FEDERATED LEARNING OF A MIXTURE OF GLOBAL AND LOCAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a new optimization formulation for training federated learning models. The standard formulation has the form of an empirical risk minimization problem constructed to find a single global model trained from the private data stored across all participating devices. In contrast, our formulation seeks an explicit trade-off between this traditional global model and the local models, which can be learned by each device from its own private data without any communication. Further, we develop several efficient variants of SGD (with and without partial participation and with and without variance reduction) for solving the new formulation and prove communication complexity guarantees. Notably, our methods are similar but not identical to federated averaging / local SGD, thus shedding some light on the essence of the elusive method. In particular, our methods do not perform full averaging steps and instead merely take steps towards averaging. We argue for the benefits of this new paradigm for federated learning.

1 INTRODUCTION

With the proliferation of mobile phones, wearable devices, tablets, and smart home devices comes an increase in the volume of data captured and stored on them. This data contains a wealth of potentially useful information to the owners of these devices, and more so if appropriate machine learning models could be trained on the heterogeneous data stored across the network of such devices. The traditional approach involves moving the relevant data to a data center where centralized machine learning techniques can be efficiently applied (Dean et al., 2012; Reddi et al., 2016). However, this approach is not without issues. First, many device users are increasingly sensitive to privacy concerns and prefer their data to never leave their devices. Second, moving data from their place of origin to a centralized location is very inefficient in terms of energy and time.

1.1 FEDERATED LEARNING

Federated learning (FL) (McMahan et al., 2016; Konečný et al., 2016b;a; McMahan et al., 2017) has emerged as an interdisciplinary field focused on addressing these issues by training machine learning models directly on edge devices. The currently prevalent paradigm (Li et al., 2019; Kairouz et al., 2019) casts supervised FL as an empirical risk minimization problem of the form

$$\min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where n is the number of devices participating in training, $x \in \mathbb{R}^d$ encodes the d parameters of a *global model* (e.g., weights of a neural network) and

$$f_i(x) := \mathbb{E}_{\xi\sim\mathcal{D}_i}\left[f(x,\xi)\right]$$

represents the aggregate loss of model x on the local data represented by distribution $\mathcal{D}_i$ stored on device i. One of the defining characteristics of FL is that the data distributions $\mathcal{D}_i$ may possess very different properties across the devices. Hence, any potential FL method is explicitly required to be able to work under the *heterogeneous* data setting.

The most popular method for solving (1) in the context of FL is the FedAvg algorithm (McMahan et al., 2016). In its most simple form, when one does not employ partial participation, model compression, or stochastic approximation, FedAvg reduces to Local Gradient Descent (LGD) (Khaled

et al., 2019; 2020), which is an extension of GD performing more than a single gradient step on each device before aggregation. FedAvg has been shown to work well empirically, particularly for non-convex problems, but comes with poor convergence guarantees compared to the non-local counterparts when data are heterogeneous.

Some issues with current approaches to FL

The first motivation for our research comes from the appreciation that data heterogeneity does not merely present challenges to the design of new provably efficient training methods for solving (1), but *also inevitably raises questions about the utility of such a global solution to individual users.* Indeed, a global model trained across all the data from all devices might be so removed from the typical data and usage patterns experienced by an individual user as to render it virtually useless. This issue has been observed before, and various approaches have been proposed to address it. For instance, the MOCHA (Smith et al., 2017) framework uses a multi-task learning approach to allow for personalization. Next, (Khodak et al., 2019) propose a generic online algorithm for gradient-based parameter-transfer meta-learning and demonstrate improved practical performance over FedAvg (McMahan et al., 2017). Approaches based on variational inference (Corinzia & Buhmann, 2019), cyclic patterns in practical FL data sampling (Eichner et al., 2019) transfer learning (Zhao et al., 2018) and explicit model mixing (Peterson et al., 2019) have been proposed.

The second motivation for our work is the realization that even very simple variants of FedAvg, such as LGD, which should be easier to analyze, fail to provide theoretical improvements in communication complexity over their non-local cousins, in this case, GD (Khaled et al., 2019; 2020). This observation is at odds with the practical success of local methods in FL. This leads us to ask the question: *if LGD does not theoretically improve upon GD as a solver for the traditional global problem* (1)*, perhaps LGD should not be seen as a method for solving* (1) *at all. In such a case, what problem does LGD solve?* A good answer to this question would shed light on the workings of LGD, and by analogy, on the role local steps play in more elaborate FL methods such as local SGD (Stich, 2020; Khaled et al., 2020) and FedAvg.

