Local SGD: Unified Theory and New Efficient Methods

Eduard Gorbunov
MIPT, Yandex, Russia
KAUST, Saudi Arabia

Filip Hanzely
KAUST, Saudi Arabia

Peter Richtárik
KAUST, Saudi Arabia

Abstract

We present a unified framework for analyzing local SGD methods in the convex and strongly convex regimes for distributed/federated training of supervised machine learning models. We recover several known methods as a special case of our general framework, including Local-SGD/FedAvg, SCAFFOLD, and several variants of SGD not originally designed for federated learning. Our framework covers both the identical and heterogeneous data settings, supports both random and deterministic number of local steps, and can work with a wide array of local stochastic gradient estimators, including shifted estimators which are able to adjust the fixed points of local iterations for faster convergence. As an application of our framework, we develop multiple novel FL optimizers which are superior to existing methods. In particular, we develop the first linearly converging local SGD method which does not require any data homogeneity or other strong assumptions.

1 Introduction

In this paper we are interested in a centralized distributed optimization problem of the form

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

where n is the number of devices/clients/nodes/workers. We assume that f_i can be represented either as a) an expectation, i.e.,

$$f_i(x) = \mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[f_{\xi_i}(x)\right], \tag{2}$$

where $\mathcal{D}_i$ describes the distribution of data on device i, or b) as a finite sum, i.e.,

$$f_i(x) = \tfrac{1}{m}\sum_{j=1}^m f_{ij}(x). \tag{3}$$

While our theory allows the number of functions m to vary across the devices, for simplicity of exposition, we restrict the narrative to this simpler case.

Federated learning (FL)—an emerging subfield of machine learning (McMahan et al., 2016; Konečný et al., 2016; McMahan et al., 2017)—is traditionally cast as an instance of problem (1) with several idiosyncrasies. First, the number of devices n is very large: tens of thousands to millions. Second, the devices (e.g., mobile phones) are often very heterogeneous in their compute, connectivity, and storage capabilities. The data defining each function f_i reflects the usage patterns of the device owner, and as such, it is either unrelated or at best related only weakly. Moreover, device owners desire to protect their local private data, and for that reason, training needs to take place with the data remaining on the devices. Finally, and this is of key importance for the development in this work, communication among the workers, typically conducted via a trusted aggregation server, is very expensive.

Communication bottleneck. There are two main directions in the literature for tackling the communication cost issue in FL. The first approach consists of algorithms that aim to reduce the number of transmitted bits by applying a carefully chosen gradient compression scheme, such as quantization (Alistarh et al., 2016; Bernstein et al., 2018; Mishchenko et al., 2019; Horváth et al., 2019; Ramezani-Kebrya et al., 2019; Reisizadeh et al., 2020), sparsification (Aji and Heafield, 2017; Lin et al., 2017; Alistarh et al., 2018; Wangni et al., 2018; Wang et al., 2018; Mishchenko et al., 2020), or other more sophisticated strategies (Karimireddy et al., 2019b; Stich and Karimireddy, 2019; Wu et al., 2018; Vogels et al., 2019; Beznosikov et al., 2020; Gorbunov et al., 2020b). The second approach—one that we investigate in this paper—instead focuses on increasing the total amount of local computation in between the

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

communication rounds in the hope that this will reduce the total number of communication rounds needed to build a model of sufficient quality (Shamir et al., 2014; Zhang and Lin, 2015; Reddi et al., 2016; Li et al., 2018; Pathak and Wainwright, 2020). These two approaches, *communication compression* and *local computation*, can be combined for a better practical performance (Basu et al., 2019).

Local first-order algorithms. Motivated by recent development in the field (Zinkevich et al., 2010; McMahan et al., 2016; Stich, 2018; Lin et al., 2018; Liang et al., 2019; Wu et al., 2019; Karimireddy et al., 2019a; Khaled et al., 2020; Woodworth et al., 2020b), in this paper we perform an in-depth and general study of *local first-order algorithms*. Contrasted with zero or higher order local methods, local first order methods perform several gradient-type steps in between the communication rounds. In particular, we consider the following family of methods:

$$x_i^{k+1} = \begin{cases} x_i^k - \gamma g_i^k, & \text{if } c_{k+1} = 0, \\ \frac{1}{n}\sum\limits_{i=1}^{n}\left(x_i^k - \gamma g_i^k\right), & \text{if } c_{k+1} = 1, \end{cases} \tag{4}$$

where x_i^k represents the local variable maintained by the i-th device, g_i^k represents local first order direction[1] and (possibly random) sequence $\{c_k\}_{k\geq 1}$ with $c_k \in \{0,1\}$ encoding the times when communication takes place.

Both the classical `Local-SGD/FedAvg` (McMahan et al., 2016; Stich, 2018; Khaled et al., 2020; Woodworth et al., 2020b) and shifted local `SGD` (Liang et al., 2019; Karimireddy et al., 2019a) methods fall into this category of algorithms. However, most of the existing methods have been analyzed with limited flexibility only, leaving many potentially fruitful directions unexplored. The most important unexplored questions include i) better understanding of the local shift that aims to correct the fixed point of local methods, ii) support for more sophisticated local gradient estimators that allow for importance sampling, variance reduction, or coordinate descent, iii) variable number of local steps, and iv) general theory supporting multiple data similarity types, including identical, heterogeneous and partially heterogeneous (ζ-heterogeneous - defined later).

Consequently, there is a need for a single framework unifying the theory of local stochastic first order methods, ideally one capable of pointing to new and more efficient variants. This is what we do in this work.

Unification of stochastic algorithms. There have been multiple recent papers aiming to unify the theory of first-order optimization algorithms. The closest to our work is the unification of (non-local) stochastic algorithms in (Gorbunov et al., 2020a) that proposes a relatively simple yet powerful framework for analyzing variants of `SGD` that allow for minibatching, arbitrary sampling,[2] variance reduction, subspace gradient oracle, and quantization. We recover this framework as a special case in a non-local regime. Next, a framework for analyzing error compensated or delayed SGD methods was recently proposed in (Gorbunov et al., 2020b). Another relevant approach covers the unification of decentralized `SGD` algorithms (Koloskova et al., 2020), which is able to recover the basic variant of `Local-SGD` as well. While our framework matches their rate for basic `Local-SGD`, we cover a broader range of local methods in this work as we focus on the centralized setting.

1.1 Our Contributions

In this paper, we propose a general framework for analyzing a broad family of local stochastic gradient methods of the form (4). Given that a particular local algorithm satisfies a specific parametric assumption (Assumption 2.3) in a certain scenario, we provide a tight convergence rate of such a method.

Let us give a glimpse of our results and their generality. A local algorithm of the form (4) is allowed to consist of an *arbitrary* local stochastic gradient estimator (see Section 4 for details), a possible *drift/shift* to correct for the non-stationarity of local methods[3] and a fixed or random local loop size. Further, we provide a tight convergence rate in both the identical and heterogeneous data regimes for strongly (quasi) convex and convex objectives. Consequently, our framework is capable of:

- **Recovering known optimizers along with their tight rates.** We recover multiple known local optimizers as a special case of our general framework, along with their convergence rates (up to small constant factors). This includes `FedAvg/Local-SGD` (McMahan et al., 2016; Stich, 2018) with currently the best-known convergence rate (Khaled et al., 2020; Woodworth et al., 2020b; Koloskova et al., 2020; Woodworth et al., 2020a) and `SCAFFOLD` (Karimireddy et al., 2019a). Moreover, in a special case we recover a general framework for an-

[1] Vector g_i^k can be a simple unbiased estimator of $\nabla f_i(x_i^k)$, but can also involve a local "shift" designed to correct the (inherently wrong) fixed point of local methods. We elaborate on this point later.

