More Communication Does Not Result in Smaller Generalization Error in Federated Learning

Romain Chor*, Milad Sefidgaran*, Abdellatif Zaidi*†
*Huawei Paris Research Center, France †Université Gustave Eiffel, France
{romain.chor, milad.sefidgaran2}@huawei.com, abdellatif.zaidi@univ-eiffel.fr

***Abstract*—We study the generalization error of statistical learning models in a Federated Learning (FL) setting. Specifically, there are K devices or clients, each holding an independent own dataset of size n. Individual models, learned locally via Stochastic Gradient Descent, are aggregated (averaged) by a central server into a global model and then sent back to the devices. We consider multiple (say $R \in \mathbb{N}^*$) rounds of model aggregation and study the effect of R on the generalization error of the final aggregated model. We establish an upper bound on the generalization error that accounts explicitly for the effect of R (in addition to the number of participating devices K and dataset size n). It is observed that, for fixed (n, K), the bound increases with R, suggesting that the generalization of such learning algorithms is negatively affected by more frequent communication with the parameter server. Combined with the fact that the empirical risk, however, generally decreases for larger values of R, this indicates that R might be a parameter to optimize to reduce the *population risk* of FL algorithms. The results of this paper, which extend straightforwardly to the heterogeneous data setting, are also illustrated through numerical examples.**

I. Introduction and Problem Setup

Consider the network statistical learning model shown in Figure 1. Also, let some *input data* Z be distributed according to an unknown distribution μ over some data space $\mathcal{Z}$. For example, in supervised learning settings $Z := (X, Y)$ where X stands for a data sample and Y stands for its associated label. There are K devices or *clients* each equipped with an individual dataset consisting of n independent and identically distributed (i.i.d.) data points, drawn according to the unknown distribution μ (the extension of the results that will follow to the heterogeneous, *i.e.,* non i.i.d. setting is straightforward). For instance, every device $k \in [K] := \{1, \dots, K\}$, has a dataset $S_k := \{Z_k^{(1)}, \dots, Z_k^{(n)}\} \subseteq \mathcal{Z}^n$. The devices collaboratively train a *(global) model* by performing both local computations and updates based on R-round, $R \in \mathbb{N}^*$, interactions with a *parameter server*. During each round $r \in [R]$, local computations at device $k \in [K]$ are performed using the popular Stochastic Gradient Descent (SGD) algorithm, which applies $\tau = n/R$ updates[1] of its local model, obtained each by taking one gradient step with respect to a sample from its local data S_k. Over all rounds, we assume for simplicity that each client performs an *epoch* over its training dataset *i.e.,* n iterations. Specifically, let for $r \in [R]$ and $t \in [\tau]$, $W_k^{(r,t)}$ denote the individual model of client k as obtained after iteration t of round r. Also, let $\overline{W}^{(r)}$ denote the model obtained by the parameter server at the end of round r, by averaging the individual models of the various devices as obtained during that round, *i.e.,*

$$\overline{W}^{(r)} = \frac{1}{K} \sum_{k=1}^{K} W_k^{(r,\tau)}. \tag{1}$$

This (intermediate) aggregated model is shared with all devices and used by them to update their own local models in the first iteration of the next round, as follows. Without loss of generality, let us denote by $S_{k,r} := \{Z_k^{((r-1)\tau+t)}\}_{t=1}^{\tau}$ the data points of dataset S_k used by device k to perform τ successive one-step SGD updates of its individual model. It is clear that $\{S_{k,1}, \dots, S_{k,R}\}$ forms a partition of S_k, *i.e.,* $\cup_{r=1}^{R} S_{k,r} = S_k$. Also, throughout we let S denote the set of all available datasets, *i.e.,* $S = \cup_{k=1}^{K} S_k \in \mathcal{Z}^{nK}$.

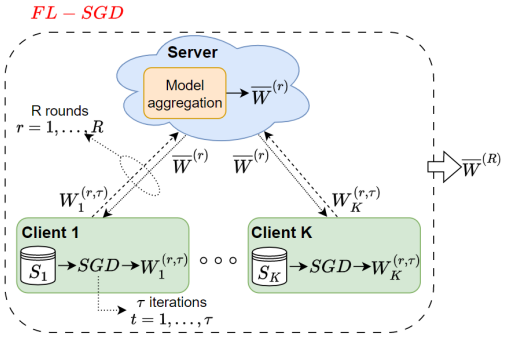

Fig. 1. Multi-round Stochastic Gradient Descent for Federated Learning.

For notational convenience, let for every k and $r \geq 2$, $W_k^{(r,0)}$ designate the previous round's aggregated model as shared back by the parameter server, *i.e.,*

$$W_k^{(r,0)} = \overline{W}^{(r-1)}. \tag{2}$$

For $t = 1, \dots, \tau$, the updates of the model of device k during round $r \in [R]$ are obtained by τ successive one gradient steps as

$$W_k^{(r,t)} = W_k^{(r,t-1)} - \eta_{r,t} \nabla \ell(Z_k^{((r-1)\tau+t)}, W_k^{(r,t-1)}), \tag{3}$$

where $\ell : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}^+$ is the used loss function (assumed to be identical for all devices) and $\eta_{r,t} > 0$ is the learning rate at iteration t of round r. For simplicity, learning rates are assumed to be identical across all devices. Also, in the first round, prior to performing any computation all models are set to some data-independent values.

[1]For ease of exposition, we assume that R divides n.

The algorithm described here is a multi-round distributed SGD for Federated Learning (FL), which we denote hereafter in short as FL-SGD and we use interchangeably the shorthand notations $\mathcal{A}$ and FL-SGD to refer to it, i.e., $\mathcal{A} =$ FL-SGD. Its output hypothesis is the final aggregated model once the R rounds are completed, *i.e.,* $\overline{W} = \overline{W}^{(R)}$; and it can be computed using the recursion equations (1), (2) and (3).

The empirical risk on dataset $s = \{s_1, \ldots, s_K\} \in \mathcal{Z}^{nK}$ of a particular hypothesis $\overline{w} = \mathcal{A}(s)$ is evaluated as the average, over all devices, of its empirical risk for each of them, computed for all used data samples during all rounds, *i.e.,*

$$\hat{\mathcal{L}}(s, \overline{w}) = \frac{1}{nK} \sum_{k=1}^{K} \sum_{r=1}^{R} \sum_{t=1}^{\tau} \ell(z_k^{((r-1)\tau+t)}, \overline{w}). \tag{4}$$

Similarly, the population risk for hypothesis $\overline{w}$ is given as $\mathcal{L}(\overline{w}) = \mathbb{E}_{Z\sim\mu}[\ell(Z, \overline{w})]$; and the generalization error for dataset $s = \{s_1, \ldots, s_K\} \in \mathcal{Z}^{nK}$ and hypothesis $\overline{w} = \mathcal{A}(s)$ is evaluated as

$$\text{gen}(s, \overline{w}) = \mathcal{L}(\overline{w}) - \hat{\mathcal{L}}(s, \overline{w}). \tag{5}$$

The expected generalization error, over all possible datasets $S = \{S_1, \ldots, S_K\} \in \mathcal{Z}^{nK}$, is defined as

$$\mathbb{E}_{S\sim\mu^{\otimes nK}}[\text{gen}(S, \mathcal{A}(S))] = \mathbb{E}_{S\sim\mu^{\otimes nK}}\left[\mathcal{L}(\mathcal{A}(S)) - \hat{\mathcal{L}}(S, \mathcal{A}(S))\right]. \tag{6}$$

where the expectation in (6) is defined also w.r.t. any other possible stochasticity in the learning algorithm.

