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

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

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.,

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

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.

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.

[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.

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-

[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.

[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.

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 $x \in \mathbb{R}^d$:*

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

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 $A, A', B, B', C, C', F, F', G, H, D_1, D_1', D_2, D_3 \geq 0, \rho \in$

[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 $\text{Be}(p)$ 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*

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

for $A_i \ge 0, B_i \ge 0, D_{1,i} \ge 0, 0 \le \rho_i \le 1, C_i \ge 0, D_{2,i} \ge 0$ 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 $0 \le \rho_i' \le 1$ 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 $A_i', D_{3,i} \ge 0$ we have

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

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*

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

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

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

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

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

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

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

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 $\mathbf{E}[f(\hat{x}) - f(x^*)] \leq \varepsilon$) 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)	$\nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(x^*) + \nabla f_i(x^*), \nabla f_i(x^*), -$	$\left(\frac{\tau L}{\mu} + \frac{\max L_{ij}}{n\mu} + \frac{\sqrt{(\tau-1)L\max L_{ij}}}{\mu}\right)\log\frac{1}{\varepsilon}$	simple, Het	G.5
`S-Local-SVRG`, Alg. 6 (NEW)	$\nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(y^k) + \nabla f_i(y^k), h_i^k - \frac{1}{n}\sum_{i=1}^n h_i^k, \nabla f_i(y^k)$	$\left(m + \frac{L}{p\mu} + \frac{\max L_{ij}}{n\mu} + \frac{\sqrt{L\max L_{ij}(1-p)}}{p\mu}\right)\log\frac{1}{\varepsilon}$	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.

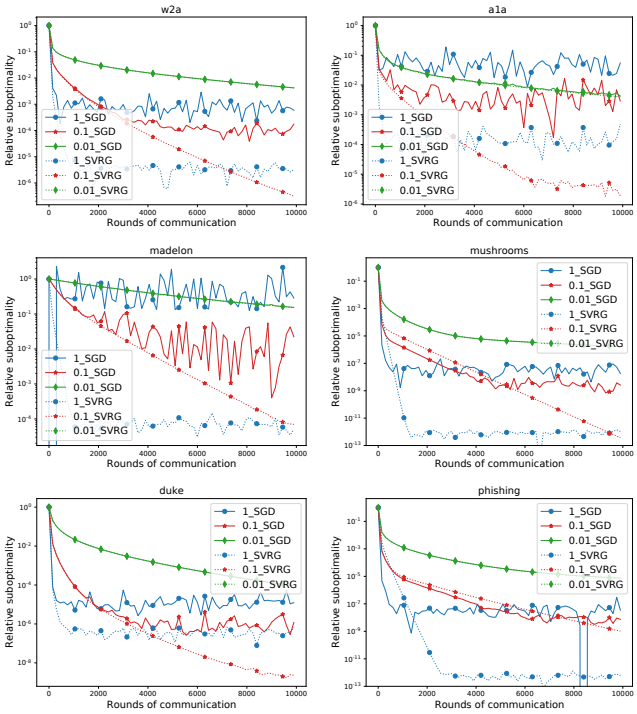

Figure 1: Comparison of standard `Local-SGD` (Alg. 1) and our `Local-SVRG` (Alg. 2) for varying γ. Logistic regression applied on LibSVM (Chang and Lin, 2011). Other parameters: $L = 1, \mu = 10^{-4}, \tau = 40$. Parameter n chosen as per Tbl. 5 in the appendix.

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).

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.

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.

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.

Appendix

Since the appendix contains substantial amount of material, we have decided to include a table of contents.

Contents

A Table of Frequently Used Notation

Table 3: Summary of frequently used notation.

Main notation		
$f:\mathbb{R}^d\to\mathbb{R}$	Objective to be minimized	(1)
$f_i:\mathbb{R}^d\to\mathbb{R}$	Local objective owned by device/worker i	(2) or (3)
x^*	Global optimum of (1); $x^*\in\mathbb{R}^d$	
d	Dimensionality of the problem space	(1)
n	Number of clients/devices/nodes/workers	(1)
x_i^k	Local iterate; $x_i^k\in\mathbb{R}^d$	(4)
g_i^k	Local stochastic direction; $g_i^k\in\mathbb{R}^d$	(4)
γ	Stepsize/learning rate; $\gamma\geq 0$	(4)
c_k	Indicator of the communication; $c_k\in\{0,1\}$	(4)
μ	Strong quasi-convexity of the local objective; $\mu\geq 0$	(5)
L	Smoothness of the local objective; $L\geq\mu$	(6)
x^k	Virtual iterate; $x^k\in\mathbb{R}^d$	Sec 2
V^k	Discrepancy between local and virtual iterates; $V^k\geq 0$	Sec 2
$\overline{x}^K$	Weighted average of historical iterates; $\overline{x}^K\in\mathbb{R}^d$	Thm 2.1
ζ	Heterogeneity parameter; $\zeta\geq 0$	(15)
τ	Size of the fixed local loop $\tau\geq 0$	Sec 3
p	Probability of aggregation fixed for the random local loop $p\in[0,1]$	Sec 3
a_i^k	Unbiased local gradient; $a_i^k\in\mathbb{R}^d$	Sec 4
b_i^k	Local shift; $b_i^k\in\mathbb{R}^d$	Sec 4
h_i^k	Delayed local gradient estimator used to construct b_i^k; $h_i^k\in\mathbb{R}^d$	Sec 4
l_i^k	Unbiased local gradient estimator used to construct b_i^k; $l_i^k\in\mathbb{R}^d$	Sec 4
$\mathcal{L}$	Expected smoothness of local objectives; $\mathcal{L}\geq 0$	(86)
$\max L_{ij}$	Smoothness constant of local summands; $\max L_{ij}\geq 0$	Sec (G.2)
σ^2	Averaged upper bound for the variance of local stochastic gradient	Tab (6)
σ_*^2	Averaged variance of local stochastic gradients at the solution	Tab (6)
ζ_*^2	$\stackrel{\text{def}}{=}\frac{1}{n}\sum_{i=1}^n\|\nabla f_i(x^*)\|^2$	Tab (6)
Parametric Assumptions		
$A, A', B, B', C, C', F, F', G, H, D_1, D_1', D_2, D_3, \rho$	Parameters of Assumption 2.3	
$A_i, B_i\, D_{1,i}, \rho_i, C_i, D_{2,i}$	Parameters of Assumption 4.1	
$A_i', D_{3,i}$	Parameters of Assumption 4.2	
$\sigma_k^2, \sigma_{i,k}^2$	Possibly random non-negative sequences from Assumptions 2.3, 4.1, E.1	
Standard		
$\mathbf{E}[\cdot]$	Expectation	
$\mathbf{E}\left[\cdot\mid x^k\right]$	$\stackrel{\text{def}}{=}\mathbf{E}\left[\cdot\mid x_1^k,\ldots,x_n^k\right]$; expectation conditioned on k-th local iterates	
$D_h(x,y)$	$\stackrel{\text{def}}{=}h(x)-h(y)-\langle\nabla h(y),x-y\rangle$; Bregman distance of x,y w.r.t. h	As 4.1

B Table with Complexity Bounds in the Weakly Convex Case

Table 4: A selection of methods that can be analyzed using our framework. 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 $\mathbf{E}[f(\hat{x}) - f(x^*)] \leq \varepsilon$) and a specific setup under which the methods are analyzed. For all algorithms we suppress constants factors. All rates are provided in the **weakly 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, $R_0 \overset{\text{def}}{=} \|x^0 - x^*\|$ – distance of the starting point x^0 from the closest solution x^*, $\Delta_0 \overset{\text{def}}{=} f(x^0) - f(x^*)$.