[2] A tight convergence rate given any sampling strategy and any smoothness structure of the objective.

[3] Basic local algorithms such as `FedAvg/Local-SGD` or `FedProx` (Li et al., 2018) have incorrect fixed points (Pathak and Wainwright, 2020). To eliminate this issue, a strategy of adding an extra "drift" or "shift" to the local gradient has been proposed recently (Liang et al., 2019; Karimireddy et al., 2019a).

alyzing non-local SGD method developed in (Gorbunov et al., 2020a), and consequently we recover multiple variants of SGD with and without variance reduction, including SAGA (Defazio et al., 2014), L-SVRG (Kovalev et al., 2019), SEGA (Hanzely et al., 2018), gradient compression methods (Mishchenko et al., 2019; Horváth et al., 2019) and many more.

• **Filling missing gaps for known methods.** Many of the recovered optimizers have only been analyzed under specific and often limiting circumstances and regimes. Our framework allows us to extend known methods into multiple hitherto unexplored settings. For instance, for each (local) method our framework encodes, we allow for a random/fixed local loop size, identical/heterogeneous/ζ-heterogeneous data (introduced soon), and convex/strongly convex objective.

• **Extending the established optimizers.** To the best of our knowledge, none of the known local methods have been analyzed under arbitrary smoothness structure of the local objectives[4] and consequently, our framework is the first to allow for the local stochastic gradient to be constructed via importance (possibly minibatch) sampling. Next, we allow for a local loop with a random length, which is a new development contrasting with the classical fixed-length regime. We discuss advantages of of the random loop in Section 3.

• **New efficient algorithms.** Perhaps most importantly, our framework is powerful enough to point to a range of novel methods. A notable example is S-Local-SVRG, which is a local variance reduced SGD method able to learn the optimal drift. This is the first time that local variance reduction is successfully combined with an on-the-fly learning of the local drift. Consequently, this is the first method which enjoys a linear convergence rate to the exact optimum (as opposed to a neighborhood of the solution only) without any restrictive assumptions and is thus superior in theory to the convergence of all existing local first order methods. We also develop another linearly converging method: S*-Local-SGD*. Albeit not of practical significance as it depends on the a-priori knowledge of the optimal solution x^*, it is of theoretical interest as it enabled us to discover S-Local-SVRG. See Table 2 which summarizes all our complexity results.

Notation. Due to its generality, our paper is heavy in notation. For the reader's convenience, we present a notation table in Sec. A of the appendix.

2 Our Framework

In this section we present the main result of the paper. Let us first introduce the key assumptions that we impose on our objective (1). We start with a relaxation of μ-strong convexity.

Assumption 2.1 ((μ, x^*)-strong quasi-convexity)**.** *Let x^* be a minimizer of f. We assume that f_i is (μ, x^*)-strongly quasi-convex for all $i \in [n]$ with $\mu \geq 0$, i.e. for all $x \in \mathbb{R}^d$:*

$$f_i(x^*) \geq f_i(x) + \langle \nabla f_i(x), x^* - x\rangle + \tfrac{\mu}{2}\|x - x^*\|^2. \quad (5)$$

Next, we require classical L-smoothness[5] of local objectives, or equivalently, L-Lipschitzness of their gradients.

Assumption 2.2 (L-smoothness)**.** *Functions f_i are L-smooth for all $i \in [n]$ with $L \geq 0$, i.e.,*

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathbb{R}^d. \quad (6)$$

In order to simplify our notation, it will be convenient to introduce the notion of virtual iterates x^k defined as a mean of the local iterates (Stich and Karimireddy, 2019): $x^k \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n x_i^k$. Despite the fact that x^k is being physically computed only for k for which $c_k = 1$, virtual iterates are a very useful tool facilitating the convergence analysis. Next, we shall measure the discrepancy between the local and virtual iterates via the quantity V_k defined as $V_k \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \|x_i^k - x^k\|^2$.

We are now ready to introduce the parametric assumption on both stochastic gradients g_i^k and function f. This is a non-trivial generalization of the assumption from (Gorbunov et al., 2020a) to the class of local stochastic methods of the form (4), and forms the heart of this work.[6]

Assumption 2.3 (Key parametric assumption)**.** *Assume that for all $k \geq 0$ and $i \in [n]$, local stochastic directions g_i^k satisfy*

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}_k\left[g_i^k\right] = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k), \quad (7)$$

where $\mathbf{E}_k[\cdot]$ defines the expectation w.r.t. randomness coming from the k-th iteration only. Further, assume that there exist non-negative constants $A, A', B, B', C, C', F, F', G, H, D_1, D_1', D_2, D_3 \geq 0, \rho \in$

[4] By this we mean that function $f_{i,j}$ from (3) is $\mathbf{M}_{i,j}$-smooth with $\mathbf{M}_{i,j} \in \mathbb{R}^{d\times d}, \mathbf{M}_{i,j} \succeq 0$, i.e., for all $x, y \in \mathbb{R}^d$ we have $f_{i,j}(x) \leq f_{i,j}(y) + \langle \nabla f_{i,j}(y), x - y\rangle + \frac{1}{2}(x - y)^\top \mathbf{M}_{i,j}(x - y)$. As an example, logistic regression possesses naturally such a structure with matrices $\mathbf{M}_{i,j}$ of rank 1.

[5] While we require L-smoothness of f_i to establish the main convergence theorem, some of the parameters of As. 2.3 can be tightened considering a more complex smoothness structure of the local objective.

[6] Recently, the assumption from (Gorbunov et al., 2020a) was generalized in a different way to cover the class of the methods with error compensation and delayed updates (Gorbunov et al., 2020b).

$(0,1]$ *and a sequence of (possibly random) variables* $\{\sigma_k^2\}_{k\geq 0}$ *such that*

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k\|^2\right] \leq 2A\mathbf{E}\left[f(x^k)-f(x^*)\right] + B\mathbf{E}\left[\sigma_k^2\right] + F\mathbf{E}\left[V_k\right] + D_1, \tag{8}$$

$$\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k\right\|^2\right] \leq 2A'\mathbf{E}\left[f(x^k)-f(x^*)\right] + B'\mathbf{E}\left[\sigma_k^2\right] + F'\mathbf{E}\left[V_k\right] + D_1', \tag{9}$$

$$\mathbf{E}\left[\sigma_{k+1}^2\right] \leq (1-\rho)\mathbf{E}\left[\sigma_k^2\right] + 2C\mathbf{E}\left[f(x^k)-f(x^*)\right] + G\mathbf{E}\left[V_k\right] + D_2, \tag{10}$$

$$2L\sum_{k=0}^K w_k\mathbf{E}[V_k] \leq \frac{1}{2}\sum_{k=0}^K w_k\mathbf{E}\left[f(x^k)-f(x^*)\right] + 2LH\mathbf{E}\sigma_0^2 + 2LD_3\gamma^2 W_K, \tag{11}$$

where sequences $\{W_K\}_{K\geq 0}$, $\{w_k\}_{k\geq 0}$ *are defined as*

$$W_K \overset{def}{=} \sum_{k=0}^K w_k, \quad w_k \overset{def}{=} \frac{1}{\left(1-\min\left\{\gamma\mu,\frac{\rho}{4}\right\}\right)^{k+1}}, \tag{12}$$