In this paper we are interested in studying the generalization error of $\mathcal{A} =$ FL-SGD. In particular, we ask the question:

How does the expected generalization error as defined by (6) *evolve with the number of rounds R?*

Such question received so far only partial answer. For example, it was shown theoretically [1]–[3], and also observed experimentally therein, that in FL-type algorithms the empirical risk decreases with the number of rounds. However, to the best of our knowledge, no work has studied this behavior for the generalization error. One central mathematical difficulty in studying the behavior of the expected generalization error as defined by (6) lies in that common tools that are generally applied in similar settings, such as the Leave-one-out Expansion Lemma of [4], do not apply easily when the empirical risk is defined as in (4) (and, so, the generalization error as in (6)). In particular, as it will become clearer throughout when the empirical risk is evaluated as given by (4) the initialization step (2) induces statistical correlations among the devices models' which become stronger with R and are not easy to handle. For example, observe that in the analysis of the contribution of a particular model $W_k^{(r,t)}$ to the overall expected generalization error of the global hypothesis $\overline{W}$ as defined by (6), one has to account for the dependence of $W_k^{(r,t)}$ on other devices' samples $Z_{k'}^{(r'\tau+t)}$ for every $k' \neq k$, $r' < r$ and $t \in [\tau]$. (See Figure 2). Perhaps this explains why while the behavior of (6) was studied in a few works [5]–[8] for the particular case of $R = 1$ (sometimes referred to as "one-shot" FL), much lesser is known in the case of multi-round FL – see Section I-B *Related Works* for few recent works on this, in some of which the mentioned correlations are sometimes eluded by defining the empirical risk differently.

A. Main Contributions

As we already mentioned, in this paper we study the expected generalization error as defined by (6). We focus on the case in which the loss function $\ell(\cdot, \cdot)$ can be expressed as a Bregman divergence [9]. This encompasses a large family of loss functions, including the squared Euclidean distance commonly used in regression problems. We establish an upper bound on the generalization error (6) that accounts explicitly for the number of rounds R. Essentially, the proof techniques involve bounding steps that account judiciously for the statistical correlations induced by (2) and which build up through the rounds. Furthermore, by studying its evolution with the number of rounds R we observe that, for fixed (n, K), the established bound can increase with R, suggesting that the generalization of FL-SGD is negatively affected by more frequent communication with the parameter server. Combined with known results about that the empirical risk, however, generally decreases for larger values of R, this indicates that R might be a parameter to optimize in order to reduce the *population risk* of FL-SGD algorithms. These results, which for simplicity are established here for the i.i.d. data setting and extend easily to the heterogeneous (non i.i.d.) setting, are also illustrated through some numerical examples in which the bound is compared to the *true* (measured) generalization error.

It is noteworthy that the results of this paper extend easily to the case of aperiodic communication with the parameter server and/or more general aggregated models $\overline{W}^{(r)}$, such as any arbitrary deterministic function of the local models $\{W_k^{(r,\tau)}\}_{k=1}^K$ among which the arithmetic average (1) that we consider here is a common choice [10]. Finally, the analysis also carries over easily for settings in which local model updates also account for additional stochasticity through added noise in the gradient steps, *i.e.,* when (3) is substituted by the more general

$$W_k^{(r,t)} = W_k^{(r,t-1)} - \eta_{r,t} \nabla \ell(Z_k^{((r-1)\tau+t)}, W_k^{(r,t-1)}) + \xi_t,$$

where ξ_t stands for some added random noise.

B. Related Works

A major focus of machine learning research over recent years has been the study of statistical learning algorithms when applied to data generated and processed in a distributed (network or graph) manner [10]–[12]. Such setups are of prime importance, especially when some degree of privacy is required [13]–[16] and/or computational power is limited [16], [17]. The study of the behavior of the generalization error, a problem which is understood only very partially even in centralized learning settings, is even more challenging in the case of distributed and multi-round algorithms such as the popular FL [10]. In particular, while a few works [1], [2] have already demonstrated that multi-round communication generally yields smaller empirical risk, only very little is known about the effect of the number of communication

rounds on the generalization error. For the special case of one-round communication, sometimes referred to as "distributed learning" or "one-shot" FL, bounds on the generalization error that improve upon the corresponding ones for the centralized setting with, essentially a factor of $1/\sqrt{K}$, are established in [8] for linear and location models and losses that can be expressed as Bregman divergence; and in [6] and [5] for a broader class of loss functions, using information-theoretic and rate-distortion theoretic approaches.

Compared with the multi-round setup that we study here, the one-round setup suffers from the lack of a joint optimization guarantee, *i.e.,* it may not be possible to make the empirical risk arbitrarily small. From a theoretical angle, however, the study of the generalization error in this latter case is less difficult comparatively, as there are no statistical couplings among the devices' models by the memoryless assumption on the dataset. Most relevant to the problem that we study in this paper is the recent work [8]. In [8], the authors study a quantity, which they argue as being a proxy to the true generalization error as defined by (6), given as

$$\Delta_{\text{SGD}}(s) = \frac{1}{R}\sum_{r=1}^{R}\left(\mathcal{L}(\overline{w}^{(r)}) - \frac{1}{\tau K}\sum_{k=1}^{K}\sum_{t=1}^{\tau}\ell(z_k^{((r-1)\tau+t)}, \overline{w}^{(r)})\right). \tag{7}$$

As the authors mention, this quantity is considered therein mainly for simplicity and in order to avoid accounting for the dependence of $W_k^{(r,t)}$ on other devices' samples $Z_{k'}^{(r'\tau+t)}$ for every $k' \neq k$, $r' < r$ and $t \in [\tau]$. In a sense, this reduces the problem to a virtual one-round setup; but at the expense of analyzing the alternate quantity (7) in place of the true generalization error (6) that we study in this paper.

C. Notations and Organization of the Paper

The rest of the paper is organized as follows. Some technical complements are given in Section II-A. Then, Section II-B presents an upper bound on the expected generalization error of a model learned in the FL setup considered in this paper. The effect of communication on the generalization error of FL algorithms and what insights our bound give on that effect are discussed in Section II-C. This is illustrated through simulations, which are provided in Section III. Finally, our result's proof is given in section IV.