Method	a_i^k, b_i^k, l_i^k	Complexity	Setting	Sec
`Local-SGD`, Alg. 1 (Woodworth et al., 2020a)	$f_{\xi_i}(x_i^k), 0, -$	$\frac{LR_0^2}{\varepsilon} + \frac{\sigma^2 R_0^2}{n\varepsilon^2} + \frac{R_0^2\sqrt{L\tau(\sigma^2+\tau\zeta^2)}}{\varepsilon^{3/2}}$	UBV, ζ-Het	G.1.1
`Local-SGD`, Alg. 1 (Koloskova et al., 2020)	$f_{\xi_i}(x_i^k), 0, -$	$\frac{\tau L R_0^2}{\varepsilon} + \frac{\sigma^2 R_0^2}{n\varepsilon^2} + \frac{R_0^2\sqrt{L(\tau-1)(\sigma^2+(\tau-1)\zeta_*^2)}}{\varepsilon^{3/2}}$	UBV, Het	G.1.1
`Local-SGD`, Alg. 1 (Khaled et al., 2020)♣	$f_{\xi_i}(x_i^k), 0, -$	$\frac{\left(L+\mathcal{L}/n+\sqrt{(\tau-1)L\mathcal{L}}\right)R_0^2}{\varepsilon} + \frac{\sigma_*^2 R_0^2}{n\varepsilon^2} + \frac{L\zeta^2(\tau-1)R_0^2}{\mu\varepsilon^2} + \frac{R_0^2\sqrt{L(\tau-1)(\sigma_*^2+\zeta_*^2)}}{\varepsilon^{3/2}}$	ES, ζ-Het	G.1.2
`Local-SGD`, Alg. 1 (Khaled et al., 2020)♣	$f_{\xi_i}(x_i^k), 0, -$	$\frac{\left(L\tau+\mathcal{L}/n+\sqrt{(\tau-1)L\mathcal{L}}\right)R_0^2}{\varepsilon} + \frac{\sigma_*^2 R_0^2}{n\varepsilon^2} + \frac{R_0^2\sqrt{L(\tau-1)(\sigma_*^2+(\tau-1)\zeta_*^2)}}{\varepsilon^{3/2}}$	ES, Het	G.1.2
`Local-SVRG`, Alg. 2 (NEW)	$\nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(y_i^k) + \nabla f_i(y_i^k), 0, -$	$\frac{\left(L+\max L_{ij}\sqrt{m/n}+\sqrt{(\tau-1)L\max L_{ij}}\right)R_0^2}{\varepsilon} + \frac{\sqrt[3]{(\tau-1)mL\max L_{ij}}R_0^2}{\varepsilon} + \frac{L\zeta^2(\tau-1)R_0^2}{\mu\varepsilon^2} + \frac{R_0^2\sqrt{L(\tau-1)\zeta_*^2}}{\varepsilon^{3/2}}$	simple, ζ-Het	G.2
`Local-SVRG`, Alg. 2 (NEW)	$\nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(y_i^k) + \nabla f_i(y_i^k), 0, -$	$\frac{\left(L\tau+\max L_{ij}\sqrt{m/n}+\sqrt{(\tau-1)L\max L_{ij}}\right)R_0^2}{\varepsilon} + \frac{\sqrt[3]{(\tau-1)mL\max L_{ij}}R_0^2}{\varepsilon} + \frac{R_0^2\sqrt{L(\tau-1)^2\zeta_*^2}}{\varepsilon^{3/2}}$	simple, Het	G.2
`S*-Local-SGD`, Alg. 3 (NEW)	$f_{\xi_i}(x_i^k), \nabla f_i(x^*), -$	$\frac{\tau L R_0^2}{\varepsilon} + \frac{\sigma^2 R_0^2}{n\varepsilon^2} + \frac{R_0^2\sqrt{L(\tau-1)\sigma^2}}{\varepsilon^{3/2}}$	UBV, Het	G.3
`SS-Local-SGD`, Alg. 4 (Karimireddy et al., 2019a)	$f_{\xi_i}(x_i^k), h_i^k - \frac{1}{n}\sum_{i=1}^n h_i^k, \nabla f_{\tilde{\xi}_i^k}(y_i^k)$	$\frac{LR_0^2}{p\varepsilon} + \frac{\sigma^2 R_0^2}{n\varepsilon^2} + \frac{R_0^2\sqrt{L(1-p)\sigma^2}}{p^{1/2}\varepsilon^{3/2}}$	UBV, Het	G.4.1
`SS-Local-SGD`, Alg. 4 (NEW)	$f_{\xi_i}(x_i^k), h_i^k - \frac{1}{n}\sum_{i=1}^n h_i^k, \nabla f_{\tilde{\xi}_i^k}(y_i^k)$	$\frac{\left(L+p\mathcal{L}/n+\sqrt{p(1-p)L\mathcal{L}}\right)R_0^2}{p\varepsilon} + \frac{\sqrt[3]{(1-p)L(L+p\mathcal{L})R_0^4\Delta_0}}{p\varepsilon} + \frac{\sqrt[3]{(1-p)L\sigma_*^2R_0^4}}{p^{2/3}\varepsilon} + \frac{\sigma_*^2R_0^2}{n\varepsilon^2} + \frac{R_0^2\sqrt{L(1-p)\sigma_*^2}}{p^{1/2}\varepsilon^{3/2}}$	ES, Het	G.4.2
`S*-Local-SGD*`, Alg. 5 (NEW)	$\nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(x^*) + \nabla f_i(x^*), \nabla f_i(x^*), -$	$\frac{\left(L\tau+\max L_{ij}/n+\sqrt{(\tau-1)L\max L_{ij}}\right)R_0^2}{\varepsilon}$	simple, Het	G.5
`S-Local-SVRG`, Alg. 6 (NEW)	$\nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(y^k) + \nabla f_i(y^k), h_i^k - \frac{1}{n}\sum_{i=1}^n h_i^k, \nabla f_i(y^k)$	$\frac{\left(L+pL\sqrt{m/n}+\sqrt{(1-p)L\max L_{ij}}\right)R_0^2}{p\varepsilon} + \frac{R_0^2\sqrt[3]{L\max L_{ij}^2}}{p^{2/3}\varepsilon}$	simple, Het	G.6