2 Contributions

In our work we argue that the two motivations mentioned in the introduction point in the same direction, i.e., we show that *a single solution can be devised addressing both problems at the same time.* Our main contributions are:

$\diamond$ **New formulation of FL which seeks an implicit mixture of global and local models.** We propose a *new optimization formulation of FL.* Instead of learning a single global model by solving (1), we propose to learn a mixture of the global model and the purely local models which can be trained by each device i using its data $\mathcal{D}_i$ only. Our formulation (see Sec. 3) lifts the problem from $\mathbb{R}^d$ to $\mathbb{R}^{nd}$, allowing each device i to learn a personalized model $x_i \in \mathbb{R}^d$. These personalized models are encouraged to not depart too much from their mean by the inclusion of a quadratic penalty ψ multiplied by a parameter $\lambda \geq 0$.[1]

$\diamond$ **Theoretical properties of the new formulation.** We study the properties of the optimal solution of our formulation, thus developing an algorithmic-free theory. When the penalty parameter is set to zero, then obviously, each device is allowed to train their own model without any dependence on the data stored on other devices. Such purely local models are rarely useful. We prove that the optimal local models converge to the traditional global model characterized by (1) at the rate $\mathcal{O}(1/\lambda)$. We also show that the total loss evaluated at the local models is never higher than the total loss evaluated at the global model (see Thm. 3.1). Moreover, we prove an insightful structural result for the optimal local models: the optimal model learned by device i arises by subtracting the gradient of the loss function stored on that device evaluated at the same point (i.e., a local model) from the average of the

[1]The idea of softly-enforced similarity of the local models was already introduced in the domain of the multi-task relationship learning (Zhang & Yeung, 2010; Liu et al., 2017; Wang et al., 2018) and distributed optimization (Lan et al., 2018; Gorbunov et al., 2019). The mixture objective we propose (see (2)) is in fact a special case of their setup, which justifies our approach from the modeling perspective. Note that Liu et al. (2017); Wang et al. (2018) provide efficient algorithms to solve the mixture objective already. However, none of the mentioned papers consider the FL application, nor they shed a light on the local gradient algorithms, which we do in our work.

optimal local models (see Thm. 3.2). As a byproduct, this theoretical result sheds new light on the key update step in the model agnostic meta-learning (MAML) method (Finn et al., 2017), which has a similar but subtly different structure.[2] The subtle difference is that the MAML update obtains the local model by subtracting the gradient evaluated at the *global* model. While MAML was originally proposed as a heuristic, we provide rigorous theoretical guarantees.

⋄ **Loopless LGD: non-uniform SGD applied to our formulation.** We then propose a randomized gradient-based method—*Loopless Local Gradient Descent (L2GD)*—for solving our new formulation (Algorithm 1). This method is, in fact, a non-standard application of SGD to our problem, and can be seen as an instance of SGD with non-uniform sampling applied to the problem of minimizing the sum of two convex functions (Zhao & Zhang, 2015; Gower et al., 2019): the average loss, and the penalty. When the loss function is selected by the randomness in our SGD method, the stochastic gradient step can be *interpreted* as the execution of a single local GD step on each device. Since we set the probability of the loss being sampled to be high, this step is typically repeated multiple times, resulting in a multiple local GD steps. In contrast to standard LGD, the number of local steps is not fixed, but random, and follows a geometric distribution. This mechanism is similar in spirit to how the recently proposed loopless variants of SVRG (Hofmann et al., 2015; Kovalev et al., 2020) work in comparison with the original SVRG (Johnson & Zhang, 2013a; Xiao & Zhang, 2014). Once the penalty is sampled by our method, the resultant SGD step can be interpreted as the execution of an aggregation step. In contrast with standard aggregation, which performs full averaging of the local models, our method merely takes a *step towards averaging*. However, the step is relatively large.

⋄ **Convergence theory.** By adapting the general theory from (Gower et al., 2019) to our setting, we obtain theoretical convergence guarantees assuming that each f_i is L-smooth and μ-strongly convex (see Thm. 4.1). Interestingly, by optimizing the probability of sampling the penalty (we get $p^\star = \frac{\lambda}{\lambda+L}$), which is an indirect way of fixing the *expected number of local steps* to $1+\frac{L}{\lambda}$, we prove an $\frac{2\lambda}{\lambda+L}\frac{L}{\mu}\log\frac{1}{\varepsilon}$ bound on the expected number of communication rounds (see Cor. 4.3). We believe that this is remarkable in several ways. By choosing λ small, we tilt our goal towards pure local models: the number of communication rounds is tending to 0 as $\lambda \to 0$. If $\lambda \to \infty$, the solution our formulation converges to is the optimal global model, and L2GD obtains the communication bound $\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$, which matches the efficiency of GD.