Random variables, their realizations, and their domains are denoted respectively by upper-case, lower-case, and calligraphy fonts, *e.g.,* X, x, and $\mathcal{X}$. Their distributions and expectations are denoted by P_X and $\mathbb{E}[X]$. For two random variables X and Y, $P_{X|Y}$ denotes the conditional distribution of X given Y. $\mathcal{Z}$, $\mathcal{W}$ are subsets of $\mathbb{R}^d$, $d \geq 1$. $\|\cdot\|$ denotes the standard Euclidean norm in $\mathbb{R}^d$. For $a, b \in \mathbb{N}$, $[a; b]$ denotes the set of integers between a and b. The set of integers from 1 to $n \in \mathbb{N}^*$ is denoted by $[n]$. Other specific notations are introduced throughout the paper, whenever they are used.

II. Main Results

In this section we establish the main result of this paper, which is an upper bound on the expected generalization error of the studied FL-SGD as given by (6); and we study its evolution with R.

A. Assumptions and Some Preliminaries

As already mentioned we focus on the case in which the loss function $\ell(\cdot,\cdot)$ is expressed as a Bregman divergence [9], which includes a large family of losses such as the squared error distance used extensively in regression problems (see [18] for more examples of such loss functions). Recall that for a continuously differentiable and strictly convex function $F : \mathbb{R}^d \to \mathbb{R}$, the associated Bregman divergence between two vectors $w, z \in \mathbb{R}^d$ is defined as

$$D_F(w, z) \coloneqq F(w) - F(z) - \langle \nabla F(z), w - z\rangle.$$

where $\langle\cdot,\cdot\rangle$ is the usual inner product.

In the results that will follow we will often make use of one or both following assumptions.

Assumption 1: Let $F : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable and strictly convex function. The loss function $\ell : \mathcal{Z} \times \mathcal{W} \to \mathbb{R}^+$ is the Bregman divergence associated to F *i.e.,* $\forall\, z, w \in \mathcal{Z} \times \mathcal{W}$ we have $\ell(z, w) = D_F(w, z)$.

Assumption 2: The function F defining the Bregman divergence D_F is L-smooth for some $L > 0$ *i.e.,* $\forall\, w, w' \in \mathcal{W}$ we have $\|\nabla F(w) - \nabla F(w')\| \leq L\|w - w'\|$.

In the proof of our upper bound on the expected generalization error (6) that will follow, we make extensive use of the so-called Leave-one-out Lemma [4, Lemma 11], a result which essentially relates the generalization error of a learning algorithm $\mathcal{B}$ to its "average stability" to the replacement of a sample in its training dataset $D \coloneqq \{Z^{(1)}, \ldots, Z^{(m)}\} \subseteq \mathcal{Z}^m$ by an i.i.d. copy of it, *i.e.,* the average difference between losses obtained using a model trained using $\mathcal{B}$ on D and one that is obtained using a model trained using $\mathcal{B}$ on an i.i.d. dataset $\tilde{D} \coloneqq \{\tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(m)}\} \subseteq \mathcal{Z}^m$, where for each $i \in [m]$, $\tilde{Z}^{(i)}$ is an i.i.d. copy of $Z^{(i)} \sim \mu$.

Lemma 1 (Leave-one-out (Expansion) Lemma [4, Lemma 11]): Let $D^{(i)} \coloneqq \{Z^{(1)}, \ldots, \tilde{Z}^{(i)}, \ldots, Z^{(m)}\}$ be a version of $D = \{Z^{(1)}, \ldots, Z^{(i)}, \ldots, Z^{(m)}\}$ in which the element $Z^{(i)}$ is replaced by an i.i.d. copy of it $\tilde{Z}^{(i)}$. Also, denote $\tilde{D} = \{\tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(m)}\}$. Then, it holds that

$$\mathbb{E}_{D\sim\mu^{\otimes m}}[\text{gen}(D, \mathcal{B}(D))] = \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}_{D,\tilde{D}}\big[\ell(\tilde{Z}^{(i)}, \mathcal{B}(D)) - \ell(\tilde{Z}^{(i)}, \mathcal{B}(D^{(i)}))\big]. \tag{8}$$

B. Upper Bound on the Expected Generalization Error

Let for every $k \in [K]$ and $i \in [n]$, $S_k^{(i)}$ designate a copy of the dataset S_k of device k in which $Z_k^{(i)}$ is replaced with an i.i.d. copy of it $\tilde{Z}_k^{(i)}$. That is, $S_k^{(i)} \coloneqq (S_k \setminus \{Z_k^{(i)}\}) \cup \{\tilde{Z}_k^{(i)}\}$. Also, recall the notation $S_{k,r} = \{Z_k^{((r-1)\tau+t)}\}_{t=1}^{\tau}$; and let $I_r \coloneqq [(r-1)\tau + 1 : r\tau]$ designate the indices of data points from $S_{k,r}$. Similarly, for every $i \in I_r$ define $S_{k,r}^{(i)} \coloneqq (S_{k,r} \setminus Z_k^{(i)}) \cup \tilde{Z}_k^{(i)}$. Also, we use the shorthand notation $S_{1:K,r} \coloneqq \cup_{k=1}^{K} S_{k,r}$.

Now, recall the FL-SGD algorithm studied in this paper and described in Section I, whose output hypothesis is $\overline{W} = \overline{W}^{(R)}$

as can be computed using (1), (2) and (3). The analysis of the associated expected generalization error as defined by (6), however, is not easy. One important difficulty is as follows: for every $k' \neq k$, $r' < r$ and $t \in [\tau]$, a change of one sample in $S_{k',r'}$ implies a change of the model $W_{k'}^{(r',t)}$ of device k'; and, in turn, of the (intermediate) aggregated model $\overline{W}^{(r')}$ and all subsequent ones $\overline{W}^{(\tilde{r})}$ for $r' < \tilde{r} \leq R$. In particular, this changes $W_k^{(r,t)}$ for all $t \in [\tau]$. (See Figure 2 for an example with $K = 2$ and $R = 2$). These induced statistical correlations then arise naturally when one applies Lemma 1, which thus becomes less amenable to easy computations.

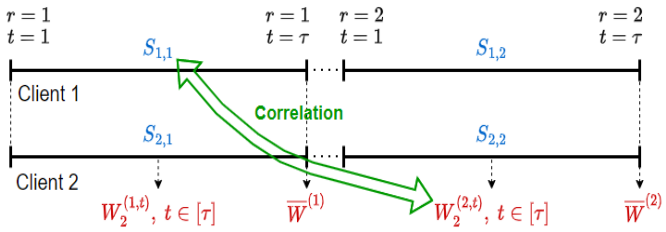

Fig. 2. Illustration of models' coupling for a two-round FL-SGD with $K = 2$.

The next theorem states the main result of this paper, which is an upper bound on the expected generalization error (6) of $\overline{W}$.