Admittedly, with its many parameters (whose meaning will become clear from the rest of the paper), As. 2.3 is not easy to parse on first reading. Several comments are due at this point. First, while the complexity of this assumption may be misunderstood as being problematic, the opposite is true. This assumption enables us to prove a single theorem (Thm. 2.1) capturing the convergence behavior, in a tight manner, of all local first-order methods described by our framework (4). So, the parametric and structural complexity of this assumption is paid for by the unification aspect it provides. Second, for each specific method we consider in this work, we *prove* that As. 2.3 is satisfied, and each such proof is based on much simpler and generally accepted assumptions. So, As. 2.3 should be seen as a "meta-assumption" forming an intermediary and abstract step in the analysis, one revealing the structure of the inequalities needed to obtain a general and tight convergence result for local first-order methods. We dedicate the rest of the paper to explaining these parameters and to describing the algorithms and the associate rates their combination encodes. We are now ready to present our main convergence result.

Theorem 2.1. *Let As. 2.1, 2.2 and 2.3 be satisfied and assume the stepsize satisfies* $0 < \gamma \leq \min\left\{\frac{1}{2(A'+\frac{4CB'}{3\rho})}, \frac{L}{F'+\frac{4GB'}{3\rho}}\right\}$. *Define* $\overline{x}^K \overset{def}{=} \frac{1}{W_K}\sum_{k=0}^K w_k x^k$, $\Phi^0 \overset{def}{=} \frac{2\|x^0-x^*\|^2+\frac{8B'}{3\rho}\gamma^2\mathbf{E}\sigma_0^2+4LH\gamma\mathbf{E}\sigma_0^2}{\gamma}$ *and* $\Psi^0 \overset{def}{=} 2\left(D_1' + \frac{4B'}{3\rho}D_2 + 2L\gamma D_3\right)$. *Let* $\theta \overset{def}{=} 1 - \min\left\{\gamma\mu, \frac{\rho}{4}\right\}$. *Then if* $\mu > 0$, *we have*

$$\mathbf{E}\left[f(\overline{x}^K)\right] - f(x^*) \leq \theta^K\Phi^0 + \gamma\Psi^0, \tag{13}$$

and in the case when $\mu = 0$, *we have*

$$\mathbf{E}\left[f(\overline{x}^K)\right] - f(x^*) \leq \frac{\Phi^0}{K} + \gamma\Psi^0. \tag{14}$$

As already mentioned, Thm. 2.1 serves as a general, unified theory for local stochastic gradient algorithms. The strongly convex case provides a linear convergence rate up to a specific neighborhood of the optimum. On the other hand, the weakly convex case yields an $\mathcal{O}(K^{-1})$ convergence rate up to a particular neighborhood. One might easily derive $\mathcal{O}(K^{-1})$ and $\mathcal{O}(K^{-2})$ convergence rates to the exact optimum in the strongly and weakly convex case, respectively, by using a particular decreasing stepsize rule. The next corollary gives an example of such a result in the strongly convex scenario, where the estimate of D_3 does not depend on the stepsize γ. A detailed result that covers all cases is provided in Section D.2 of the appendix.

Corollary 2.1. *Consider the setup from Thm. 2.1 and by* $\frac{1}{\nu}$ *denote the resulting upper bound on* γ.[7] *Suppose that* $\mu > 0$ *and* D_3 *does not depend on* γ. *Let*

$$\gamma = \min\left\{\frac{1}{\nu}, \frac{\ln\left(\max\left\{2,\min\left\{\frac{\Upsilon_1\mu^2K^2}{\Upsilon_2},\frac{\Upsilon_1\mu^3K^3}{\Upsilon_3}\right\}\right\}\right)}{\mu K}\right\},$$

where $\Upsilon_1 = 2\|x^0-x^*\|^2 + \frac{8B'\mathbf{E}\sigma_0^2}{3\nu^2\rho} + \frac{4LH\mathbf{E}\sigma_0^2}{\nu}$, $\Upsilon_2 = 2D_1' + \frac{4B'D_2}{3\rho}$, $\Upsilon_3 = 4LD_3$. *Then, the procedure* (4) *achieves*

$$\mathbf{E}\left[f(\overline{x}^K)\right] - f(x^*) \leq \varepsilon$$

as long as

$$K \geq \widetilde{\mathcal{O}}\left(\left(\frac{1}{\rho}+\frac{\nu}{\mu}\right)\log\left(\frac{\nu\Upsilon_1}{\varepsilon}\right) + \frac{\Upsilon_2}{\mu\varepsilon} + \sqrt{\frac{\Upsilon_3}{\mu^2\varepsilon}}\right).$$

Remark 2.1. *Admittedly, Thm. 2.1 does not yield the tightest known convergence rate in the heterogeneous setup under As. 2.1. Specifically, the neighborhood to which* `Local-SGD` *converges can be slightly smaller (Koloskova et al., 2020). While we provide a tighter theory that matches the best-known results, we have deferred it to the appendix for the sake of clarity. In particular, to get the tightest rate, one shall replace the bound on the second moment of the stochastic direction (8) with two analogous bounds – first one for the variance and the second one for the squared expectation. See As. E.1 for details. Fortunately, Thm. 2.1 does not need to change as it does not require parameters from (8); these are only used later to derive* D_3, H, γ *based on the data type. Therefore, only a few extra parameters should be determined in the specific scenario to get the tightest rate.*

[7] In order to get tight estimate of D_3 and H, we will impose further bounds on γ (see Tbl. 1). Assume that these extra bounds are included in parameter h.

Remark 2.2. *As we show in the appendix when looking at particular special cases, local gradient methods are only as good as their non-local counterparts (i.e., when $\tau = 1$) in terms of the communication complexity in the fully heterogeneous setup. Furthermore, the non-local methods outperform local ones in terms of computation complexity. While one might think that this observation is a byproduct of our analysis, our observations are supported by findings in recent literature on this topic (Karimireddy et al., 2019a; Khaled et al., 2020). To rise to the defense of local methods, we remark that they might be preferable to their non-local cousins in the homogeneous data setup (Woodworth et al., 2020b) or for personalized federated learning (Hanzely and Richtárik, 2020).*

The parameters that drive both the convergence speed and the neighborhood size are determined by As. 2.3. In order to see through the provided rates, we shall discuss the value of these parameters in various scenarios. In general, we would like to have $\rho \in (0, 1]$ as large as possible, while all other parameters are desired to be small so as to make the inequalities as tight as possible.

Let us start with studying data similarity and inner loop type as these can be decoupled from the type of the local direction that the method (4) takes.

3 Data Similarity and Local Loop

We now explain how our framework supports fixed and random local loop, and several data similarity regimes.