⋄ **What problem do local methods solve?** Noting that L2GD is a (mildly nonstandard) version of LGD, which is a key method most local methods for FL are based on, and noting that, as we show, L2GD solves our new formulation of FL, we offer a new and surprising interpretation of the role of local steps in FL. In particular, the role of local steps in gradient type methods, such as GD, is not to reduce communication complexity, as is generally believed. Indeed, there is no theoretical result supporting this claim in the key heterogeneous data regime. Instead, their role is to steer the method towards finding a mixture of the traditional global and the purely local models. The more local steps are taken, the more we bias the method towards the purely local models. Our new optimization formulation of FL formalizes this as it defines the problem that local methods, in this case L2GD, solve. There is an added benefit here: the more we want our formulation to be biased towards purely local models (i.e., the smaller the penalty parameter λ is), the more local steps does L2GD take, and the better the total communication complexity of L2GD becomes. Hence, despite a lot of research on this topic, our paper provides *the first proof that a local method (e.g., L2GD) can be better than its non-local counterpart (e.g., GD) in terms of total communication complexity in the heterogeneous data setting*. We are able to do this by noting that local methods should better be seen as methods for solving the new FL formulation proposed here.

⋄ **Generalizations: partial participation, local SGD and variance reduction.** We further generalize and improve our method by allowing for (i) stochastic *partial participation* of devices in each communication round,(ii) *subsampling* on each device which means we can perform local SGD steps instead of local GD steps, and (iii) *total variance reduction mechanism* to tackle the variance coming from three sources: locality of the updates induced by non-uniform sampling (already present in L2GD), partial participation and local subsampling. Due to its level of generality, this method, which we call L2SGD++, is presented in the Appendix only, alongside the associated complexity results. In the main body of this paper, we instead present a simplified version thereof, which we call L2SGD+ (Algorithm 3). The convergence theory for it is presented in Thm. 5.1 and Cor. 5.2.

[2]The connection of FL and multi-task meta learning is discussed in (Kairouz et al., 2019), for example.

◇ **Heterogeneous data.** All our methods and convergence results allow for fully heterogeneous data and do not depend on any assumptions on data similarity across the devices.

◇ **Superior empirical performance.** We show through ample numerical experiments that our theoretical predictions can be observed in practice.

3 New Formulation of FL

We now introduce our new formulation for training supervised FL models:

$$\min_{x_1,\ldots,x_n\in\mathbb{R}^d} \{F(x) := f(x) + \lambda\psi(x)\}$$
$$f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x_i), \quad \psi(x) := \frac{1}{2n}\sum_{i=1}^{n} \|x_i - \bar{x}\|^2, \tag{2}$$

where $\lambda \geq 0$ is a penalty parameter, $x_1, \ldots, x_n \in \mathbb{R}^d$ are local models, $x := (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{nd}$ and $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$ is the average of the local models.

Due to the assumptions on f_i we will make in Sec. 3.1, F is strongly convex and hence (2) has a unique solution, which we denote $x(\lambda) := (x_1(\lambda), \ldots, x_n(\lambda)) \in \mathbb{R}^{nd}$. We further let $\bar{x}(\lambda) := \frac{1}{n}\sum_{i=1}^{n} x_i(\lambda)$. We now comment on the rationale behind the new formulation.

Local models ($\lambda = 0$). Note that for each i, $x_i(0)$ solves the *local problem* $\min_{x_i\in\mathbb{R}^d} f_i(x_i)$. That is, $x_i(0)$ is the local model based on data $\mathcal{D}_i$ stored on device i only. This model can be computed by device i without any communication whatsoever. Typically, $\mathcal{D}_i$ is not rich enough for this local model to be useful. In order to learn a better model, one has to take into account the date from other clients as well. This, however, requires communication.

Mixed models ($\lambda \in (0, \infty)$). As λ increases, the penalty $\lambda\psi(x)$ has an increasingly more substantial effect, and communication is needed to ensure that the models are not too dissimilar, as otherwise ψ would be too large.

Global model ($\lambda = \infty$). Let us now look at the limit case $\lambda \to \infty$. Intuitively, this limit case should force the optimal local models to be mutually identical, while minimizing the loss f. In particular, this limit case will solve[3] $\min\left\{f(x) \,:\, x_1, \ldots, x_n \in \mathbb{R}^d,\ x_1 = x_2 = \cdots = x_n\right\}$, which is equivalent to the global formulation (2). Because of this, let us define $x_i(\infty)$ for each i to be the optimal global solution of (1), and let $x(\infty) := (x_1(\infty), \ldots, x_n(\infty))$.