Theorem 1: Under Assumptions 1 and 2, it holds that

$$\mathbb{E}_S\big[\mathrm{gen}(S, \overline{W})\big] \leq \frac{1}{RK^2} \sum_{r=1}^{R} \sum_{k=1}^{K} \mathbb{E}_{S_{k,r}}[\mathrm{gen}(S_{k,r}, \mathcal{A}'(r-1, S_{k,r}))] + \sum_{r=1}^{R-1} \frac{L b_{r+1}}{nK^2} \sum_{i \in I_r} \sum_{k=1}^{K} \mathbb{E}\Big[\|\nabla F(\tilde{Z}_k^{(i)})\| \, \|W_{k\setminus i}^{(r,\tau)} - W_k^{(r,\tau)}\|\Big] \quad (9)$$

where:

(i) $\mathcal{A}'(r-1, S_{k,r}) := \mathbb{E}[\mathrm{SGD}(r-1, S_{k,r})]$ where: the expectation is over the distribution of $\cup_{q=1}^{r-1} S_{1:K,q}$ and for $r \geq 2$, $\mathrm{SGD}(r-1, S_{k,r})$ denotes the output of the (*centralized*) SGD algorithm when initialized with $\overline{W}^{(r-1)}$ and applied on samples of $S_{k,r}$ – for $r = 1$, $\mathrm{SGD}(0, S_{k,1}) := W_k^{(1,0)}$.

(ii) $W_{k\setminus i}^{(r,\tau)}$ is the model obtained by client k at the end of the last iteration τ of round r when gradient steps are applied on data points of dataset $S_{k,r}^{(i)}$.

(iii) $b_{r+1} := \sum_{q=r+1}^{R} \sum_{t=1}^{\tau} \eta_{q,t} (\prod_{h=1}^{t-1} 1 + L\eta_{q,h})$, satisfying $\prod_{h=1}^{0} 1 + L\eta_{r,h} = 1$, where L is the smoothness constant of Assumption 2.

(iv) The expectation in the second term of the RHS of (9) is over the joint distribution of $(S_{k,r}, S'_{k,r})$.

Proof: The proof of Theorem 1 is given in Section IV. ∎

We now pause to discuss the result of the theorem. The RHS of (9) may appear as somewhat not amenable to an easy interpretation at first glance. A closer investigation, however, reveals that it is not. In particular, one utility of the result is that it somewhat decouples the aforementioned statistical correlations. Indeed, the first sum term of the RHS of (9) is an average over all devices and rounds of the expected generalization error of a (centralized) *modified* SGD applied at round r by device k on the part $S_{k,r}$ of its local dataset S_k. The modification is in that while the gradient steps are all computed only w.r.t. to samples of $S_{k,r}$ the model learned by this modified SGD is an average (over all parts $S_{k',r'}$ of all devices and all rounds prior to round r), i.e., the term inside the first sum of the RHS of (9) is

$$\mathbb{E}_{S_{k,r}}\Big[\mathrm{gen}\Big(S_{k,r}, \mathbb{E}_{S_{1:K,r'<r}}\big[\mathrm{SGD}(\overline{W}^{(r-1)}, S_{k,r})\big]\Big)\Big]. \quad (10)$$

Equivalently, recalling that SGD iterates essentially consist of an initialization term added to an *innovation* term obtained by application of the gradient of the loss function on a new sample, for every pair $(k, r) \in [K] \times [R]$, (10) captures the statistical correlations caused by the local models' innovation parts till round r. Similarly, for every $(k, r) \in [K] \times [R-1]$ the second sum term of the RHS of (9) captures the statistical correlations caused by the local models' innovation parts from round $(r+1)$ to R. It is noteworthy that these correlations are eluded if instead of the true generalization error (6) one considers the proxy (7) of [8], a setting for which the second term of the RHS of (9) vanishes.

The following corollary is an easy consequence of Theorem 1.

Corollary 1: For "one-shot" FL-SGD, *i.e.,* $R = 1$, if the loss $\ell(\cdot, \cdot)$ satisfies the condition of Assumption 1, then

$$\mathbb{E}_S\Big[\mathrm{gen}(S, \overline{W}^{(1)})\Big] = \frac{1}{K^2} \sum_{k=1}^{K} \mathbb{E}_{S_k}\Big[\mathrm{gen}(S_k, W_k^{(1,\tau)})\Big]. \quad (11)$$

Proof: Observe that in the proof of Theorem 1 in Section IV, in this case ($R = 1$) the second term of (19) is zero. The rest of the proof follows by substituting $\tau = n$ and applying Lemma 3. ∎

It is interesting to observe that in (11) the expected generalization error decays with the number of clients K faster than the average (over clients) of their individual expected generalization errors defined w.r.t. to only their own datasets. The convergence boost is of the order of $1/K$.

C. Effect of Communication Rounds R

For simplicity we assume identical learning rates, *i.e.,* $\eta_{r,t} = \eta$, $\forall (r, t) \in [R] \times [\tau]$. Previous works [19]–[21] have shown that SGD with n iterations on mini-batches of size b, and with learning rate η, has an expected generalization error that (roughly) evolves as $\mathcal{O}(f(b/\eta)/\sqrt{n})$, where $f(b/\eta)$ is a function that captures the dependency on the mini-batch size and learning rate. Moreover, several works [22], [23] have reported that the function $f(\cdot)$ increases with increasing values of the ratio b/η. Hereafter, in particular, we investigate two extreme cases, $R = 1$ and $R = n$ – see the next section for results with other, intermediate, values of R.

For $R = 1$, the setup reduces to one-shot FL [17] in which the local models are all trained in n iterations and aggregated once. In this case, by application of Corollary 1 (see also [8]), we get

$$\mathbb{E}[\text{gen}(S, \overline{W}^{(1)})] = \mathcal{O}\left(\frac{1}{K}\mathbb{E}[\text{gen}(S_1, W_1^{(1,n)})]\right) = \mathcal{O}(f(1/\eta)/\sqrt{nK^2}).$$

For the case $R = n$, the models are aggregated after each local iteration. Thus, the expected generalization error coincides with that of SGD with mini-batch of size $b = K$ and learning rate η. A bound on the generalization error in this case then behaves, roughly, as $\mathcal{O}(f(K/\eta)/\sqrt{nK})$. Since $f(\cdot)$ is an increasing function as observed, *e.g.,* in [22], [23], in particular, this means that for the FL-SGD that we study in this paper the expected generalization error (6) increases with the number of rounds R. This is in line with the findings of [8] and [5] (see also [6]) which have established bounds on the generalization error of a distributed setting that are smaller than the corresponding one of the centralized learning. This, combined with the intuition that more communication generally induces further "*homogeneity*" among the individual devices' models, which then account better for variations in each local dataset, is in accordance with our observation here that (6) increases with R.

The result of Theorem 1 also reflects this evolution with R. Indeed, the first term of (9) computed for $R = n$ *seems* to be larger than the corresponding one for $R = 1$. Moreover, it is easily seen that for $R = 1$ the second term of (9) equals zero whereas it is positive for $R = n$. The observation that (6) increases with R is also illustrated numerically through experiments in the next section.