C Extra Experiments

C.1 Missing details from Section 6 and an extra figure

In Section 6 we study the effect of local variance reduction on the communication complexity of local methods. We consider the regularized logistic regression objective, i.e., we choose

$$f_i(x) \overset{\text{def}}{=} \frac{1}{m}\sum_{j=1}^{m} \log\left(1+\exp\left(\langle a_{(i-1)m+j}, x\rangle \cdot b_{(i-1)m+j}\right)\right) + \frac{\mu}{2}\|x\|^2,$$

where $a_j \in \mathbb{R}^d, b_j \in \{-1, 1\}$ for $j \leq nm$ are the training data and labels.

Number of the clients. We select a different number of clients for each dataset in order to capture a variety of scenarios. See Table 5 for details.

Table 5: Number of clients per dataset (Figures 1 and 2).

Dataset	n	**# datapoints** $(= mn)$	d
`a1a`	5	1 605	123
`mushrooms`	12	8 124	112
`phishing`	11	11 055	68
`madelon`	50	2 000	500
`duke`	4	44	7 129
`w2a`	10	3 470	300

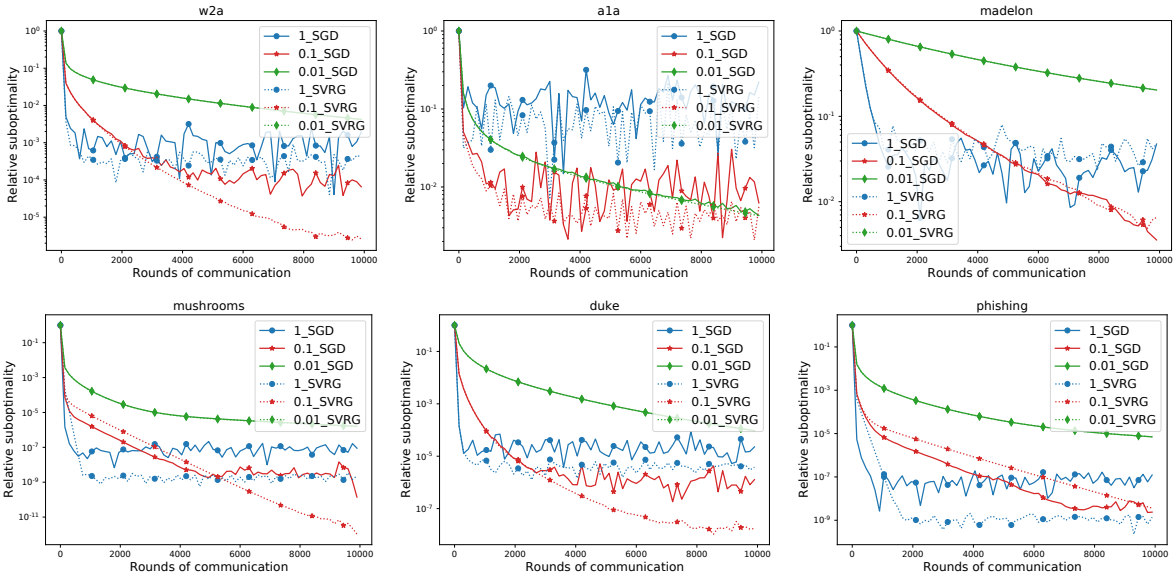

Figure 2: Comparison of standard `Local-SGD` (Algorithm 1), and `Local-SVRG` (Algorithm 2) with various stepsizes γ. Logistic regression applied on LibSVM data (Chang and Lin, 2011) with heterogenously splitted data. Other parameters: $L = 1, \mu = 10^{-4}, \tau = 40$. Parameter n chosen as per Table 5. (Same as Fig. 1, but with the heterogenous data split)

Data split. The experiment from Figure 1 in the main body of the paper splits the data among the clients uniformly at random (i.e., split according to the the order given by a random permutation). However, in a typical FL scenario, the local data might significantly differ from the population average. For this reason, we also test on a different split of the data: we first sort the data according to the labels, and then split them among the clients. Figure 2 shows the results. We draw a conclusions identical to Figure 1. We see that `Local-SVRG` was at least as good as `Local-SGD` for every stepsize choice and every dataset. Further, the prediction that the smaller stepsize yields the smaller of the optimum neighborhood for the price of slower convergence was confirmed.

Table 6: Instances of (17).

Type	m	z_i^*
0	1	$\sim \mathcal{N}(0, \mathbf{I})$
1	10	$\sim \mathcal{N}(0, \mathbf{I})$
2	1	$\sim \mathcal{N}(0, \mathbf{I})$
3	10	$\sim \mathcal{N}(0, \mathbf{I})$

Environment. All experiments were performed in a simulated environment on a single machine.

C.2 The effect of local shift/drifts

The experiment presented in Section 6 examined the effect of the noise on the performance of local methods and demonstrated that control variates can be efficiently employed to reduce that noise. In this section, we study the second factor that influences the neighborhood to which `Local-SGD` converges: non-stationarity of `Local-GD`.

We have already shown that the mentioned non-stationarity of `Local-GD` can be fixed using a carefully designed idealized/optimal shift that depends on the solution x^* (see Algorithm 3). Furthermore, we have shown that this idealized shift can be learned on-the-fly at the small price of slightly slower convergence rate (see Algorithm 4 – `SS-Local-SGD`/`SCAFFOLD`).[11]

In this experiment, we therefore compare `Local-SGD`, `S*-Local-SGD` and `SCAFFOLD`. In order to decouple the local variance with the non-stationarity of the local methods, we let each algorithm access the full local gradients. Next, in order to have a full control of the setting, we let the local objectives to be artificially generated quadratic problems. Specifically, we set

$$f_i(x) = \frac{\mu}{2}\|x\|^2 + \frac{1-\mu}{2}(x - z_i^*)^\top \left(\sum_{j=1}^m a_i a_i^\top\right)(x - z_i^*), \tag{17}$$

where a_i are mutually orthogonal vectors of norm 1 with $m < d$ (generated by orthogonalizing Gaussian vectors), z_i^* are Gaussian vectors and $\mu = 10^{-3}$. We consider four different instances of (17) given by Table 17. Figures 3, 4, 5, 6 show the result.