3.1 Technical Preliminaries

We make the following assumption on the functions f_i:

Assumption 3.1 *For each i, the function $f_i : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and μ-strongly convex.*

For $x_i, y_i \in \mathbb{R}^d$, $\langle x_i, y_i\rangle$ denotes the standard inner product and $\|x\| := \langle x_i, x_i\rangle^{1/2}$ is the standard Euclidean norm. For vectors $x = (x_1, \ldots, x_n) \in \mathbb{R}^{nd}$, $y = (y_1, \ldots, y_n) \in \mathbb{R}^{nd}$ we define the standard inner product and norm via $\langle x, y\rangle := \sum_{i=1}^{n}\langle x_i, y_i\rangle$, $\|x\|^2 := \sum_{i=1}^{n}\|x_i\|^2$. Note that the separable structure of f implies that $(\nabla f(x))_i = \frac{1}{n}\nabla f_i(x_i)$, i.e., $\nabla f(x) = \frac{1}{n}(\nabla f_1(x_1), \nabla f_2(x_2), \ldots, \nabla f_n(x_n))$. Hence, the norm of $\nabla f(x) \in \mathbb{R}^{nd}$ decomposes as $\|\nabla f(x)\|^2 = \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x_i)\|^2$.

Note that Assumption 3.1 implies that f is L_f-smooth with $L_f := \frac{L}{n}$ and μ_f-strongly convex with $\mu_f := \frac{\mu}{n}$. Clearly, ψ is convex by construction. It can be shown that ψ is L_ψ-smooth with $L_\psi = \frac{1}{n}$ (see Appendix). We can also easily see that $(\nabla\psi(x))_i = \frac{1}{n}(x_i - \bar{x})$(see Appendix), which implies $\psi(x) = \frac{n}{2}\sum_{i=1}^{n}\|(\nabla\psi(x))_i\|^2 = \frac{n}{2}\|\nabla\psi(x)\|^2$.

[3] If $\lambda = \infty$ and $x_1 = x_2 = \cdots = x_n$ does not hold, we have $F(x) = \infty$. Therefore, we can restrict ourselves on set $x_1 = x_2 = \cdots = x_n$ without loss of generality.

3.2 CHARACTERIZATION OF OPTIMAL SOLUTIONS

Our first result describes the behavior of $f(x(\lambda))$ and $\psi(x(\lambda))$ as a function of λ.

Theorem 3.1 *The function $\lambda \to \psi(x(\lambda))$ is non-increasing, and for all $\lambda > 0$ we have*

$$\psi(x(\lambda)) \leq \frac{f(x(\infty)) - f(x(0))}{\lambda}. \tag{3}$$

Moreover, the function $\lambda \to f(x(\lambda))$ is non-decreasing, and for all $\lambda \geq 0$ we have

$$f(x(\lambda)) \leq f(x(\infty)). \tag{4}$$

Ineq. (3) says that the penalty decreases to zero as λ grows, and hence the optimal local models $x_i(\lambda)$ are increasingly similar as λ grows. The second statement suggest that the loss $f(x(\lambda))$ increases with λ, but never exceeds the optimal global loss $f(x(\infty))$ of the standard FL formulation (1).

We now characterize the optimal local models which connect our model to the MAML framework (Finn et al., 2017), as mentioned in the introduction.

Theorem 3.2 *For each $\lambda > 0$ and $1 \leq i \leq n$ we have*

$$x_i(\lambda) = \overline{x}(\lambda) - \tfrac{1}{\lambda}\nabla f_i(x_i(\lambda)). \tag{5}$$

Further, we have $\sum_{i=1}^{n} \nabla f_i(x_i(\lambda)) = 0$ and $\psi(x(\lambda)) = \frac{1}{2\lambda^2}\left\|\nabla f(x(\lambda))\right\|^2$.

The optimal local models (5) are obtained from the average model by subtracting a multiple of the local gradient. Observe that the local gradients always sum up to zero at optimality. This is obviously true for $\lambda = \infty$, but it is a bit less obvious that this holds for any $\lambda > 0$.

4 L2GD: LOOPLESS LOCAL GD

In this section we describe a new randomized method for solving the formulation (2). Our method is a non-uniform SGD for (2) seen as a 2-sum problem, sampling either ∇f or $\nabla \psi$ to estimate ∇F. Letting $0 < p < 1$, we define a stochastic gradient of F at $x \in \mathbb{R}^{nd}$ as follows

$$G(x) := \begin{cases} \frac{\nabla f(x)}{1-p} & \text{with probability} \quad 1-p \\ \frac{\lambda \nabla \psi(x)}{p} & \text{with probability} \quad p \end{cases}. \tag{6}$$

Clearly, $G(x)$ is an unbiased estimator of $\nabla F(x)$. This leads to the following method for minimizing F, which we call L2GD: $x^{k+1} = x^k - \alpha G(x^k)$. Plugging the formulas for $\nabla f(x)$ and $\nabla \psi(x)$ into (6), and writing the resulting method in a distributed manner, we arrive at Algorithm 1. In each iteration, a coin ξ is tossed and lands 1 with probability p and 0 with probability $1-p$. If $\xi = 0$, all Devices perform one local GD step (7), and if $\xi = 1$, Master shifts each local model towards the average via (8). As we shall see in Sec. 4.2, our theory limits the value of the stepsize α, which has the effect that the ratio $\frac{\alpha\lambda}{np}$ cannot exceed $\frac{1}{2}$. Hence, (8) is a convex combination of x_i^k and $\bar{x}^k$.