III. Experimental results

We consider the ordinary least squares (OLS) regression problem. Precisely, the loss function is the squared Euclidean distance *i.e.,* $\forall z := (x, y) \in \mathbb{R}^d \times \mathbb{R}, \forall w \in \mathbb{R}^d$: $\ell(z, w) = (w^t x - y)^2$, which is a Bregman divergence D_F for $F : y \mapsto y^2$. In our experiments, for a dataset S of size nK, we measure the expected generalization error (6) experimentally, compare it to the result of our upper bound of Theorem 1, and depict the evolution of both of them as functions of the number of rounds R.

Each of the K clients is equipped with a subset of S of size n. We implement FL-SGD on a single machine. We train models $\overline{w}^{(R)}$ in that setup for various values of the number of communication rounds R. The population risk $\mathbb{E}_Z[\ell(Z, \overline{w}^{(R)})]$ is estimated by the risk calculated over a test dataset of size $N_{test} = 10^3$. The expectations that are involved in Theorem 1 are approximated by Monte-Carlo simulations for $M = 10^3$. For more details on the experiments, the reader may refer to Appendix B.

The results shown in Figure 3 are obtained with the following numerical values: $d = 10$ (number of features), $n = 500$ and $K = 10$. As visible from the figure, the bound of Theorem 1 captures the increasing behavior of the *true* (measured) generalization error with R. Combined with the known observation that the empirical risk, however, generally decreases for larger values of R (a fact that is also observed in our experiments), this indicates that R might be a parameter to optimize in order to reduce the *population risk* of FL-SGD. The observed gap between the bound (which is tight for large n) and the true values of the generalization error is an indicator that the latter decays faster than $1/n$ for small values of n.

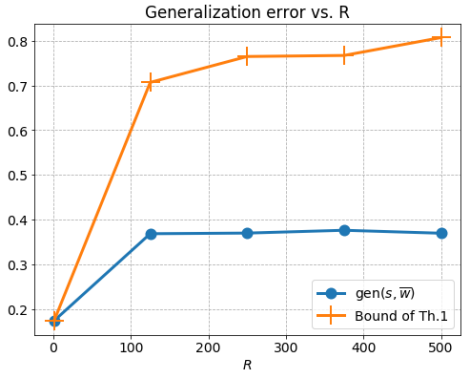

Fig. 3. Evolution of the expected generalization error (6) of FL-SGD and the upper bound in Theorem 1 with the number of communication rounds R.

IV. Proof of Theorem 1

Let the following shorthand notations and substitutions, be used throughout. Define for every pair $(j, k) \in [K]^2$, $r \in [R]$ and $i \in [n]$:

(i) $S^{(k,i)} := (S \backslash \{Z_k^{(i)}\}) \cup \{\tilde{Z}_k^{(i)}\}$ and $S_k^{(i)} := (S_k \backslash \{Z_k^{(i)}\}) \cup \{\tilde{Z}_k^{(i)}\}$.

(ii) $\forall t \in [n]$, $Z_{k\backslash i}^{(t)}$ denotes $\tilde{Z}_k^{(i)}$ if $t = i$, and $Z_k^{(t)}$ otherwise. Similarly, $Z_{j\backslash k,i}^{(t)} = Z_j^{(t)}$ if $j \neq k$, and $Z_{j\backslash k,i}^{(t)} = Z_{k\backslash i}^{(t)}$ if $j = k$. Also, $g_k^{(i)} := \nabla F(\tilde{Z}_k^{(i)})$.

(iii) $\overline{W}_{\backslash k,i}^{(r)} := \frac{1}{K}\sum_{j=1}^K W_{j\backslash k,i}^{(r,\tau)}$, where $W_{j\backslash k,i}^{(r,\tau)}$ is the j-th client's model at of iteration τ of round r and iteration τ when the k-th client's model $W_{k\backslash i}^{(r,\tau)}$ is obtained using the dataset $S_k^{(i)}$.

(iv) For every u, let the following denotes the SGD "innovations" during round $q \geq 2$

$$V_u^{(q)} := \sum_{t=1}^{\tau} \eta_{q,t} \nabla \ell(Z_u^{((q-1)\tau+t)}, W_u^{(q,t-1)}).$$

Finally, throughout we let $S' := \{\tilde{Z}_k^{(i)} : k \in [K], i \in [n]\}$.

In what follows, for convenience we first provide the proof for the specific case of $R = 2$; and then extend it to general R. The proofs of some lemmas, used hereafter, are deferred to Appendix A.

A. Case $R = 2$

First, we state the following lemma which allows to decompose the expected generalization error into two terms that we analyze separately.

Lemma 2: Under Assumption 1, it holds that

$$\mathbb{E}_S[\mathrm{gen}(S,\overline{W}^{(2)})] = \tag{12}$$
$$\frac{1}{nK}\sum_k\sum_{i=1}^{\tau}\mathbb{E}_{S,S'}\left[\left\langle g_k^{(i)}, \overline{W}_{\backslash k,i}^{(1)} - \overline{W}^{(1)}\right\rangle\right]$$
$$+\frac{1}{nK^2}\sum_{i,k,j}\sum_{t=1}^{\tau}\eta_{2,t}\mathbb{E}_{S,S'}\left[\left\langle g_k^{(i)}, \nabla\ell(Z_j^{(\tau+t)}, W_j^{(2,t-1)})\right.\right.$$
$$\left.\left.-\nabla\ell(Z_{j\backslash k,i}^{(\tau+t)}, W_{j\backslash k,i}^{(2,t-1)})\right\rangle\right].$$

Let A and B denote respectively the first and second sum term of the RHS of (12). A accounts for the iterations of the first round, while B accounts for those of the second round; and, so, the devices' models coupling during that round. Recall that for $k\in[K]$, the dataset S_k is partitioned (in this case) into two subsets of equal size $\tau = n/2$: $S_{k,1}$ and $S_{k,2}$, respectively used during the first round and second round. Then, we have the following lemma, derived using the Leave-one-out Lemma.

Lemma 3: For every $k\in[K]$, it holds that

$$\frac{1}{\tau}\sum_{i=1}^{\tau}\mathbb{E}_{S,S'}\left[\left\langle g_k^{(i)}, \overline{W}_{\backslash k,i}^{(1)} - \overline{W}^{(1)}\right\rangle\right] = \tag{13}$$
$$\mathbb{E}_{S_{k,1}}[\mathrm{gen}(S_{k,1}, W_k^{(1,\tau)})].$$

Using Lemma 3 it is easy to see that the first sum term of the RHS of (12) is

$$A = \frac{1}{nK}\sum_{k=1}^{K}\mathbb{E}_{S_{k,1}}[\mathrm{gen}(S_{k,1}, W_k^{(1,\tau)})]. \tag{14}$$