Through most of the plots across all combinations of type, τ, n, we can see that `Local-SGD` suffers greatly from the fact that it is attracted to an incorrect fixed point and as a result, it never converges to the exact optimum. On the other hand, both `S*-Local-SGD` and `SCAFFOLD` converge to the exact optimum and therefore outperform `Local-SGD` in most examples. We shall note that the rate of `SCAFFOLD` involves slightly worse constants than those in `Local-SGD` and `S*-Local-SGD`, and therefore it sometimes performs worse in the early stages of the optimization process when compared to the other methods. Furthermore, notice that our method `S*-Local-SGD` always performed best.

To summarize, our results demonstrate that

(i) the incorrect fixed point of used by standard local methods is an issue not only theory but also in practice, and should be addressed if better performance is required,

(ii) the theoretically optimal shift employed by `S*-Local-SGD` is ideal from a performance perspective if it was available (however, this strategy is impractical to implement as the optimal shift presumes the knowledge of the optimal solution), and

(iii) `SCAFFOLD`/`SS-Local-SGD` is a practical solution to fixing the incorrect fixed point problem – it converges to the exact optimum at a price of a slightly worse initial convergence speed.

[11] In fact, `SCAFFOLD` can be coupled together with `Local-SVRG` given that the local objectives are of a finite-sum structure, resulting in Algorithm 6.

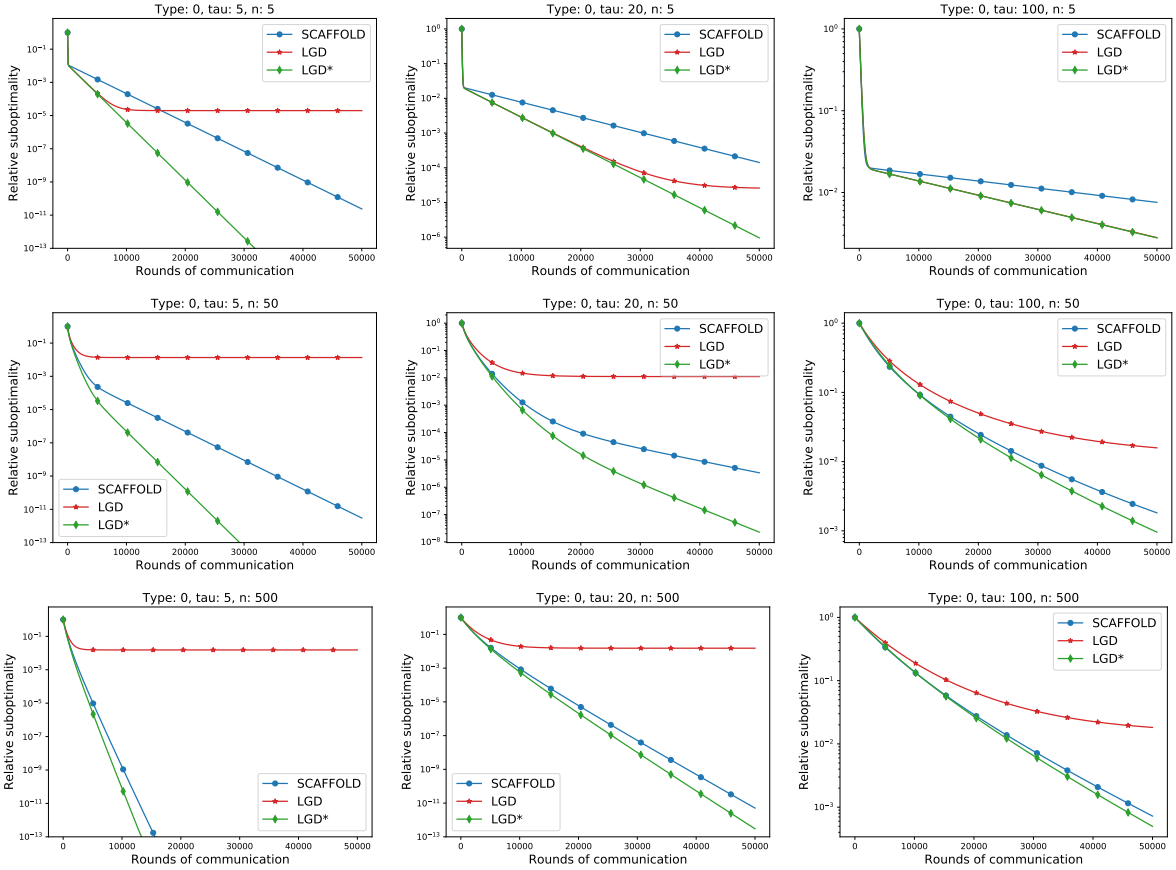

Figure 3: Comparison of the following noiseless algorithms `Local-SGD` (`LGD`, Algorithm 1 with no local noise) and `SCAFFOLD` (Karimireddy et al., 2019a) (Algorithm 4 without "Loopless") and `S*-Local-SGD` (`LGD*`, Algorithm 3). Quadratic minimization, problem type 0 (see Table 6).