Note that Algorithm 1 is only required to communicate when a two consecutive coin tosses land a different value (see the detailed explanation in Sec. C.1 of the appendix). Consequently, the expected number of communication rounds in k iterations of L2GD is $p(1-p)k$.

4.1 THE DYNAMICS OF LOCAL GD AND AVERAGING STEPS

Notice that *the average of the local models does not change* during an aggregation step. Indeed, $\bar{x}^{k+1}$ is equal to $\frac{1}{n}\sum_{i=1}^{n} x_i^{k+1} \overset{(8)}{=} \frac{1}{n}\sum_{i=1}^{n}\left[\left(1 - \frac{\alpha\lambda}{np}\right)x_i^k + \frac{\alpha\lambda}{np}\bar{x}^k\right] = \bar{x}^k$.

If several averaging steps take place in a sequence, the point $a = \bar{x}^k$ in (8) remains unchanged, and each local model x_i^k merely moves along the line joining the initial value of the local model at the start of the sequence and a, with each step pushing x_i^k closer to the average a.

In summary, the more local GD steps are taken, the closer the local models get to the pure local models; and the more averaging steps are taken, the closer the local models get to their average value. The relative number of local GD vs. averaging steps is controlled by the parameter p: the expected # of local GD steps is $\frac{1}{p}$, and the expected number of consecutive aggregation steps is $\frac{1}{1-p}$.

Algorithm 1 L2GD: Loopless Local Gradient Descent

Input: $x_1^0 = \cdots = x_n^0 \in \mathbb{R}^d$, stepsize α, probability p
for $k = 0, 1, \ldots$ **do**
  $\xi = 1$ with probability p and 0 with probability $1-p$
  if $\xi = 0$ **then**
    All Devices $i = 1, \ldots, n$ perform a local GD step:

$$x_i^{k+1} = x_i^k - \tfrac{\alpha}{n(1-p)} \nabla f_i(x_i^k) \tag{7}$$

  else
    Master computes the average $\bar{x}^k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$
    Master for each i computes step towards aggregation

$$x_i^{k+1} = \left(1 - \tfrac{\alpha\lambda}{np}\right) x_i^k + \tfrac{\alpha\lambda}{np} \bar{x}^k \tag{8}$$

  end if
end for

4.2 Convergence theory

We now present our convergence result for L2GD.

Theorem 4.1 *Let Assumption 3.1 hold. If $\alpha \leq \frac{1}{2\mathcal{L}}$, then*

$$\mathbb{E}\left[\left\|x^k - x(\lambda)\right\|^2\right] \leq \left(1 - \tfrac{\alpha\mu}{n}\right)^k \left\|x^0 - x(\lambda)\right\|^2 + \tfrac{2n\alpha\sigma^2}{\mu},$$

where $\mathcal{L} := \frac{1}{n} \max\left\{\frac{L}{1-p}, \frac{\lambda}{p}\right\}$ and $\sigma^2 := \frac{1}{n^2} \sum_{i=1}^{n} \left(\frac{1}{1-p} \|\nabla f_i(x_i(\lambda))\|^2 + \frac{\lambda^2}{p} \|x_i(\lambda) - \overline{x}(\lambda)\|^2\right)$.

Remark 4.2 (Full averaging not supported) *Is a setup such that conditions of Thm. 4.1 are satisfied and the aggregation update (8) is identical to full averaging? This is equivalent requiring $0 < p < 1$ such that $\alpha\lambda = np$. However, we have $\alpha\lambda \leq \frac{\lambda}{2\mathcal{L}} \leq np$, which means that full averaging is not supported by our theory.*

Let us find the parameters p, α which lead to the fastest rate, in terms of the iterations and communication rounds, to push the error within ε-neighborhood of the optimum[4], i.e., to achieve

$$\mathbb{E}\left[\left\|x^k - x(\lambda)\right\|^2\right] \leq \varepsilon \left\|x^0 - x(\lambda)\right\|^2 + \tfrac{2n\alpha\sigma^2}{\mu}. \tag{9}$$

Corollary 4.3 *The value $p^\star = \frac{\lambda}{L+\lambda}$ minimizes both the number of iterations and the expected number of communications for achieving (9). In particular, the optimal number of iterations is $2\frac{L+\lambda}{\mu} \log \frac{1}{\varepsilon}$, and the optimal expected number of communications is $\frac{2\lambda}{\lambda+L} \frac{L}{\mu} \log \frac{1}{\varepsilon}$.*