We now analyze the second sum term (B) of the RHS of (12). First recall that at the end of the first round, the local models are aggregated as $\overline{W}^{(1)} = (\sum_{k=1}^K W_k^{(1,\tau)})/K$. Also, $W_k^{(2,0)} = \overline{W}^{(1)}$. Then, the term B can be re-written as

$$B \coloneqq \frac{1}{nK^2}\sum_{i,k,j}\sum_{t=1}^{\tau}\eta_{2,t}\mathbb{E}_{S,S'}\left[\left\langle g_k^{(i)}, \nabla\ell(Z_j^{(\tau+t)}, W_j^{(2,t-1)})\right.\right.$$
$$\left.\left.-\nabla\ell(Z_{j\backslash k,i}^{(\tau+t)}, W_{j\backslash k,i}^{(2,t-1)})\right\rangle\right]$$
$$= \frac{1}{nK^2}\sum_{k,j}\sum_{i=1}^{\tau}\mathbb{E}_{S,S'}\left[\left\langle g_k^{(i)}, V_j^{(2)} - V_{j\backslash k,i}^{(2)}\right\rangle\right]$$
$$+\frac{1}{nK^2}\sum_{k}\sum_{i=\tau+1}^{n}\mathbb{E}_{S,S'}\left[\left\langle g_k^{(i)}, V_k^{(2)} - V_{k\backslash i}^{(2)}\right\rangle\right], \tag{15}$$

where in the second equality we used that:

(i) $Z_{j\backslash k,i}^{(\tau+t)} = Z_j^{(\tau+t)}$, for $i\in[\tau],\ j\in[K]$.

(ii) the sum over j vanishes because $\nabla\ell(Z_j^{(\tau+t)}, W_j^{(2,t-1)}) - \nabla\ell(Z_{j\backslash k,i}^{(\tau+t)}, W_{j\backslash k,i}^{(2,t-1)}) = 0$ for $i>\tau,\ j\neq k$.

The second sum term of the RHS of (15) can be computed using the following lemma.

Lemma 4: For every $k\in[K]$, it holds that

$$\frac{1}{\tau}\sum_{i=\tau+1}^{n}\mathbb{E}_{S,S'}\left[\left\langle g_k^{(i)}, V_k^{(2)} - V_{k\backslash i}^{(2)}\right\rangle\right]$$
$$= \mathbb{E}_{S_{k,2}}[\mathrm{gen}(S_{k,2}, \mathcal{A}'(1, S_{k,2}))]. \tag{16}$$

The first sum term of the RHS of (15), as for it, is upper-bounded using the following lemma.

Lemma 5: Under Assumption 2, $\forall\, k\in[K], \forall\, i\in[\tau]$, we have

$$\sum\nolimits_j \mathbb{E}_{S,S'}\left[\left\langle g_k^{(i)}, V_j^{(2)} - V_{j\backslash k,i}^{(2)}\right\rangle\right]$$
$$\leq Lb_2\mathbb{E}_{S_{k,1},S'_{k,1}}\left[\|g_k^{(i)}\|\,\|W_{k\backslash i}^{(1,\tau)} - W_k^{(1,\tau)}\|\right] \tag{17}$$

where $b_2 \coloneqq \sum_{t=1}^{\tau}\eta_{2,t}(\prod_{h=1}^{t-1}1 + L\eta_{2,h})$.
Continuing from (12) using Lemma 4 and 5 we get

$$B \leq \frac{Lb_2}{nK^2}\sum\nolimits_k\sum\nolimits_{i=1}^{\tau}\mathbb{E}_{S_{k,1},S'_{k,1}}\left[\|g_k^{(i)}\|\,\|W_{k\backslash i}^{(1,\tau)} - W_k^{(1,\tau)}\|\right]$$
$$+\frac{1}{2K^2}\sum\nolimits_k\mathbb{E}_{S_{k,2}}[\mathrm{gen}(S_{k,2}, \mathcal{A}'(1, S_{k,2}))]. \tag{18}$$

Summarizing: substituting the terms in (12) using (14) and (18) completes the proof of the theorem for $R=2$.

B. Extension to Arbitrary R

First note that Lemma 2 and its proof can be generalized easily to arbitrary $R\geq 2$. That is, under Assumption 1,

$$\mathbb{E}_S[\mathrm{gen}(S,\overline{W}^{(R)})]$$
$$= \frac{1}{nK^2}\sum_{i,k}\mathbb{E}_{S,S'}[\langle g_k^{(i)}, W_{k\backslash i}^{(1,\tau)} - W_k^{(1,\tau)}\rangle]$$
$$+\frac{1}{nK^2}\sum_{i,k,j}\sum_{q=2}^{R}\mathbb{E}_{S,S'}\left[\left\langle g_k^{(i)}, V_j^{(q)} - V_{j\backslash k,i}^{(q)}\right\rangle\right] \tag{19}$$

Recalling that for $r\in[R]$ we have $I_r = [(r-1)\tau+1 : r\tau]$, the second sum term of (19) can be written equivalently as

$$\frac{1}{nK^2}\sum_{\substack{k\in[K]\\ j\in[K]}}\sum_{\substack{r\in[R]\\ i\in I_r}}\sum_{q\in[2;R]}\mathbb{E}_{S,S'}\left[\left\langle g_k^{(i)}, V_j^{(q)} - V_{j\backslash k,i}^{(q)}\right\rangle\right]. \tag{20}$$

For fixed $(k,j)\in[K]^2$ and $r\in[R]$, denote for $q\in[2;R]$, $i\in I_r$, $C_{q,i} \coloneqq \mathbb{E}_{S,S'}[\langle g_k^{(i)}, V_j^{(q)} - V_{j\backslash k,i}^{(q)}\rangle]$ (notational dependence on (j,k,r) is omitted for simplicity). Also, we have:

(i) For $q<r$, $C_{q,i}=0$ since $\mathbb{E}[V_j^{(q)}] = \mathbb{E}[V_{j\backslash k,i}^{(q)}]$.

(ii) For $q=r$, $\sum_{i\in I_r}C_{q,i} = \mathbb{E}_{S_{k,r}}[\mathrm{gen}(S_{k,r}, \mathcal{A}'(r-1, S_{k,r}))]$. The proof of this equality uses an easy extension of Lemma 4 to any R and is omitted for brevity.

(iii) The term $\sum_{q=r+1}^{R}\sum_{i\in I_r}C_{q,i}$ is bounded by

$$Lb_{r+1}\mathbb{E}_{S_{k,r},S'_{k,r}}[\|\nabla F(\tilde{Z}_k^{(i)})\|\,\|W_{k\backslash i}^{(r,\tau)} - W_k^{(r,\tau)}\|],$$

where $b_{r+1} \coloneqq \sum_{q=r+1}^{R}\sum_{t=1}^{\tau}\eta_{q,t}(\prod_{h=1}^{t-1}1 + L\eta_{q,h})$ and $\prod_{h=1}^{0}1+L\eta_{q,h} = 1$. This uses an extension of Lemma 5.

Finally, using (19) and substituting using (20) and the above completes the proof of Theorem 1.

REFERENCES

[1] S. U. Stich, "Local sgd converges fast and communicates little," in *2019 International Conference on Learning Representations (ICLR)*, 2019.