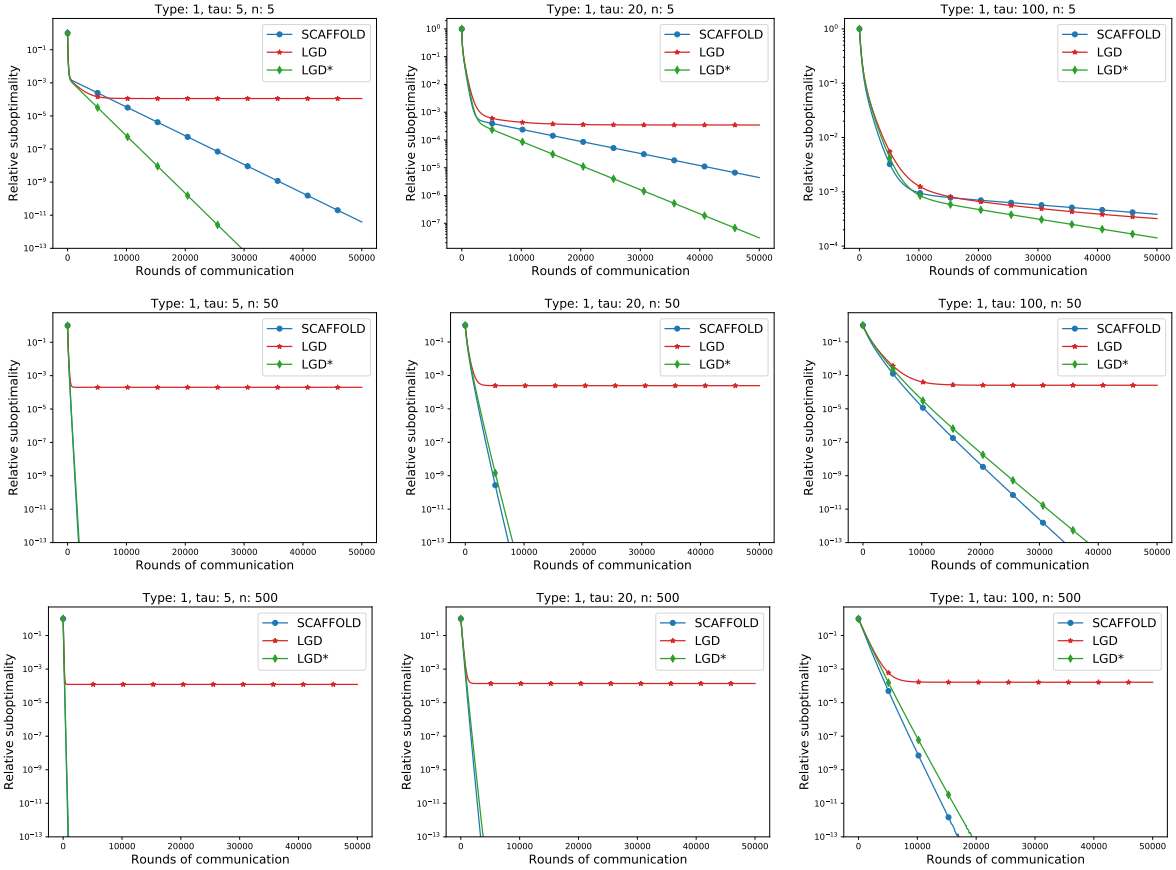

Figure 4: Comparison of the following noiseless algorithms `Local-SGD` (`LGD`, Algorithm 1 with no local noise) and `SCAFFOLD` (Karimireddy et al., 2019a) (Algorithm 4 without "Loopless") and `S*-Local-SGD` (`LGD*`, Algorithm 3). Quadratic minimization, problem type 1 (see Table 6).

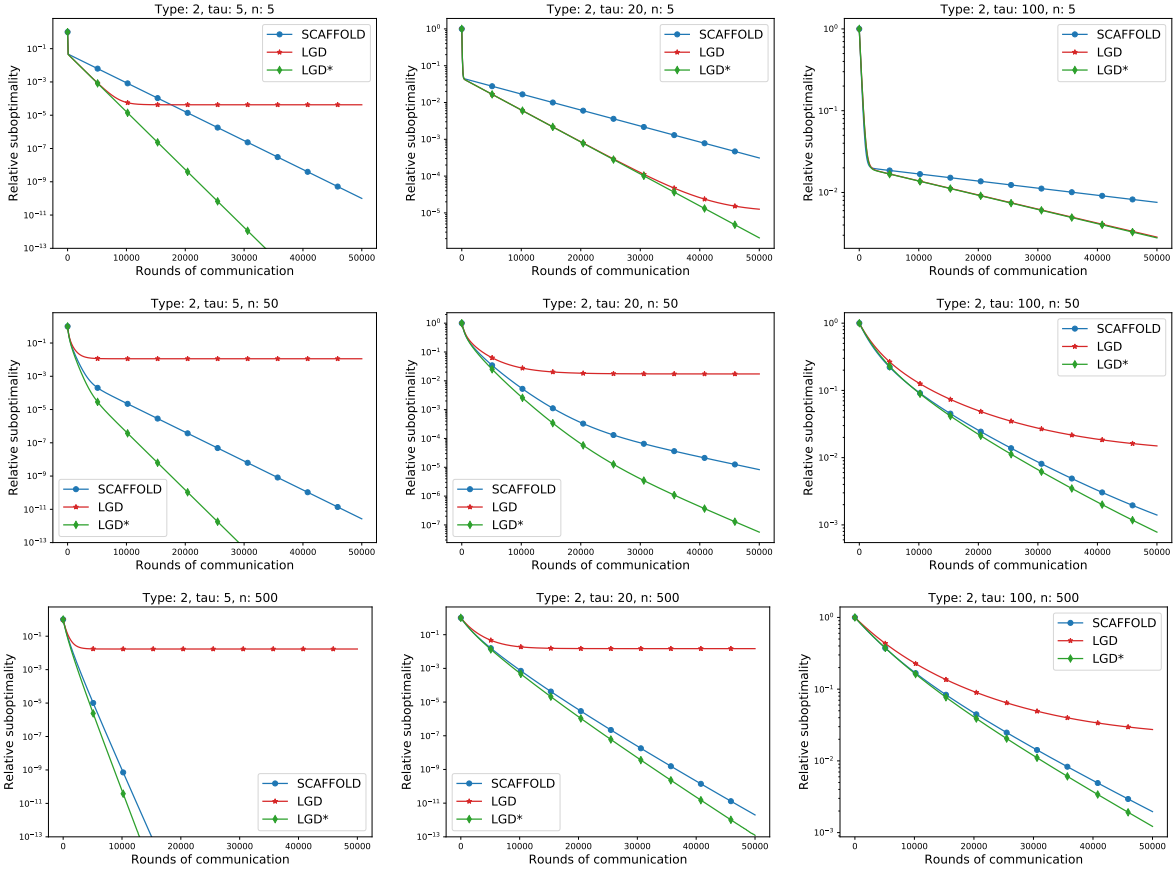

Figure 5: Comparison of the following noiseless algorithms `Local-SGD` (`LGD`, Algorithm 1 with no local noise) and `SCAFFOLD` (Karimireddy et al., 2019a) (Algorithm 4 without "Loopless") and `S*-Local-SGD` (`LGD*`, Algorithm 3). Quadratic minimization, problem type 2 (see Table 6).