If we choose $p = p^\star$, then $\frac{\alpha\lambda}{np} = \frac{1}{2}$, and the aggregation rule (8) in Algorithm 1 becomes

$$x_i^{k+1} = \tfrac{1}{2}\left(x_i^k + \bar{x}^k\right) \tag{10}$$

while the local GD step (7) becomes $x_i^{k+1} = x_i^k - \frac{1}{2L} \nabla f_i(x_i^k)$. Notice that while our method does not support full averaging as that is too unstable, (10) suggests that one should take a large step *towards* averaging. As λ get smaller, the solution to the optimization problem (2) will increasingly favour pure local models, i.e., $x_i(\lambda) \to x_i(0) := \arg\min f_i$ for all i as $\lambda \to 0$. Pure local models can be computed without any communication whatsoever and Cor. 4.3 confirms this intuition: the optimal number of communication round decreases to zero as $\lambda \to 0$. On the other hand, as $\lambda \to \infty$, the optimal number of communication rounds converges to $2\frac{L}{\mu} \log \frac{1}{\varepsilon}$, which recovers the performance of GD for finding the globally optimal model (see Fig. 1).

[4] In Sec. 5 we propose a variance reduced algorithm which removes the neighborhood from Thm. 4.1. In that setting, our goal will be to achieve $\mathbb{E}\left[\left\|x^k - x(\lambda)\right\|^2\right] \leq \varepsilon \left\|x^0 - x(\lambda)\right\|^2$.

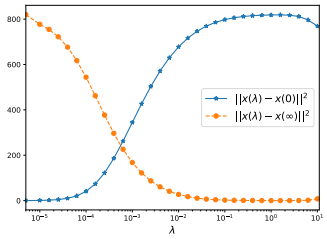

Figure 1: Distance of solution $x(\lambda)$ of (2) to pure local solution $x(0)$ and global solution $x(\infty)$ as a function of λ. Logistic regression on a1a dataset. See Appendix for experimental setup.

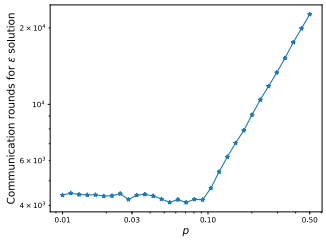

Figure 2: Communication rounds to get $\frac{F(x^k)-F(x^*)}{F(x^0)-F(x^*)} \leq 10^{-5}$ as a function of p with $p^* \approx 0.09$ (for L2SGD+). Logistic regression on a1a dataset with $\lambda = 0.1$; details in the Appendix.

In summary, we recover the communication efficiency of GD for finding the globally optimal model as $\lambda \to \infty$. However, for other values of λ, the communication complexity of L2GD is better and decreases to 0 as $\lambda \to 0$. Hence, our communication complexity result interpolates between the communication complexity of GD for finding the global model and the zero communication complexity for finding the pure local models.

5 Loopless Local SGD with Variance Reduction

As we have seen in Sec. 4.2, L2GD is a specific instance of SGD, thus only converges linearly to a neighborhood of the optimum. In this section, we resolve the mentioned issue by incorporating control variates to the stochastic gradient (Johnson & Zhang, 2013b; Defazio et al., 2014). We go further: we assume that each local objective has a finite-sum structure and propose an algorithm, L2SGD+, which takes *local stochastic* gradient steps, while maintaining (global) linear convergence rate. As a consequence, L2SGD+ is the first local SGD with linear convergence.[5] For convenience, we present variance reduced local GD (i.e., no local subsampling) in the Appendix.

Assumption 5.1 *Assume that f_i has a finite-sum structure: $f_i(x_i) = \frac{1}{m}\sum_{j=1}^m f'_{i,j}(x_i)$. Let $f'_{i,j}$ be convex, L'-smooth while f_i is μ-strongly convex (for each $1 \leq j \leq m, 1 \leq i \leq n$).*

5.1 Convergence theory

We are now ready to present a convergence rate of L2SGD+ (the algorithm, along with the efficient implementation is presented in Appendix C.3).

Theorem 5.1 *Let Assumption 5.1 hold and choose $\alpha = n \min\left\{\frac{(1-p)}{4L'+\mu m}, \frac{p}{4\lambda+\mu}\right\}$. Then the iteration complexity of Algorithm 3 is $\max\left\{\frac{4L'+\mu m}{(1-p)\mu}, \frac{4\lambda+\mu}{p\mu}\right\}\log\frac{1}{\varepsilon}$.*

Next, we find the value of p that yields both the best iteration and communication complexity.