[2] F. Haddadpour, M. M. Kamani, M. Mahdavi, and V. Cadambe, "Local sgd with periodic averaging: Tighter analysis and adaptive synchronization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[3] T. Qin, S. R. Etesami, and C. A. Uribe, "Faster convergence of local sgd for over-parameterized models," *arXiv preprint arXiv:2201.12719*, 2022.

[4] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, "Learnability, stability and uniform convergence," *The Journal of Machine Learning Research*, vol. 11, pp. 2635–2670, 2010.

[5] M. Sefidgaran, R. Chor, and A. Zaidi, "Rate-distortion theoretic bounds on generalization error for distributed learning," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 19 687–19 702.

[6] S. Yagli, A. Dytso, and H. V. Poor, "Information-theoretic bounds on the generalization error and privacy leakage in federated learning," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2020, pp. 1–5.

[7] Y. Bu, S. Zou, and V. V. Veeravalli, "Tightening mutual information-based bounds on generalization error," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 121–130, 2020.

[8] L. Barnes, A. Dytso, and H. V. Poor, "Improved information theoretic generalization bounds for distributed and federated learning," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 1465–1470.

[9] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR computational mathematics and mathematical physics*, vol. 7, no. 3, pp. 200–217, 1967.

[10] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[11] O. Gupta and R. Raskar, "Distributed learning of deep neural network over multiple agents," *Journal of Network and Computer Applications*, vol. 116, pp. 1–8, 2018.

[12] M. Moldoveanu and A. Zaidi, "In-network learning for distributed training and inference in networks," in *2021 IEEE Globecom Workshops (GC Wkshps)*. IEEE, 2021, pp. 1–6.

[13] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek, and H. V. Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[14] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang, and Y. Zhou, "A hybrid approach to privacy-preserving federated learning," in *Proceedings of the 12th ACM workshop on artificial intelligence and security*, 2019, pp. 1–11.

[15] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Generation Computer Systems*, vol. 115, pp. 619–640, 2021.

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

[17] M. Zinkevich, M. Weimer, L. Li, and A. Smola, "Parallelized stochastic gradient descent," *Advances in neural information processing systems*, vol. 23, 2010.

[18] A. Banerjee, S. Merugu, I. S. Dhillon, J. Ghosh, and J. Lafferty, "Clustering with bregman divergences," *Journal of machine learning research*, vol. 6, no. 10, 2005.

[19] M. Wang and C. Ma, "Generalization error bounds for deep neural networks trained by sgd," *arXiv preprint arXiv:2206.03299*, 2022.

[20] Y. Cao and Q. Gu, "Generalization bounds of stochastic gradient descent for wide and deep neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

[21] G. Neu, G. K. Dziugaite, M. Haghifam, and D. M. Roy, "Information-theoretic generalization bounds for stochastic gradient descent," in *Conference on Learning Theory*. PMLR, 2021, pp. 3526–3545.

[22] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," in *2017 International Conference on Learning Representations (ICLR)*, 2017.

[23] S. Jastrzebski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey, "Three factors influencing minima in sgd," *arXiv preprint arXiv:1711.04623*, 2017.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

APPENDIX A
MISSING PROOFS IN MAIN TEXT

In this Appendix, we provide proofs for the technical lemmas used in the main proof *i.e.,* the one of Theorem 1.

A. Proof of Lemma 2

$$
\begin{aligned}
&\mathbb{E}_S\Big[\mathrm{gen}(S,\overline{W}^{(2)})\Big]\\
&\overset{(a)}{=}\frac{1}{nK}\sum_{k=1}^{K}\sum_{i=1}^{n}\mathbb{E}_{S,S'}[\ell(\tilde{Z}_k^{(i)},\overline{W}^{(2)})-\ell(\tilde{Z}_k^{(i)},\overline{W}^{(2)}_{\setminus k,i})]\\
&\overset{(b)}{=}\frac{1}{nK}\sum_{k,i}\mathbb{E}_{S,S'}\Big[F(\overline{W}^{(2)})-F(\tilde{Z}_k^{(i)})\\
&\qquad\qquad -\langle\nabla F(\tilde{Z}_k^{(i)}),\overline{W}^{(2)}-\tilde{Z}_k^{(i)}\rangle\\
&\qquad\qquad -F(\overline{W}^{(2)}_{\setminus k,i})+F(\tilde{Z}_k^{(i)})\\
&\qquad\qquad +\langle\nabla F(\tilde{Z}_k^{(i)}),\overline{W}^{(2)}_{\setminus k,i}-\tilde{Z}_k^{(i)}\rangle\Big]\\
&\overset{(c)}{=}\frac{1}{nK}\sum_{i,k}\mathbb{E}_{S,S'}\Big[\langle\nabla F(\tilde{Z}_k^{(i)}),\overline{W}^{(2)}_{\setminus k,i}-\overline{W}^{(2)}\rangle\Big]\\
&\overset{(d)}{=}\frac{1}{nK}\sum_{i,k}\mathbb{E}_{S,S'}\Big[\langle g_k^{(i)},\overline{W}^{(1)}_{\setminus k,i}-\overline{W}^{(1)}\rangle\Big]\\
&\quad+\frac{1}{nK^2}\sum_{i,k,j}\sum_{t=1}^{\tau}\eta_{2,t}\mathbb{E}_{S,S'}\Big[\langle g_k^{(i)},\nabla\ell(Z_j^{(\tau+t)},W_j^{(2,t-1)})\\
&\qquad\qquad -\nabla\ell(Z_{j\setminus k,i}^{(\tau+t)},W_{j\setminus k,i}^{(2,t-1)})\rangle\Big]
\end{aligned}
$$

where

- (a) uses the leave-one-out lemma (Lemma 1), applied to S and $\overline{W}^{(2)}$.
- (b) comes from the definition of the loss function.
- (c) uses that $F(\overline{W}^{(2)})$ and $F(\overline{W}^{(2)}_{\setminus k,i})$ have the same expected value.
- (d) uses:

$$\overline{W}^{(2)}=\overline{W}^{(1)}-\frac{1}{K}\sum_{j=1}^{K}\sum_{t=\tau+1}^{T}\eta_t\nabla\ell(Z_j^{(\tau+t)},W_j^{(2,t-1)}).$$

$Z_{j\setminus k,i}^{(\tau+t)}\coloneqq Z_{k\setminus i}^{(\tau+t)}$ if $j=k$, where $Z_{k\setminus i}^{(\tau+t)}$ is the t-th sample of $S_k^{(i)}$. Moreover $W_{j\setminus k,i}^{(2,t-1)}$ is the model of client j given that client k trains its model $W_{k\setminus i}^{(2,t-1)}$ with the dataset $S_k^{(i)}$.

