2) \sigma_{\eta,t}^2
\]  

(108)

where we introduced constants \( a, b, c \) to make the notation simpler. Then, taking again the expectation to remove the conditioning and using Assumption 2 2

\[
\mathbb{E} \left\{ \left\| \mathbf{q}_i \right\|^2 \right\} \leq \frac{\sigma^2}{L} \sum_{k=1}^{K} \frac{1}{K^2 \mathbb{P}(k)} \left( a^2 \mu^2 \mathbb{E} \left\{ \left\| \mathbf{w}_{t-1} \right\|^2 \right\} \right)
+ a^2 \mu^2 \left\{ \left\| \mathbf{w}_o - \mathbf{w}_o^2 \right\|^2 \right\}
+ \sum_{k=1}^{K} \frac{1}{K^2 \mathbb{P}(k)} \left( b\mu^2 + c\mu^2 \right) \sigma_{\eta,t}^2
\]

(109)

Further simplifying the notation gives us the desired result. Thus, since \( a = O(\mu^{-1}) \), \( b = O(\mu^{-2}) \) and \( c = O(1) \), we get

\[
\mathbb{E} \left\{ \left\| \mathbf{q}_i \right\|^2 \right\} = O(\mu).
\]

REFERENCES

[1] E. Rizk, S. Vlaski, and A. H. Sayed, “Dynamic federated learning,” in Proc. IEEE SPAWC, Atlanta, Georgia, USA, 26–29 May 2020, pp. 1–5.

[2] S. Miyato, M. Ishikawa, M. Yamada, and Y. Tolstikhin, “Virtual Adversarial Training,” in Proc. International Conference on Machine Learning, New York, NY, 2017, pp. 1135–1144.

[3] S. Gud飢, M. Welling, J. Shlens, and S. Tenenbaum, “Deep learning for computer vision,” in Proc. International Conference on Machine Learning, 2013, pp. 1135–1144.

[4] P. Kairouz, H. B. McMahan, and S. Rane, “Advances in Neural Information Processing Systems 26,” 2013, pp. 3285–3293.

[5] S. Gud飢, M. Welling, J. Shlens, and S. Tenenbaum, “Deep learning for computer vision,” in Proc. International Conference on Machine Learning, 2013, pp. 1135–1144.

[6] S. Gud飢, M. Welling, J. Shlens, and S. Tenenbaum, “Deep learning for computer vision,” in Proc. International Conference on Machine Learning, 2013, pp. 1135–1144.

[7] S. Gud飢, M. Welling, J. Shlens, and S. Tenenbaum, “Deep learning for computer vision,” in Proc. International Conference on Machine Learning, 2013, pp. 1135–1144.

[8] S. Gud飢, M. Welling, J. Shlens, and S. Tenenbaum, “Deep learning for computer vision,” in Proc. International Conference on Machine Learning, 2013, pp. 1135–1144.

[9] S. Gud飢, M. Welling, J. Shlens, and S. Tenenbaum, “Deep learning for computer vision,” in Proc. International Conference on Machine Learning, 2013, pp. 1135–1144.

[10] S. Gud飢, M. Welling, J. Shlens, and S. Tenenbaum, “Deep learning for computer vision,” in Proc. International Conference on Machine Learning, 2013, pp. 1135–1144.

[11] S. Gud飢, M. Welling, J. Shlens, and S. Tenenbaum, “Deep learning for computer vision,” in Proc. International Conference on Machine Learning, 2013, pp. 1135–1144.

[12] S. Gud飢, M. Welling, J. Shlens, and S. Tenenbaum, “Deep learning for computer vision,” in Proc. International Conference on Machine Learning, 2013, pp. 1135–1144.

[13] S. Gud飢, M. Welling, J. Shlens, and S. Tenenbaum, “Deep learning for computer vision,” in Proc. International Conference on Machine Learning, 2013, pp. 1135–1144.
[28] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, “Adaptive federated learning in resource constrained edge computing systems,” IEEE Journal on Selected Areas in Communications, vol. 37, no. 6, pp. 1205–1221, 2019.

[29] H. Yu, R. Jin, and S. Yang, “On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization,” in Proc. International Conference on Machine Learning, Long Beach, California, Jun 2019, pp. 7184–7193.

[30] C. Xie, S. Koyejo, and I. Gupta, “Asynchronous federated optimization,” arXiv:1903.03934, 2019.

[31] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, “On the convergence of fedavg on non-iid data,” in International Conference on Learning Representations, Addis Ababa, Ethiopia, 2020, pp. 1–26.

[32] L. Liu, J. Zhang, S. H. Song, and K. B. Letaief, “Client-edge-cloud hierarchical federated learning,” arXiv:1905.06641, 2019.

[33] T. Nishio and R. Yonetani, “Client selection for federated learning with heterogeneous resources in mobile edge,” in Proc. IEEE International Conference on Communications (ICC), Shanghai, China, 2019, pp. 1–7.

[34] N. Yoshida, T. Nishio, M. Morikura, K. Yamamoto, and R. Yonetani, “Hybrid-fl for wireless networks: Cooperative learning mechanism using non-iid data,” in Proc. IEEE International Conference on Communications (ICC), Dublin, Ireland, 2020, pp. 1–7.

[35] H. T. Nguyen, V. Sehwag, S. Hosseinalipour, C. G. Brinton, M. Chiang, and H. V. Poor, “Fast-convergent federated learning,” arXiv:2007.13137, 2020.

[36] M. Mohri, G. Sivek, and A. T. Suresh, “Agnostic federated learning,” in Proc. International Conference on Machine Learning (ICML), vol. 97, Long Beach, California, USA, 09–15 Jun 2019, pp. 4615–4625.

[37] T. Li, M. Sanjabi, A. Beirami, and V. Smith, “Fair resource allocation in federated learning,” in Proc. International Conference on Learning Representations, Addis Ababa, Ethiopia, 2020, pp. 1–27.

[38] K. Yuan, B. Ying, S. Vlaski, and A. H. Sayed, “Stochastic gradient descent with finite samples sizes,” in Proc. International Workshop on Machine Learning for Signal Processing (MLSP), Salerno, Italy, 2016, pp. 1–6.

[39] G. Alain, A. Lamb, C. Sankar, A. Courville, and Y. Bengio, “Variance reduction in SGD by distributed importance sampling,” arXiv:1511.06481, 2015.

[40] D. Needell, R. Ward, and N. Srebro, “Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm,” in Proc. Advances in Neural Information Processing Systems, Montreal, Canada, 2014, pp. 1017–1025.

[41] P. Zhao and T. Zhang, “Stochastic optimization with importance sampling for regularized loss minimization,” in Proc. International Conference on Machine Learning (ICML), Lille, France, 2015, pp. 1355–1363.

[42] S. U. Stich, A. Raj, and M. Jaggi, “Safe adaptive importance sampling,” in Proc. International Conference on Machine Learning (ICML), Lille, France, 2015, pp. 1355–1363.

[43] T. D. Needell, R. Ward, and N. Srebro, “Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm,” in Proc. Advances in Neural Information Processing Systems, Montreal, Canada, 2014, pp. 1017–1025.

[44] H. O. Hartley and J. N. K. Rao, “Sampling with unequal probabilities and without replacement,” Ann. Math. Statist., vol. 33, no. 2, pp. 350–374, 06 1962.

[45] D. Prokhorov, “Jcnn 2001 neural network competition,” 2001. [Online]. Available: http://www.csie.ntu.edu.tw/~cj1in/libsvmtools/

[46] K. R. W. Brewer and M. Hanif, Sampling with Unequal Probabilities. New York: Springer-Verlag, 1983.