Corollary 5.2 *Both communication and iteration complexity of L2SGD+ are minimized for $p = \frac{4\lambda+\mu}{4\lambda+4L'+(m+1)\mu}$. The resulting iteration complexity is $\left(4\frac{\lambda}{\mu} + 4\frac{L'}{\mu} + m + 1\right)\log\frac{1}{\varepsilon}$, while the communication complexity is $\frac{4\lambda+\mu}{4L'+4\lambda+(m+1)\mu}\left(4\frac{L'}{\mu} + m\right)\log\frac{1}{\varepsilon}$.*

Note that with $\lambda \to \infty$, the communication complexity of L2SGD+ tends to $\left(4\frac{L'}{\mu} + m\right)\log\frac{1}{\varepsilon}$, which is communication complexity of minibatch SAGA to find the globally optimal

[5] We are aware that a linearly converging local SGD (with $\lambda = \infty$) can be obtained as a particular instance of the decoupling method (Mishchenko & Richtárik, 2019). Other variance reduced local SGD algorithms (Liang et al., 2019; Karimireddy et al., 2019; Wu et al., 2019) do not achieve linear convergence.

model (Hanzely & Richtárik, 2019). On the other hand, in the pure local setting ($\lambda = 0$), the communication complexity becomes $\log \frac{1}{\epsilon}$ – this is because the Lyapunov function involves a term that measures the distance of local models, which requires communication to be estimated.

Remark 5.3 *L2SGD+ is the simplest local SGD method with variance reduction. In the Appendix, we present L2SGD++ which allows for 1) an arbitrary number of data points per client and arbitrary local subsampling, 2) partial participation of clients, and 3) local SVRG-like updates of control variates (thus potentially better memory). Lastly, L2SGD++ exploits the complex smoothness structure of the local objectives, resulting in tighter rates.*

6 EXPERIMENTS

In this section, we numerically verify the theoretical claims from this paper. We only present a single experiment here, all remaining ones along with the missing details about the setup are in the Appendix. In particular, the Appendix includes two more experiments. The first one studies how p (communication) influences the convergence of L2SGD+. The second experiment aims to examine the effect of parameter λ on the convergence rate of L2SGD+.

We consider logistic regression problem with LibSVM data (Chang & Lin, 2011). The data were normalized so that $f'_{i,j}$ is 1-smooth for each j, while the local objectives are 10^{-4}-strongly convex. In order to cover a range of possible scenarios, we have chosen a different number of clients for each dataset (see the Appendix). Lastly, the stepsize was always chosen according to Thm. 5.1.

We compare three different methods: L2SGD+, L2GD with local subsampling (L2SGD in the Appendix), and L2GD with local subsampling and control variates constructed for ψ only (L2SGD2 in the Appendix; similar to (Liang et al., 2019)). We expect L2SGD+ to converge to the global optimum linearly, while both L2SGD and L2SGD2 to converge to certain neighborhood. Each method is applied to two objectives constructed by a different split of the data among the devices. For the homogeneous split, we randomly reshuffle the data. For heterogeneous split, we first sort the data based on the labels and then construct the local objectives according to the current order.

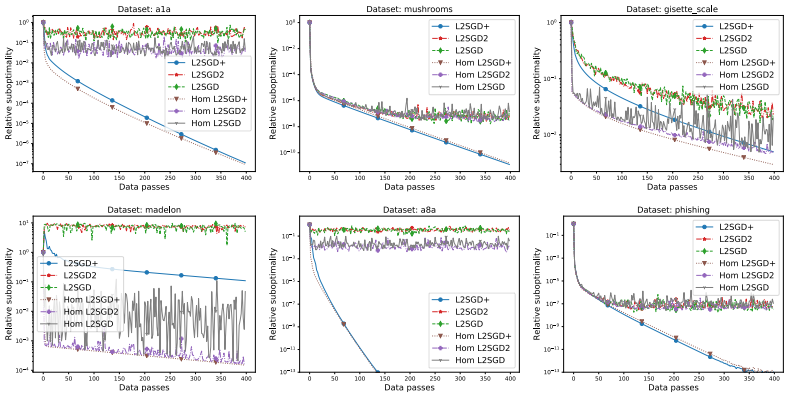

Figure 3: L2SGD+, vs L2SGD vs L2SGD2 with identical stepsize (details in the Appendix).

Fig. 3 demonstrates the importance of variance reduction – it ensures a fast global convergence of L2SGD+, while the neighborhood is slightly smaller for L2SGD2 compared to L2SGD. As predicted, data heterogeneity does not affect the convergence speed of the proposed methods.

REFERENCES

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 1200–1205. ACM, 2017.

Chih-Chung Chang and Chih-Jen Lin. LibSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

L. Corinzia and J. M. Buhmann. Variational federated multi-task learning. *arXiv preprint arXiv:1906.06268*, 2019.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V Le, and et al. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems*, pp. 1223–1231, 2012.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pp. 1646–1654, 2014.