Local loop. Our framework supports *local loop of a fixed length* $\tau \geq 1$ (i.e., we support local methods performing τ local iterations in between communications). This option, which is the de facto standard for local methods in theory and practice (McMahan et al., 2016), is recovered by setting $c_{a\tau} = 1$ for all non-negative integers a and $c_k = 0$ for k that are not divisible by τ in (4). However, our framework also captures the very rarely considered *local loop with a random length.* We recover this when c_k are random samples from the Bernoulli distribution $\mathrm{Be}(p)$ with parameter $p \in (0, 1]$.

Data similarity. We look at various possible data similarity regimes. The first option we consider is the fully heterogeneous setting where we do not assume any similarity between the local objectives whatsoever. Secondly, we consider the identical data regime with $f_1 = \ldots = f_n$. Lastly, we consider the ζ-heterogeneous data setting, which bounds the dissimilarity between the full and the local gradients (Woodworth et al., 2020a) (see Def. 3.1).

Definition 3.1 (ζ-heterogeneous functions)**.** *We say that functions $f_1, \ldots, f_n$ are ζ-heterogeneous for some $\zeta \geq 0$ if the following inequality holds for all $x \in \mathbb{R}^d$:*

$$\frac{1}{n}\sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f(x)\|^2 \leq \zeta^2. \tag{15}$$

The ζ-heterogeneous data regime recovers the heterogeneous data for $\zeta = \infty$ and identical data for $\zeta = 0$.

In Sec. E of the appendix, we show that the local loop type and the data similarity type affect parameters H and D_3 from As. 2.3 only. However, in order to obtain an efficient bound on these parameters, we impose additional constraints on the stepsize γ. While we do not have space to formally state our results in the main body, we provide a comprehensive summary in Tbl. 1.

Methods with a random loop communicate once per p^{-1} iterations on average, while the fixed loop variant communicates once every τ iterations. Consequently, we shall compare the two loop types for $\tau = p^{-1}$. In such a case, parameters D_3 and H and the extra conditions on stepsize γ match exactly, meaning that the loop type does not influence the convergence rate. Having said that, random loop choice provides more flexibility compared to the fixed loop. Indeed, one might want the local direction g_i^k to be synchronized with the communication time-stamps in some special cases. However, our framework does not allow such synchronization for a fixed loop since we assume that the local direction g_i^k follows some stationary distribution over stochastic gradients. The random local loop comes in handy here; the random variable that determines the communication follows a stationary distribution, thus possibly synchronized with the local computations.

4 Local Stochastic Direction

This section discusses how the choice of g_i^k allows us to obtain the remaining parameters from As. 2.3 that were not covered in the previous section. To cover the most practical scenarios, we set g_i^k to be a difference of two components $a_i^k, b_i^k \in \mathbb{R}^d$, which we explain next. We stress that the construction of g_i^k is very general: we recover various state-of-the-art methods along with their rates while covering many new interesting algorithms. We will discuss this in more detail in Sec. 5.

4.1 Unbiased local gradient estimator a_i^k

The first component of the local direction that the method (4) takes is a_i^k – an unbiased, possibly variance reduced, estimator of the local gradient, i.e., $\mathbf{E}_k[a_i^k] = \nabla f_i(x_i^k)$. Besides the unbiasedness, a_i^k is allowed to be anything that satisfies the parametric recursive relation from (Gorbunov et al., 2020a), which tightly covers many variants of `SGD` including non-uniform, minibatch, and variance reduced stochastic

Table 1: The effect of data similarity and local loop on As. 2.3. Constant factors are ignored. Homogeneous data are recovered as a special case of ζ-heterogeneous data with $\zeta = 0$. Heterogeneous case is slightly loose in light of Remark 2.1. If one replaces the bound on the second moments (8) with a analogous bound on variance squared expectation (see As. E.1), the bounds on γ, D_3 and H will have $(\tau - 1)$ times better dependence on the variance parameters (or $\frac{1-p}{p}$ times for the random loop). See Sec. E.1.1 and E.2.1 of appendix for more details.

Data	Loop	Extra upper bounds on γ	D_3	H
het	fixed	$\frac{1}{\tau\mu}, \frac{1}{\tau\sqrt{\left(F+\frac{BG}{\rho(1-\rho)}\right)}}, \frac{1}{\tau\sqrt{2L\left(A+\frac{BC}{\rho(1-\rho)}\right)}}$	$(\tau-1)^2\left(D_1+\frac{BD_2}{\rho}\right)$	$\frac{B(\tau-1)^2\gamma^2}{\rho}$
ζ-het	fixed	$\frac{1}{\tau\mu}, \frac{1}{\sqrt{\tau\left(F+\frac{BG}{\rho(1-\rho)}\right)}}, \frac{1}{\sqrt{L\tau\left(A+\frac{BC}{\rho(1-\rho)}\right)}}$	$(\tau-1)\left(D_1+\frac{\zeta^2}{\gamma\mu}+\frac{BD_2}{\rho}\right)$	$\frac{B(\tau-1)\gamma^2}{\rho}$
het	random	$\frac{p}{\mu}, \frac{p}{\sqrt{(1-p)F}}, \frac{p\sqrt{\rho(1-\rho)}}{\sqrt{BG(1-p)}}, \frac{p}{\sqrt{L(1-p)\left(A+\frac{BC}{\rho(1-\rho)}\right)}}$	$\frac{(1-p)\left(D_1+\frac{BD_2}{\rho}\right)}{p^2}$	$\frac{B(1-p)\gamma^2}{p^2\rho}$
ζ-het	radnom	$\frac{p}{\mu}, \sqrt{\frac{p}{F(1-p)}}, \sqrt{\frac{p\rho(1-\rho)}{BG(1-p)}}, \sqrt{\frac{p}{L(1-p)\left(A+\frac{BC}{\rho(1-\rho)}\right)}}$	$\frac{(1-p)}{p}\left(D_1+\frac{\zeta^2}{\gamma\mu}+\frac{BD_2}{\rho}\right)$	$\frac{B(1-p)\gamma^2}{p\rho}$

gradient. The parameters of such a relation are capable of encoding both the general smoothness structure of the objective and the gradient estimator's properties that include a diminishing variance, for example. We state the adapted version of this recursive relation as As. 4.1.

Assumption 4.1. *Let the unbiased local gradient estimator a_i^k be such that*

$$\mathbf{E}_k\left[\|a_i^k - \nabla f_i(x^*)\|^2\right] \leq 2A_i D_{f_i}(x_i^k, x^*) + B_i\sigma_{i,k}^2 + D_{1,i},$$
$$\mathbf{E}_k\left[\sigma_{i,k+1}^2\right] \leq (1-\rho_i)\sigma_{ik}^2 + 2C_i D_{f_i}(x_i^k, x^*) + D_{2,i}$$

for $A_i \geq 0, B_i \geq 0, D_{1,i} \geq 0, 0 \leq \rho_i \leq 1, C_i \geq 0, D_{2,i} \geq 0$ and a non-negative sequence $\{\sigma_{i,k}^2\}_{k=0}^\infty$.[8]

Note that the parameters of As. 4.1 can be taken directly from (Gorbunov et al., 2020a) and offer a broad range of unbiased local gradient estimators a_i^k in different scenarios. The most interesting setups covered include minibatching, importance sampling, variance reduction, all either under the classical smoothness assumption or under a uniform bound on the stochastic gradient variance.