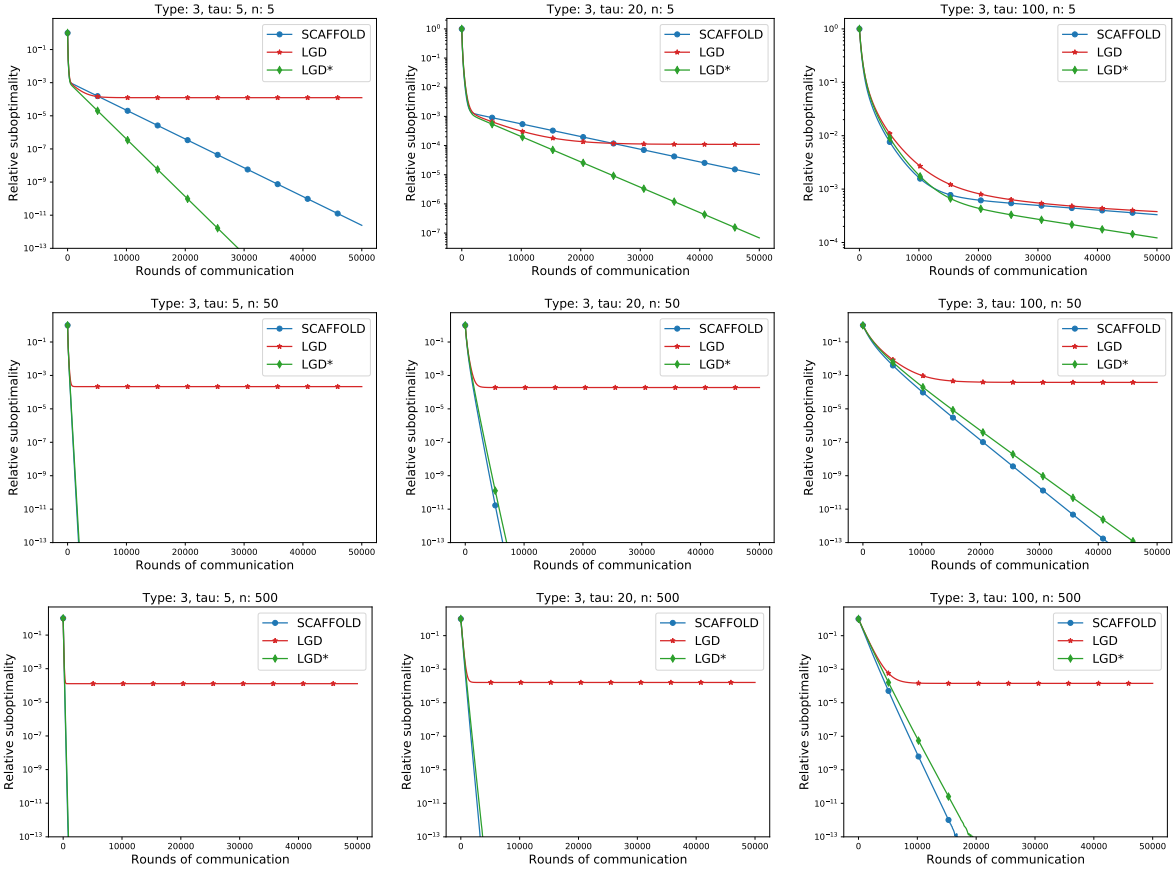

Figure 6: Comparison of the following noiseless algorithms: `Local-SGD` (`LGD`, Algorithm 1 with no local noise) and `SCAFFOLD` (Karimireddy et al., 2019a) (Algorithm 4 without "Loopless") and `S*-Local-SGD` (`LGD*`, Algorithm 3). Quadratic minimization, problem type 3 (see Table 6).

D Missing Proofs for Section 2

Let us first state some well-known consequences of L-smoothness. Specifically, if f_i is L-smooth, we must have

$$f_i(y) \leq f_i(x) + \langle \nabla f_i(x), y - x\rangle + \frac{L}{2}\|x-y\|^2, \qquad \forall x, y \in \mathbb{R}^d. \tag{18}$$

If in addition to this we assume that f_i is convex, the following bound holds:

$$\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq 2L(f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y\rangle) \overset{\text{def}}{=} 2LD_{f_i}(x,y), \qquad \forall x, y \in \mathbb{R}^d \tag{19}$$

We next proceed with the proof of Theorem 2.1. Following the technique of virtual iterates from (Stich and Karimireddy, 2019; Khaled et al., 2020), notice that the sequence $\{x^k\}_{k\geq 0}$ satisfies the recursion

$$x^{k+1} = x^k - \frac{\gamma}{n}\sum_{i=1}^n g_i^k. \tag{20}$$

This observation forms the backbone of the key lemma of our paper, which we present next.

Lemma D.1. *Let As. 2.1, 2.2 and 2.3 be satisfied and $\gamma \leq \min\left\{1/2(A'+MC), L/(F'+MG)\right\}$, where $M = \frac{4B'}{3\rho}$. Let $\eta \overset{def}{=} \min\left\{\gamma\mu, \frac{\rho}{4}\right\}$. Then for all $k \geq 0$ we have*

$$\gamma \mathbf{E}\left[f(x^k) - f(x^*)\right] \leq (1-\eta)\mathbf{E}T^k - \mathbf{E}T^{k+1} + \gamma^2(D_1' + MD_2) + 2L\gamma\mathbf{E}V_k, \tag{21}$$

where $\eta \overset{def}{=} \min\left\{\gamma\mu, \frac{\rho}{4}\right\}$, $T^k \overset{def}{=} \|x^k - x^*\|^2 + M\gamma^2\sigma_k^2$.

Proof. First of all, to simplify the proofs we introduce new notation: $g^k \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n g_i^k$. Using this and (20) we get

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &\overset{(20)}{=} \left\|x^k - x^* - \gamma g^k\right\|^2 \\ &= \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, g^k\rangle + \gamma^2\|g^k\|^2.\end{aligned}$$

Taking conditional mathematical expectation $\mathbf{E}_k[\cdot] = \mathbf{E}[\cdot \mid x^k] \overset{\text{def}}{=} \mathbf{E}[\cdot \mid x_1^k, \ldots, x_n^k]$ on both sides of the previous inequality we get

$$\mathbf{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] \overset{(7)}{=} \|x^k - x^*\|^2 - \frac{2\gamma}{n}\sum_{i=1}^n \left\langle x^k - x^*, \nabla f_i(x_i^k)\right\rangle + \gamma^2\mathbf{E}\left[\|g^k\|^2 \mid x^k\right],$$

hence

$$\begin{aligned}\mathbf{E}\left[\|x^{k+1} - x^*\|^2\right] &\overset{(140)}{\leq} \mathbf{E}\left[\|x^k - x^*\|^2\right] - \frac{2\gamma}{n}\sum_{i=1}^n \mathbf{E}\left[\left\langle x^k - x^*, \nabla f_i(x_i^k)\right\rangle\right] + \gamma^2\mathbf{E}\left[\|g^k\|^2\right] \\ &\overset{(8)}{\leq} \mathbf{E}\left[\|x^k - x^*\|^2\right] - \frac{2\gamma}{n}\sum_{i=1}^n \mathbf{E}\left[\left\langle x^k - x^*, \nabla f_i(x_i^k)\right\rangle\right] + B'\gamma^2\mathbf{E}\left[\sigma_k^2\right] \\ &\quad + 2A'\gamma^2\mathbf{E}\left[f(x^k) - f(x^*)\right] + F'\gamma^2\mathbf{E}\left[V_k\right] + \gamma^2 D_1'.\end{aligned} \tag{22}$$