B. Proof of Lemma 3

$\forall k\in[K]$:

$$
\begin{aligned}
&\frac{1}{\tau}\sum_{i=1}^{\tau}\mathbb{E}_{S,S'}\Big[\langle g_k^{(i)},\overline{W}^{(1)}_{\setminus k,i}-\overline{W}^{(1)}\rangle\Big]\\
&\quad=\frac{1}{\tau}\sum_{i=1}^{\tau}\mathbb{E}_{S,S'}\Big[\langle g_k^{(i)},W_{k\setminus i}^{(1,\tau)}-W_k^{(1,\tau)}\rangle\Big]\\
&\quad=\frac{1}{\tau}\sum_{i=1}^{\tau}\mathbb{E}_{S,S'}\Big[\ell(\tilde{Z}_k^{(i)},W_k^{(1,\tau)})-\ell(\tilde{Z}_k^{(i)},W_{k\setminus i}^{(1,\tau)})\Big]\\
&\quad=\mathbb{E}_{S_{k,1}}[\mathrm{gen}(S_{k,1},W_k^{(1,\tau)})]
\end{aligned}
$$

- Before the first communication round, local models $W_j^{(1,t)}$ and $W_{j\setminus k,i}^{(1,t)}$ are the same (because trained on the same datapoints) excepted for client k which yields the first equation.
- Second inequality follows by adding the appropriate cancelled out terms in the loss function ℓ.
- Last equality comes from an application of Lemma 1 to $S_{k,1}$ and $W_k^{(1,\tau)}$. Moreover, $|S_{k,1}|=\tau=n/2$.

C. Proof of Lemma 4

$\forall k\in[K]$:

$$
\begin{aligned}
&\frac{1}{\tau}\sum_{i=\tau+1}^{n}\mathbb{E}_{S,S'}\Big[\langle g_k^{(i)},V_k^{(2)}-V_{k\setminus i}^{(2)}\rangle\Big]\\
&\quad\overset{(a)}{=}\frac{1}{\tau}\sum_{i=\tau+1}^{n}\mathbb{E}_{S_{1:K,2},S'_{1:K,2}}\Big[\Big\langle g_k^{(i)},\mathbb{E}_{S_{1:K,1}}[V_k^{(2)}]\\
&\qquad\qquad -\mathbb{E}_{S_{1:K,1}}[V_{k\setminus i}^{(2)}]\Big\rangle\Big]\\
&\quad\overset{(b)}{=}\frac{1}{\tau}\sum_{i=\tau+1}^{n}\mathbb{E}_{S_{1:K,2},S'_{1:K,2}}\Big[\ell(\tilde{Z}_k^{(i)},\mathcal{A}'(1,S_{k,2}))\\
&\qquad\qquad -\ell(\tilde{Z}_k^{(i)},\mathcal{A}'(1,S_{k,2}^{(i)}))\Big]\\
&\quad=\mathbb{E}_{S_{k,2}}[\mathrm{gen}(S_{k,2},\mathcal{A}'(1,S_{k,2}))]
\end{aligned}
$$

where

- $S=S_{1:K,1}\cup S_{1:K,2}$, $S'=S'_{1:K,1}\cup S'_{1:K,2}$ which are all independent and $V_{k\setminus i}^{(2)}$ independent of $S'_{1:K,1}$ for $i\in[\tau+1,n]$. Using Fubini-Lebesgue's theorem gives (a).
- $\mathcal{A}'(1,S_{k,2})=\mathbb{E}_{S_{1:K,1}}[V_k^{(2)}]$ denotes the output of $\mathcal{A}'$, which is initialized with $\overline{W}^{(1)}$ and uses $S_{k,2}$. This yields (b).
- Applying Lemma 1 to $S_{k,2}$ and $\mathcal{A}'_k(1,S_{k,2})$ gives the last equality.

D. Proof of Lemma 5

$\forall k\in[K],\ \forall i\in[\tau]$:

$$
\begin{aligned}
&\sum_j\mathbb{E}_{S,S'}\Big[\Big\langle g_k^{(i)},V_j^{(2)}-V_{j\setminus k,i}^{(2)}\Big\rangle\Big]\\
&\quad\overset{(c)}{=}\sum_j\sum_{t=1}^{\tau}\eta_{2,t}\mathbb{E}_{S,S'}\Big[\Big\langle g_k^{(i)},\nabla F(W_j^{(2,t-1)})\\
&\qquad\qquad -\nabla F(W_{j\setminus k,i}^{(2,t-1)})\Big\rangle\Big]\\
&\quad\overset{(d)}{\leq}L\sum_j\sum_{t=1}^{\tau}\eta_{2,t}\mathbb{E}_{S,S'}\Big[\|g_k^{(i)}\|.\|W_{j\setminus k,i}^{(2,t-1)}-W_j^{(2,t-1)}\|\Big]\\
&\quad\overset{(e)}{\leq}Lb_2\mathbb{E}_{S_{k,1},S'_{k,1}}\Big[\|g_k^{(i)}\|.\|W_{k\setminus i}^{(1,\tau)}-W_k^{(1,\tau)}\|\Big]
\end{aligned}
$$

where

- (c) is due to $\forall z, w$:

$$\begin{aligned}\nabla \ell(z, w) &= \nabla_w (F(w) - F(z) - \langle \nabla F(z), w - z \rangle) \\ &= \nabla F(w) - \nabla F(z).\end{aligned}$$

- (d) uses Cauchy-Schwarz inequality and Assumption 2.
- (e) uses the following inequality, obtained by recursion: $\forall j \in [K],\ \forall t \in [\tau], \forall i \in [\tau]$:

$$\begin{aligned}&\|W_{j\setminus k,i}^{(2,t-1)} - W_j^{(2,t-1)}\| \\ &= \|W_{j\setminus k,i}^{(2,t-2)} - W_j^{(2,t-2)} - \eta_{2,t-1}(\nabla \ell(Z_j^{(\tau+t)}, W_{j\setminus k,i}^{(2,t-2)}) \\ &\qquad\qquad - \nabla \ell(Z_j^{(\tau+t)}, W_j^{(2,t-2)}))\| \\ &\leq (1 + L\eta_{2,t-1}) \|W_{j\setminus k,i}^{(2,t-2)} - W_j^{(2,t-2)}\| \\ &\leq \left(\prod_{h=1}^{t-1} 1 + L\eta_{2,h} \right) \|\overline{W}_{\setminus k,i}^{(1)} - \overline{W}^{(1)}\| \\ &= \left(\prod_{h=1}^{t-1} 1 + L\eta_{2,h} \right) \|W_{k\setminus i}^{(1)} - W_k^{(1)}\|.\end{aligned}$$

APPENDIX B
DETAILS OF THE EXPERIMENTS

a) Datasets: We conducted the simulations on datasets implemented in the open source Machine Learning library *Scikit-Learn* [24], which are "california_housing" and "friedman1". The results of our numerical experiments were consistent with both datasets; Figure 3 presents simulations using "friedman1" dataset.

b) Model & algorithm implementation: OLS and LocalSGD are implemented using *SGDRegressor* from Scikit-Learn and custom Python classes.

c) Training and hyperparameters: The models were trained for one epoch, to coincide with the theoretical setup of this paper. The learning rate was set to $\eta = 0.01$.

d) Hardware and other resources: We performed our experiments on a machine equipped with 56 CPUs Intel Xeon E5-2690v4 2.60GHz. The experiments are conducted using Python language.