Our next goal is to derive the parameters of As. 2.3 from the parameters of As. 4.1. However, let us first discuss the second component of the local direction – the local shift b_i^k.

4.2 Local shift b_i^k

The local update rule (4) can include the local shift/drift b_i^k allowing us to eliminate the infamous non-stationarity of the local methods. The general requirement for the choice of b_i^k is so that it sums up to zero ($\sum_{i=1}^n b_i^k = 0$) to avoid unnecessary extra bias. For the sake of simplicity (while maintaining generality), we will consider three choices of b_i^k – zero, ideal shift ($= \nabla f_i(x^*)$) and on-the-fly shift via a possibly outdated local stochastic non-variance reduced gradient estimator that satisfies a similar bound as As. 4.1.

Assumption 4.2. *Consider the following choices:*
Case I: $b_i^k = 0$,
Case II: $b_i^k = \nabla f_i(x^*)$,
Case III: $b_i^k = h_i^k - \frac{1}{n}\sum_{i=1}^n h_i^k$ where $h_i^k \in \mathbb{R}^d$ is a delayed local gradient estimator defined recursively as

$$h_i^{k+1} = \begin{cases} h_i^k & \text{with probability } 1-\rho_i' \\ l_i^k & \text{with probability } \rho_i' \end{cases},$$

where $0 \leq \rho_i' \leq 1$ and $l_i^k \in \mathbb{R}^d$ is an unbiased non-variance reduced possibly stochastic gradient estimator of $\nabla f_i(x^k)$ such that for some $A_i', D_{3,i} \geq 0$ we have

$$\mathbf{E}_k\left[\|l_i^k - \nabla f_i(x^*)\|^2\right] \leq 2A_i' D_{f_i}(x_i^k, x^*) + D_{3,i}. \quad (16)$$

Let us look closer at Case III as this one is the most interesting. Note that what we assume about l_i^k (i.e., (16)) is essentially a variant of As. 4.2 with $\sigma_{i,k}^2$ parameters set to zero. This is achievable for a broad range of non-variance reduced gradient estimators that includes minibatching and importance sampling (Gower et al., 2019). An intuitive choice of l_i^k is to set it to a_i^k given that a_i^k is not variance reduced. In such a case, the scheme (4) reduces to `SCAFFOLD` (Karimireddy et al., 2019a) along with its rate.

However, our framework can do much more beyond this example. First, we cover the local variance reduced gradient a_i^k with l_i^k constructed as its non-variance reduced part. In such a case, the neighborhood of the optimum from Thm. 2.1 to which the method (4) converges shrinks. There is a way to get rid of this

[8] By $D_{f_i}(x_i^k, x^k)$ we mean Bregman distance between x_i^k, x^k defined as $D_{f_i}(x_i^k, x^k) \overset{\text{def}}{=} f_i(x_i^k) - f_i(x^k) - \langle\nabla f_i(x^k), x_i^k - x^k\rangle$.

neighborhood, noticing that l_i^k is used only once in a while. Indeed, the combination of the full local gradient l_i^k together with the variance reduced a_i^k leads to a linear rate in the strongly (quasi) convex case or $\mathcal{O}(K^{-1})$ rate in the weakly convex case. We shall remark that the variance reduced gradient might require a sporadic computation of the full local gradient – it makes sense to synchronize it with the update rule for h_i^k. In such a case, the computation of l_i^k is for free. We have just described the `S-Local-SVRG` method (Algorithm 6).

4.3 Parameters of Assumption 2.3

We proceed with a key lemma that provides us with the remaining parameters of As. 2.3 that were not covered in Sec. 3. These parameters will be chosen purely based on the selection of a_i^k and b_i^k discussed earlier.

Lemma 4.1. *For all $i \in [n]$ suppose that a_i^k satisfies As. 4.1, while b_i^k was chosen as per As. 4.2. Then,* (8), (9) *and* (10) *hold with*

$$A = 4\max_i A_i, B = 2, F = 4L\max_i A_i,$$

$$D_1 = \begin{cases} \frac{2}{n}\sum_{i=1}^n \left(D_{1,i} + \|\nabla f_i(x^*)\|^2\right) & \textit{Case I,} \\ \frac{2}{n}\sum_{i=1}^n D_{1,i} & \textit{Case II, III,} \end{cases}$$

$$B' = \frac{1}{n}, F' = \frac{2L\max_i A_i}{n} + 2L^2, D_1' = \frac{1}{n^2}\sum_{i=1}^n D_{1,i}$$

$$A' = \frac{2\max_i A_i}{n} + L, G = {}^{CL}\!/_2,$$

$$\rho = \begin{cases} \min_i \rho_i & \textit{Case I, II,} \\ \min_i \min\{\rho_i, \rho_i'\} & \textit{Case III,} \end{cases}$$

$$D_2 = \begin{cases} \frac{2}{n}\sum\limits_{i=1}^n B_i D_{2,i}, & \textit{Case I, II,} \\ \frac{1}{n}\sum\limits_{i=1}^n (2B_i D_{2,i} + \rho_i' D_{3,i}) & \textit{Case III,} \end{cases}$$

$$C = \begin{cases} 4\max_i\{B_i C_i\} & \textit{Case I, II,} \\ 4\max_i\{B_i C_i\} + 4\max_i\{\rho_i' A_i'\} & \textit{Case III.} \end{cases}$$

We have just broken down the parameters of As. 2.3 based on the optimization objective and the particular instance of (4). However, it might still be hard to understand particular rates based on these choices. In the appendix, we state a range of methods and decouple their convergence rates. A summary of the key parameters from As. 2.3 is provided in Tbl. 7.

5 Special Cases

Our theory covers a broad range of local stochastic gradient algorithms. While we are able to recover multiple known methods along with their rates, we also introduce several new methods along with extending the analysis of known algorithms. As already mentioned, our theory covers convex and strongly convex cases, identical and heterogeneous data regimes. From the algorithmic point of view, we cover the fixed and random loop, various shift types, and arbitrary local stochastic gradient estimator. We stress that our framework gives a tight convergence rate under any circumstances.

While we might not cover all of these combinations in a deserved detail, we thoroughly study a subset of them in Sec. G of the appendix. An overview of these methods is presented in Tbl. 2 together with their convergence rates in the strongly convex case (see Tbl. 4 in the appendix for the rates in the weakly convex setting). Next, we describe a selected number of special cases of our framework.

- **Non-local stochastic methods.** Our theory recovers a broad range of non-local stochastic methods. In particular, if $n = 1$, we have $V_k = 0$, and consequently we can choose $A = A', B = B', D_1 = D_1', F = F' = G = H = D_3 = 0$. With such a choice, our theory matches[9] the general analysis of stochastic gradient methods from (Gorbunov et al., 2020a) for $\tau = 1$. Consequently, we recover a broad range of algorithms as a special case along with their convergence guarantees, namely `SGD` (Robbins and Monro, 1951) with its best-known rate on smooth objectives (Nguyen et al., 2018; Gower et al., 2019), variance reduced finite sum algorithms such as `SAGA` (Defazio et al., 2014), `SVRG` (Johnson and Zhang, 2013), `L-SVRG` (Hofmann et al., 2015; Kovalev et al., 2019), variance reduced subspace descent methods such as `SEGA/SVRCD` (Hanzely et al., 2018; Hanzely and Richtárik, 2019), quantized methods (Mishchenko et al., 2019; Horváth et al., 2019) and others.