H. Eichner, T. Koren, H. B. McMahan, N. Srebro, and K. Talwar. Semi-cyclic stochastic gradient descent. In *International Conference on Machine Learning*, 2019.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, 2017.

Nidham Gazagnadou, Robert M Gower, and Joseph Salmon. Optimal mini-batch and step sizes for SAGA. In *International Conference on Machine Learning*, 2019.

Eduard Gorbunov, Darina Dvinskikh, and Alexander Gasnikov. Optimal decentralized distributed algorithms for stochastic convex optimization. *arXiv preprint arXiv:1911.07363*, 2019.

Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In *The 23rd International Conference on Artificial Intelligence and Statistics*, 2020.

Robert Mansel Gower, Peter Richtárik, and Francis Bach. Stochastic quasi-gradient methods: variance reduction via Jacobian sketching. *arXiv:1805.02632*, 2018.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5200–5209, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/qian19b.html.

Filip Hanzely and Peter Richtárik. One method to rule them all: Variance reduction for data, parameters and many new methods. *arXiv preprint arXiv:1905.11266*, 2019.

Thomas Hofmann, Aurelien Lucchi, Simon Lacoste-Julien, and Brian McWilliams. Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems*, pp. 2305–2313, 2015.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pp. 315–323, 2013a.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pp. 315–323, 2013b.

Peter Kairouz, H. Brendan McMahan, and et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977v1*, 2019.

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

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. First analysis of local GD on heterogeneous data. In *NeurIPS Workshop on Federated Learning for Data Privacy and Confidentiality*, pp. 1–11, 2019.

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

M. Khodak, M.-F. Balcan, and A. Talwalkar. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems 32*, 2019.

Jakub Konečný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: distributed machine learning for on-device intelligence. *arXiv:1610.02527*, 2016a.

Jakub Konečný, H. Brendan McMahan, Felix Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: strategies for improving communication efficiency. In *NIPS Private Multi-Party Machine Learning Workshop*, 2016b.

Dmitry Kovalev, Samuel Horváth, and Peter Richtárik. Dont jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, 2020.

Guanghui Lan, Soomin Lee, and Yi Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*, pp. 1–48, 2018.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: challenges, methods, and future directions. *arXiv preprint arXiv:1908.07873*, 2019.

Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Variance reduced local SGD with lower communication complexity. *arXiv preprint arXiv:1912.12844*, 2019.

Sulin Liu, Sinno Jialin Pan, and Qirong Ho. Distributed multi-task relationship learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 937–946, 2017.

Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*, 2016.

H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Konstantin Mishchenko and Peter Richtárik. A stochastic decoupling method for minimizing the sum of smooth and non-smooth functions. *arXiv preprint arXiv:1905.11535*, 2019.

Yurii Nesterov. *Introductory lectures on convex optimization: a basic course (Applied Optimization)*. Kluwer Academic Publishers, 2004.

Daniel Peterson, Pallika Kanani, and Virendra J Marathe. Private federated learning with domain adaptation. *arXiv preprint arXiv:1912.06733*, 2019.

Xun Qian, Zheng Qu, and Peter Richtárik. L-SVRG and L-Katyusha with arbitrary sampling. *arXiv preprint arXiv:1906.01481*, 2019a.

Xun Qian, Zheng Qu, and Peter Richtárik. SAGA with arbitrary sampling. In *The 36th International Conference on Machine Learning*, 2019b.

Zheng Qu and Peter Richtárik. Coordinate descent with arbitrary sampling II: Expected separable overapproximation. *Optimization Methods and Software*, 31(5):858–884, 2016. doi: 10.1080/10556788.2016.1190361. URL http://dx.doi.org/10.1080/10556788.2016.1190361.

Sashank J. Reddi, Jakub Konečný, Peter Richtárik, Barnabás Póczos, and Alex Smola. AIDE: fast and communication efficient distributed optimization. *arXiv:1608.06879*, 2016.

Peter Richtárik and Martin Takáč. Parallel coordinate descent methods for big data optimization. *Mathematical Programming*, 156(1-2):433–484, 2016.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4424–4434. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7029-federated-multi-task-learning.pdf.

Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2020.

Weiran Wang, Jialei Wang, Mladen Kolar, and Nathan Srebro. Distributed stochastic multi-task learning with graph regularization. *arXiv preprint arXiv:1802.03830*, 2018.

Zhaoxian Wu, Qing Ling, Tianyi Chen, and Georgios B Giannakis. Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks. *arXiv preprint arXiv:1912.12716*, 2019.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Yu Zhang and Dit-Yan Yeung. A convex formulation for learning task relationships in multi-task learning. *Uncertainty in Artificial Intelligence*, pp. 733–442, 2010.

Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *Proceedings of the 32nd International Conference on Machine Learning, PMLR*, volume 37, pp. 1–9, 2015.

Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.