Next, we derive an upper bound for the second term on the right-hand side of the previous inequality:

$$\begin{aligned}-\frac{2\gamma}{n}\sum_{i=1}^n \left\langle x^k - x^*, \nabla f_i(x_i^k)\right\rangle &= \frac{2\gamma}{n}\sum_{i=1}^n \left(\left\langle x^* - x_i^k, \nabla f_i(x_i^k)\right\rangle + \left\langle x_i^k - x^k, \nabla f_i(x_i^k)\right\rangle\right) \\ &\overset{(5),(18)}{\leq} \frac{2\gamma}{n}\sum_{i=1}^n \left(f_i(x^*) - f_i(x_i^k) - \frac{\mu}{2}\|x_i^k - x^*\|^2\right) \\ &\quad + \frac{2\gamma}{n}\sum_{i=1}^n \left(f_i(x_i^k) - f_i(x^k) + \frac{L}{2}\|x^k - x_i^k\|^2\right) \\ &\overset{(136)}{\leq} -2\gamma\left(f(x^k) - f(x^*)\right) - \mu\gamma\|x^k - x^*\|^2 + L\gamma V_k.\end{aligned} \tag{23}$$

Plugging (23) in (22), we obtain

$$\begin{aligned}\mathbf{E}\left[\|x^{k+1}-x^*\|^2\right] &\overset{(22),(23)}{\leq} (1-\gamma\mu)\mathbf{E}\left[\|x^k-x^*\|^2\right] - 2\gamma\left(1-A'\gamma\right)\mathbf{E}\left[f(x^k)-f(x^*)\right] \\ &\quad +B'\gamma^2\mathbf{E}\left[\sigma_k^2\right] + \gamma\left(L+F'\gamma\right)\mathbf{E}\left[V_k\right] + \gamma^2 D_1'. \end{aligned} \tag{24}$$

It implies that

$$\begin{aligned}\mathbf{E}T^{k+1} &= \mathbf{E}\left[\|x^{k+1}-x^*\|^2\right] + M\gamma^2\mathbf{E}\left[\sigma_{k+1}^2\right] \\ &\overset{(24),(10)}{\leq} (1-\gamma\mu)\mathbf{E}\|x^k-x^*\|^2 + \left(1+\frac{B'}{M}-\rho\right)M\gamma^2\mathbf{E}\sigma_k^2 \\ &\quad -2\gamma\left(1-\left(A'+MC\right)\gamma\right)\mathbf{E}\left[f(x^k)-f(x^*)\right] \\ &\quad +\gamma\left(L+\left(F'+MG\right)\gamma\right)\mathbf{E}V_k + \gamma^2\left(D_1'+MD_2\right).\end{aligned}$$

Since $M=\frac{4B'}{3\rho}$, $\eta=\min\left\{\gamma\mu,\frac{\rho}{4}\right\}$ and $\gamma\leq\min\left\{1/2(A'+MC), L/(F'+MG)\right\}$, we get

$$\begin{aligned}\mathbf{E}T^{k+1} &\leq (1-\gamma\mu)\mathbf{E}\|x^k-x^*\|^2 + \left(1-\frac{\rho}{4}\right)M\gamma^2\mathbf{E}\sigma_k^2 - \gamma\mathbf{E}\left[f(x^k)-f(x^*)\right] \\ &\quad +2L\gamma\mathbf{E}V_k + \gamma^2\left(D_1'+MD_2\right) \\ &\leq (1-\eta)\mathbf{E}T^k - \gamma\mathbf{E}\left[f(x^k)-f(x^*)\right] + 2L\gamma\mathbf{E}V_k + \gamma^2\left(D_1'+MD_2\right).\end{aligned}$$

Rearranging the terms we get (21). □

Using the above lemma we derive the main complexity result.

D.1 Proof of Theorem 2.1

From Lemma D.1 we have that

$$\gamma\mathbf{E}\left[f(x^k)-f(x^*)\right] \leq (1-\eta)\mathbf{E}T^k - \mathbf{E}T^{k+1} + \gamma^2(D_1'+MD_2) + 2L\gamma\mathbf{E}V_k.$$

Summing up previous inequalities for $k=0,\ldots,K$ with weights w_k defined in (12) we derive

$$\begin{aligned}\gamma\sum_{k=0}^K w_k\mathbf{E}\left[f(x^k)-f(x^*)\right] &\leq \sum_{k=0}^K\left(w_k(1-\eta)\mathbf{E}T^k - w_k\mathbf{E}T^{k+1}\right) + \gamma^2(D_1'+MD_2)W_K \\ &\quad +2L\gamma\sum_{k=0}^K w_k\mathbf{E}V_k \\ &\overset{(12),(11)}{\leq} \sum_{k=0}^K\left(w_{k-1}\mathbf{E}T^k - w_k\mathbf{E}T^{k+1}\right) + \gamma^2\left(D_1'+MD_2\right)W_K \\ &\quad +\frac{\gamma}{2}\sum_{k=0}^K w_k\mathbf{E}\left[f(x^k)-f(x^*)\right] + 2LH\gamma\mathbf{E}\sigma_0^2 + 2L\gamma^3 D_3 W_K.\end{aligned}$$

Relations $T^k\geq 0$ and $w_{-1}=1$ imply that

$$\frac{\gamma}{2}\sum_{k=0}^K w_k\mathbf{E}\left[f(x^k)-f(x^*)\right] \leq T^0 + 2LH\gamma\mathbf{E}\sigma_0^2 + \gamma^2\left(D_1'+MD_2+2L\gamma D_3\right)W_K.$$

Using the definition of $\overline{x}^K$ and convexity of f, we get

$$\mathbf{E}\left[f(\overline{x}^K)-f(x^*)\right] \leq \frac{2T^0+4LH\gamma\mathbf{E}\sigma_0^2}{\gamma W_K} + 2\gamma\left(D_1'+MD_2+2L\gamma D_3\right). \tag{25}$$

It remains to consider two cases: $\mu>0$ and $\mu=0$. If $\mu>0$ we have $W_K\geq w_K\geq(1-\eta)^{-K}$, where $\eta\overset{\text{def}}{=}\min\left\{\gamma\mu,\frac{\rho}{4}\right\}$ which implies (13). Finally, when $\mu=0$, we have $w_k=1$ for all $k\geq 0$, which implies $W_K=K+1\geq K$ and (14).

D.2 Corollaries

We state the full complexity results that can be obtained from Theorem 2.1. These results can be obtained as a direct consequence of Lemmas I.2 and I.3.