- **"Star"-shifted local methods.** As already mentioned, local methods have inherently incorrect fixed points (Pathak and Wainwright, 2020); and one can fix these by shifting the local gradients. Star-shifted local methods employ the ideal stationary shift using the local gradients at the optimum $b_i^k = \nabla f_i(x^*)$ (i.e., Case II from As. 4.2) and serve as a transition from the plain local methods (Case I from As. 4.2) to the local methods that shift using past gradients such as `SCAFFOLD` (Case III from As. 4.2). In the appendix, we present two such methods: `S*-Local-SGD` (Algorithm 3) and `S*-Local-SGD*` (Algorithm 5). While being impractical in most cases since $\nabla f_i(x^*)$ is not known, star-shifted local methods give new insights into the role and effect of the shift for local algorithms. Specifically, these methods enjoy superior convergence rate when compared to methods without local shift (Case I) and methods with a shift constructed from observed gradients (Case III), while their rate serves as an aspiring goal for local methods in general. For-

[9] Up to the non-smooth regularization/proximal steps and small constant factors.

Table 2: A selection of methods that can be analyzed using our framework, which we detail in the appendix. A choice of a_i^k, b_i^k and l_i^k is presented along with the established complexity bounds (= number of iterations to find such $\hat{x}$ that $\mathbf{E}[f(\hat{x}) - f(x^*)] \le \varepsilon$) and a specific setup under which the methods are analyzed. For Algorithms 1-4 we suppress constants and $\log \frac{1}{\varepsilon}$ factors. Since Algorithms 5 and 6 converge linearly, we suppress constants only while keeping $\log \frac{1}{\varepsilon}$ factors. All rates are provided in the **strongly convex** setting. UBV stands for the "Uniform Bound on the Variance" of local stochastic gradient, which is often assumed when f_i is of the form (2). ES stands for the "Expected Smoothness" (Gower et al., 2019), which does not impose any extra assumption on the objective/noise, but rather can be derived given the sampling strategy and the smoothness structure of f_i. Consequently, such a setup allows us to obtain local methods with importance sampling. Next, the simple setting is a special case of ES when we uniformly sample a single index on each node each iteration. ♣: `Local-SGD` methods have never been analyzed under ES assumption. Notation: σ^2 – averaged (within nodes) uniform upper bound for the variance of local stochastic gradient, σ_*^2 – averaged variance of local stochastic gradients at the solution, $\zeta_*^2 \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^*)\|^2$, $\max L_{ij}$ – the worst smoothness of $f_{i,j}, i \in [n], j \in [m]$, $\mathcal{L}$ – the worst ES constant for all nodes.

Method	a_i^k, b_i^k, l_i^k	Complexity	Setting	Sec
`Local-SGD`, Alg. 1 (Woodworth et al., 2020a)	$f_{\xi_i}(x_i^k), 0, -$	$\frac{L}{\mu} + \frac{\sigma^2}{n\mu\varepsilon} + \sqrt{\frac{L\tau(\sigma^2+\tau\zeta^2)}{\mu^2\varepsilon}}$	UBV, ζ-Het	G.1.1
`Local-SGD`, Alg. 1 (Koloskova et al., 2020)	$f_{\xi_i}(x_i^k), 0, -$	$\frac{\tau L}{\mu} + \frac{\sigma^2}{n\mu\varepsilon} + \sqrt{\frac{L(\tau-1)(\sigma^2+(\tau-1)\zeta_*^2)}{\mu^2\varepsilon}}$	UBV, Het	G.1.1
`Local-SGD`, Alg. 1 (Khaled et al., 2020)♣	$f_{\xi_i}(x_i^k), 0, -$	$\frac{L+\mathcal{L}/n+\sqrt{(\tau-1)L\mathcal{L}}}{\mu} + \frac{\sigma_*^2}{n\mu\varepsilon}$ $+\frac{L\zeta^2(\tau-1)}{\mu^2\varepsilon} + \sqrt{\frac{L(\tau-1)(\sigma_*^2+\zeta_*^2)}{\mu^2\varepsilon}}$	ES, ζ-Het	G.1.2
`Local-SGD`, Alg. 1 (Khaled et al., 2020)♣	$f_{\xi_i}(x_i^k), 0, -$	$\frac{L\tau+\mathcal{L}/n+\sqrt{(\tau-1)L\mathcal{L}}}{\mu} + \frac{\sigma_*^2}{n\mu\varepsilon}$ $+\sqrt{\frac{L(\tau-1)(\sigma_*^2+(\tau-1)\zeta_*^2)}{\mu^2\varepsilon}}$	ES, Het	G.1.2
`Local-SVRG`, Alg. 2 (NEW)	$\nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(y_i^k)$ $+\nabla f_i(y_i^k)$, $0, -$	$m + \frac{L+\max L_{ij}/n+\sqrt{(\tau-1)L\max L_{ij}}}{\mu}$ $+\frac{L\zeta^2(\tau-1)}{\mu^2\varepsilon} + \sqrt{\frac{L(\tau-1)\zeta_*^2}{\mu^2\varepsilon}}$	simple, ζ-Het	G.2
`Local-SVRG`, Alg. 2 (NEW)	$\nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(y_i^k)$ $+\nabla f_i(y_i^k)$, $0, -$	$m + \frac{L\tau+\max L_{ij}/n+\sqrt{(\tau-1)L\max L_{ij}}}{\mu}$ $+\sqrt{\frac{L(\tau-1)^2\zeta_*^2}{\mu^2\varepsilon}}$	simple, Het	G.2
`S*-Local-SGD`, Alg. 3 (NEW)	$f_{\xi_i}(x_i^k), \nabla f_i(x^*), -$	$\frac{\tau L}{\mu} + \frac{\sigma^2}{n\mu\varepsilon} + \sqrt{\frac{L(\tau-1)\sigma^2}{\mu^2\varepsilon}}$	UBV, Het	G.3
`SS-Local-SGD`, Alg. 4 (Karimireddy et al., 2019a)	$f_{\xi_i}(x_i^k), h_i^k - \frac{1}{n}\sum_{i=1}^n h_i^k$, $\nabla f_{\tilde{\xi}_i^k}(y_i^k)$	$\frac{L}{p\mu} + \frac{\sigma^2}{n\mu\varepsilon} + \sqrt{\frac{L(1-p)\sigma^2}{p\mu^2\varepsilon}}$	UBV, Het	G.4.1
`SS-Local-SGD`, Alg. 4 (NEW)	$f_{\xi_i}(x_i^k), h_i^k - \frac{1}{n}\sum_{i=1}^n h_i^k$, $\nabla f_{\tilde{\xi}_i^k}(y_i^k)$	$\frac{L}{p\mu} + \frac{\mathcal{L}}{n\mu} + \frac{\sqrt{L\mathcal{L}(1-p)}}{p\mu}$ $+\frac{\sigma_*^2}{n\mu\varepsilon} + \sqrt{\frac{L(1-p)\sigma_*^2}{p\mu^2\varepsilon}}$	ES, Het	G.4.2
`S*-Local-SGD*`, Alg. 5 (NEW)	$\nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(x^*)$ $+\nabla f_i(x^*)$, $\nabla f_i(x^*)$, $-$	$\Big(\frac{\tau L}{\mu} + \frac{\max L_{ij}}{n\mu}$ $+\frac{\sqrt{(\tau-1)L\max L_{ij}}}{\mu}\Big)\log\frac{1}{\varepsilon}$	simple, Het	G.5
`S-Local-SVRG`, Alg. 6 (NEW)	$\nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(y^k)$ $+\nabla f_i(y^k)$, $h_i^k - \frac{1}{n}\sum_{i=1}^n h_i^k, \nabla f_i(y^k)$	$\Big(m + \frac{L}{p\mu} + \frac{\max L_{ij}}{n\mu}$ $+\frac{\sqrt{L\max L_{ij}(1-p)}}{p\mu}\Big)\log\frac{1}{\varepsilon}$	simple, Het	G.6

tunately, in several practical scenarios, one can match the rate of star methods using an approach from Case III, as we shall see in the next point.

- **Shifted Local** `SVRG` **(**`S-Local-SVRG`**).** As already mentioned, local `SGD` suffers from convergence to a neighborhood of the optimum only, which is credited to i) inherent variance of the local stochastic gradient, and ii) incorrect fixed point of local `GD`. We propose a way to correct both issues. To the best of our knowledge, this is the first time that on-device variance reduction was combined with the trick for reducing the non-stationarity of local methods. Specifically, the latter is achieved by selecting b_i^k as a particular instance of Case III from As. 4.2 such that l_i^k is the full local gradient, which in turns yields $D'_{1,i} = 0, A'_i = L$. In order to not waste local computation, we synchronize the evaluation of l_i^k with the computation of the full local gradient for the `L-SVRG` (Hofmann et al., 2015; Kovalev et al., 2019) estimator, which we use to construct a_i^k. Consequently, some terms cancel out, and we obtain a simple, fast, linearly converging local `SGD` method, which we present as Algorithm 6 in the appendix. We believe that this is remarkable since only a very few local methods converge linearly to the exact optimum.[10]

[10] A linearly converging local `SGD` variant can be recovered from stochastic decoupling (Mishchenko and Richtárik,

6 Experiments

We perform multiple experiments to verify the theoretical claims of this paper. Due to space limitations, we only present a single experiment in the main body; the rest can be found in Section C of the appendix.

We demonstrate the benefit of on-device variance reduction, which we introduce in this paper. For that purpose, we compare standard `Local-SGD` (Algorithm 1) with our `Local-SVRG` (Algorithm 2) on a regularized logistic regression problem with LibSVM data (Chang and Lin, 2011). For each problem instance, we compare the two algorithms with the stepsize $\gamma \in \{1, 0.1, 0.01\}$ (we have normalized the data so that $L = 1$). The remaining details for the setup are presented in Section C.1 of the appendix.

Our theory predicts that both `Local-SGD` and `Local-SVRG` have identical convergence rate early on. However, the neighborhood of the optimum to which `Local-SVRG` converges is smaller comparing to `Local-SGD`. For both methods, the neighborhood is controlled by the stepsize: the smaller the stepsize is, the smaller the optimum neighborhood is. The price to pay is a slower rate at the beginning.

The results are presented in Fig. 1. As predicted, `Local-SVRG` always outperforms `Local-SGD` as it converges to a better neighborhood. Fig. 1 also demonstrates that one can trade the smaller neighborhood for the slower convergence by modifying the stepsize.

7 Conclusions and Future Work

This paper develops a unified approach to analyzing and designing a wide class of local stochastic first order algorithms. While our framework covers a broad range of methods, there are still some types of algorithms that we did not include but desire attention in future work. First, it would be interesting to study algorithms with *biased* local stochastic gradients; these are popular for minimizing finite sums; see `SAG` (Schmidt et al., 2017) or `SARAH` (Nguyen et al., 2017). The second hitherto unexplored direction is including Nesterov's acceleration (Nesterov, 1983) in our framework. This idea is gaining traction in the area of local methods already (Pathak and Wainwright, 2020; Yuan and Ma, 2020). However, it is not at all clear how this should be done and several attempts at achieving this unification goal failed. The third direction is allowing for a regularized local objective, which has been underexplored in the FL community so far. Other compelling directions that we do not cover are the local higher-order or proximal methods (Li et al., 2018; Pathak and Wainwright, 2020) and methods supporting partial participation (McMahan et al., 2016).

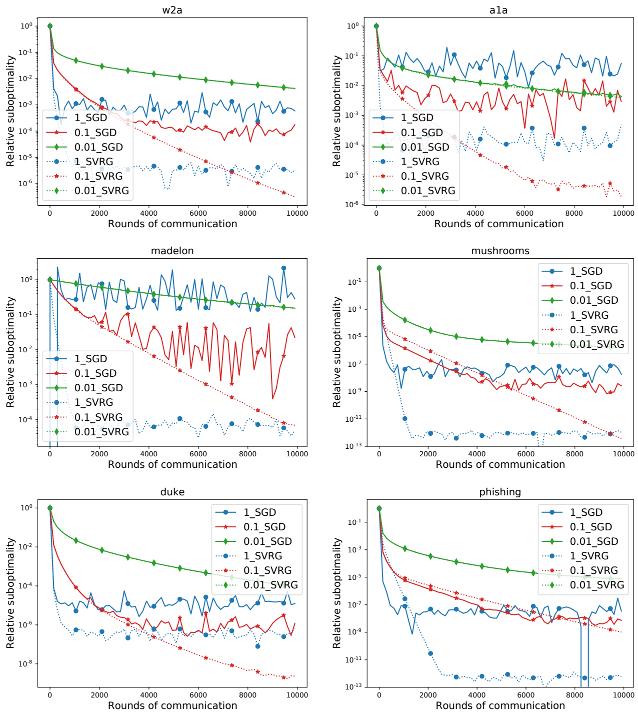

Figure 1: Comparison of standard `Local-SGD` (Alg. 1) and our `Local-SVRG` (Alg. 2) for varying γ. Logistic regression applied on LibSVM (Chang and Lin, 2011). Other parameters: $L = 1, \mu = 10^{-4}, \tau = 40$. Parameter n chosen as per Tbl. 5 in the appendix.

Acknowledgements

This work was supported by the KAUST baseline research grant of P. Richtárik. Part of this work was done while E. Gorbunov was a research intern at KAUST. The research of E. Gorbunov in Lemmas E.1, E.3 was also supported by the Ministry of Science and Higher Education of the Russian Federation (Goszadaniye) 075-00337-20-03, and in Lemmas E.2, E.4 – by RFBR, project number 19-31-51001.

References

Aji, A. F. and Heafield, K. (2017). Sparse communication for distributed gradient descent. *arXiv preprint arXiv:1704.05021*.

Alistarh, D., Hoefler, T., Johansson, M., Konstantinov, N., Khirirat, S., and Renggli, C. (2018). The convergence of sparsified gradient methods. In *Advances in Neural Information Processing Systems*, pages 5973–5983.

2019), although this was not considered therein. Besides that, FedSplit (Pathak and Wainwright, 2020) achieves a linear rate too, however, with a much stronger local oracle.

Alistarh, D., Li, J., Tomioka, R., and Vojnovic, M. (2016). QSGD: Randomized quantization for communication-optimal stochastic gradient descent. *arXiv preprint arXiv:1610.02132*.

Basu, D., Data, D., Karakus, C., and Diggavi, S. (2019). Qsparse-local-SGD: Distributed SGD with quantization, sparsification and local computations. In *Advances in Neural Information Processing Systems*, pages 14695–14706.

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. (2018). signSGD: Compressed optimisation for non-convex problems. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 560–569, Stockholmsmässan, Stockholm Sweden. PMLR.

Beznosikov, A., Horváth, S., Richtárik, P., and Safaryan, M. (2020). On biased compression for distributed learning. *arXiv preprint arXiv:2002.12410*.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654.

Gorbunov, E., Hanzely, F., and Richtárik, P. (2020a). A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 680–690. PMLR.

Gorbunov, E., Kovalev, D., Makarenko, D., and Richtárik, P. (2020b). Linearly converging error compensated sgd. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20889–20900. Curran Associates, Inc.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). SGD: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209.

Hanzely, F., Mishchenko, K., and Richtárik, P. (2018). SEGA: Variance reduction via gradient sketching. In *Advances in Neural Information Processing Systems*, pages 2082–2093.

Hanzely, F. and Richtárik, P. (2019). One method to rule them all: variance reduction for data, parameters and many new methods. *arXiv preprint arXiv:1905.11266*.

Hanzely, F. and Richtárik, P. (2020). Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*.

Hofmann, T., Lucchi, A., Lacoste-Julien, S., and McWilliams, B. (2015). Variance reduced stochastic gradient descent with neighbors. In *Advances in Neural Information Processing Systems*, pages 2305–2313.

Horváth, S., Kovalev, D., Mishchenko, K., Stich, S., and Richtárik, P. (2019). Stochastic distributed learning with gradient quantization and variance reduction. *arXiv preprint arXiv:1904.05115*.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. (2019a). Scaffold: Stochastic controlled averaging for on-device federated learning. *arXiv preprint arXiv:1910.06378*.

Karimireddy, S. P., Rebjock, Q., Stich, S. U., and Jaggi, M. (2019b). Error feedback fixes signSGD and other gradient compression schemes. *arXiv preprint arXiv:1901.09847*.

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

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. U. (2020). A unified theory of decentralized SGD with changing topology and local updates. *arXiv preprint arXiv:2003.10422*.

Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Kovalev, D., Horváth, S., and Richtárik, P. (2019). Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. *arXiv preprint arXiv:1901.08689*.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. (2018). Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*.

Liang, X., Shen, S., Liu, J., Pan, Z., Chen, E., and Cheng, Y. (2019). Variance reduced local SGD with lower communication complexity. *arXiv preprint arXiv:1912.12844*.

Lin, T., Stich, S. U., Patel, K. K., and Jaggi, M. (2018). Don't use large mini-batches, use local SGD. *arXiv preprint arXiv:1808.07217*.

Lin, Y., Han, S., Mao, H., Wang, Y., and Dally, W. J. (2017). Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR.

McMahan, H. B., Moore, E., Ramage, D., and y Arcas, B. A. (2016). Federated learning of deep networks using model averaging. *arXiv preprint arXiv:1602.05629*.

Mishchenko, K., Gorbunov, E., Takáč, M., and Richtárik, P. (2019). Distributed learning with compressed gradient differences. *arXiv preprint arXiv:1901.09269*.

Mishchenko, K., Hanzely, F., and Richtárik, P. (2020). 99% of worker-master communication in distributed optimization is not needed. In *Conference on Uncertainty in Artificial Intelligence*, pages 979–988. PMLR.

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

Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.

Nesterov, Y. E. (1983). A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547.

Nguyen, L., Nguyen, P. H., Dijk, M., Richtárik, P., Scheinberg, K., and Takáč, M. (2018). SGD and Hogwild! convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pages 3750–3758.

Nguyen, L. M., Liu, J., Scheinberg, K., and Takáč, M. (2017). Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2613–2621. JMLR. org.

Pathak, R. and Wainwright, M. J. (2020). FedSplit: An algorithmic framework for fast federated optimization. *arXiv preprint arXiv:2005.05238*.

Ramezani-Kebrya, A., Faghri, F., and Roy, D. M. (2019). NUQSGD: Improved communication efficiency for data-parallel SGD via nonuniform quantization. *arXiv preprint arXiv:1908.06077*.

Reddi, S. J., Konečný, J., Richtárik, P., Póczós, B., and Smola, A. (2016). AIDE: Fast and communication efficient distributed optimization. *arXiv preprint arXiv:1608.06879*.

Reisizadeh, A., Mokhtari, A., Hassani, H., Jadbabaie, A., and Pedarsani, R. (2020). Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International Conference on Artificial Intelligence and Statistics*, pages 2021–2031.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407.

Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112.

Shamir, O., Srebro, N., and Zhang, T. (2014). Communication-efficient distributed optimization using an approximate newton-type method. In *International Conference on Machine Learning*, pages 1000–1008.

Stich, S. U. (2018). Local SGD converges fast and communicates little. *arXiv preprint arXiv:1805.09767*.

Stich, S. U. (2019). Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*.

Stich, S. U. and Karimireddy, S. P. (2019). The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication. *arXiv preprint arXiv:1909.05350*.

Vogels, T., Karimireddy, S. P., and Jaggi, M. (2019). PowerSGD: Practical low-rank gradient compression for distributed optimization. In *Advances in Neural Information Processing Systems*, pages 14259–14268.

Wang, H., Sievert, S., Liu, S., Charles, Z., Papailiopoulos, D., and Wright, S. (2018). Atomo: Communication-efficient learning via atomic sparsification. In *Advances in Neural Information Processing Systems*, pages 9850–9861.

Wangni, J., Wang, J., Liu, J., and Zhang, T. (2018). Gradient sparsification for communication-efficient distributed optimization. In *Advances in Neural Information Processing Systems*, pages 1299–1309.

Woodworth, B., Patel, K. K., and Srebro, N. (2020a). Minibatch vs local SGD for heterogeneous distributed learning. *arXiv preprint arXiv:2006.04735*.

Woodworth, B., Patel, K. K., Stich, S. U., Dai, Z., Bullins, B., McMahan, H. B., Shamir, O., and Srebro, N. (2020b). Is local SGD better than minibatch SGD?

In *Proceedings of the 37th International Conference on Machine Learning*.

Wu, J., Huang, W., Huang, J., and Zhang, T. (2018). Error compensated quantized SGD and its applications to large-scale distributed optimization. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5325–5333, Stockholmsmässan, Stockholm Sweden. PMLR.

Wu, Z., Ling, Q., Chen, T., and Giannakis, G. B. (2019). Federated variance-reduced stochastic gradient descent with robustness to byzantine attacks. *arXiv preprint arXiv:1912.12716*.

Yuan, H. and Ma, T. (2020). Federated accelerated stochastic gradient descent. *arXiv preprint arXiv:2006.08950*.

Zhang, Y. and Lin, X. (2015). DiSCO: Distributed optimization for self-concordant empirical loss. In *International Conference on Machine Learning*, pages 362–370.

Zinkevich, M., Weimer, M., Li, L., and Smola, A. J. (2010). Parallelized stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 2595–2603.