Corollary D.1. *Consider the setup from Theorem 2.1 and denote $\frac{1}{h}$ to be the resulting upper bound on γ[12] and $\mu > 0$.*

1. *If D_3 does not depend on γ, then for all K such that*

$$\begin{aligned} \text{either} \quad & \frac{\ln\left(\max\{2, \min\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\}\}\right)}{K} \leq \rho \\ \text{or} \quad & \frac{1}{h} \leq \frac{\ln\left(\max\{2, \min\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\}\}\right)}{\mu K}, \end{aligned}$$

$a = 2\|x^0 - x^*\|^2 + \frac{8B'\mathbf{E}\sigma_0^2}{3h^2\rho} + \frac{4LH\mathbf{E}\sigma_0^2}{h}$, $c_1 = 2D_1' + \frac{4B'D_2}{3\rho}$, $c_2 = 4LD_3$ and

$$\begin{aligned} \gamma &= \min\left\{\frac{1}{h}, \gamma_K\right\}, \\ \gamma_K &= \frac{\ln\left(\max\left\{2, \min\left\{\frac{a\mu^2K^2}{c_1}, \frac{a\mu^3K^3}{c_2}\right\}\right\}\right)}{\mu K}, \end{aligned}$$

we have[13]

$$\mathbf{E}\left[f(\overline{x}^K)\right] - f(x^*) = \widetilde{\mathcal{O}}\left(ha\exp\left(-\min\left\{\frac{\mu}{h}, \rho\right\}K\right) + \frac{c_1}{\mu K} + \frac{c_2}{\mu^2K^2}\right).$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K)\right] - f(x^*) \leq \varepsilon$, the method requires[14]*:*

$$K = \widetilde{\mathcal{O}}\left(\left(\frac{1}{\rho} + \frac{h}{\mu}\right)\log\left(\frac{ha}{\varepsilon}\right) + \frac{c_1}{\mu\varepsilon} + \sqrt{\frac{c_2}{\mu^2\varepsilon}}\right).$$

2. *If $D_3 = D_{3,1} + \frac{D_{3,2}}{\gamma}$, then the same bounds hold with $c_1 = 2D_1' + \frac{4B'D_2}{3\rho} + 2LD_{3,2}$ and $c_2 = 4LD_{3,1}$.*

Corollary D.2. *Let assumptions of Theorem 2.1 be satisfied with any $\gamma \leq \frac{1}{h}$ and $\mu = 0$.*

1. *If D_3 does not depend on γ, then for all K and*

$$\gamma = \min\left\{\frac{1}{h}, \sqrt{\frac{a}{b_1}}, \sqrt[3]{\frac{a}{b_2}}, \sqrt{\frac{a}{c_1K}}, \sqrt[3]{\frac{a}{c_2K}}\right\},$$

where $a = 2\|x^0 - x^*\|^2$, $b_1 = 4LH\mathbf{E}\sigma_0^2$, $b_2 = \frac{8B'\mathbf{E}\sigma_0^2}{3\rho}$, $c_1 = 2D_1' + \frac{4B'D_2}{3\rho}$, $c_2 = 4LD_3$, we have

$$\mathbf{E}\left[f(\overline{x}^K)\right] - f(x^*) = \mathcal{O}\left(\frac{ha}{K} + \frac{\sqrt{ab_1}}{K} + \frac{\sqrt[3]{a^2b_2}}{K} + \sqrt{\frac{ac_1}{K}} + \frac{\sqrt[3]{a^2c_2}}{K^{2/3}}\right).$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K)\right] - f(x^*) \leq \varepsilon$, the method requires

$$K = \mathcal{O}\left(\frac{ha}{\varepsilon} + \frac{\sqrt{ab_1}}{\varepsilon} + \frac{\sqrt[3]{a^2b_2}}{\varepsilon} + \frac{ac_1}{\varepsilon^2} + \frac{a\sqrt{c_2}}{\varepsilon^{3/2}}\right).$$

2. *If $D_3 = D_{3,1} + \frac{D_{3,2}}{\gamma}$, then the same bounds hold with $c_1 = 2D_1' + \frac{4B'D_2}{3\rho} + 2LD_{3,2}$ and $c_2 = 4LD_{3,1}$.*

[12] In order to obtain tight estimate of parameters D_3 and H, we shall impose further bounds on γ (see Section 3 and Table 1 therein).

[13] $\widetilde{\mathcal{O}}$ hides numerical constants and logarithmical factors depending on K and parameters of the problem.

[14] If $c_1 = c_2 = 0$, then one can replace $\widetilde{\mathcal{O}}$ by $\mathcal{O}$.

E Missing Proofs and Details for Section 3

E.1 Constant Local Loop

In this section we show how our results can be applied to analyze (4) in the case when

$$c_k = \begin{cases} 1, & \text{if } k \mod \tau = 0, \\ 0, & \text{if } k \mod \tau \neq 0, \end{cases}$$

where τ is number of local steps between two neighboring rounds of communications. This corresponds to the setting in which the local loop size on each device has a fixed length.

E.1.1 Heterogenous Data

First of all, we need to assume more about g_i^k.

Assumption E.1. *We assume that inequalities (8)-(10) hold and additionally there exist such non-negative constants $\widetilde{A}, \widehat{A}, \widetilde{B}, \widehat{B}, \widetilde{F}, \widehat{F}, \widetilde{D}_1, \widehat{D}_1$ that for all $k \geq 0$*

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\bar{g}_i^k\|^2\right] \leq 2\widetilde{A}\mathbf{E}\left[f(x^k) - f(x^*)\right] + \widetilde{B}\mathbf{E}\left[\sigma_k^2\right] + \widetilde{F}\mathbf{E}\left[V_k\right] + \widetilde{D}_1, \tag{26}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2\right] \leq 2\widehat{A}\mathbf{E}\left[f(x^k) - f(x^*)\right] + \widehat{B}\mathbf{E}\left[\sigma_k^2\right] + \widehat{F}\mathbf{E}\left[V_k\right] + \widehat{D}_1, \tag{27}$$

where $\bar{g}_i^k = \mathbf{E}\left[g_i^k \mid x_1^k, \ldots, x_n^k\right]$.

We notice that inequalities (26)-(27) imply (8) and vice versa. Indeed, if (26)-(27) hold then inequality (8) holds with $A = \widetilde{A} + \widehat{A}$, $B = \widetilde{B} + \widehat{B}$, $F = \widetilde{F} + \widehat{F}$, $D_1 = \widetilde{D}_1 + \widehat{D}_1$ due to variance decomposition formula (139), and if (8) is true then (26)-(27) also hold with $\widetilde{A} = \widehat{A} = A$, $\widetilde{B} = \widehat{B} = B$, $\widetilde{F} = \widehat{F} = F$, $\widetilde{D}_1 = \widehat{D}_1 = D_1$.

We start our analysis without making any assumption on homogeneity of data that workers have an access to. Next lemma provides an upper bound for the weighted sum of $\mathbf{E}V_k$.

Lemma E.1. *Let As. 2.1, 2.2 and E.1 hold and*[15]

$$\gamma \leq \min\left\{\frac{1}{4(\tau-1)\mu}, \frac{1}{2\sqrt{e(\tau-1)\left(\widetilde{F}(\tau-1) + \widehat{F} + \frac{2G(\widetilde{B}(\tau-1)+\widehat{B})}{\rho(1-\rho)}\right)}}\right\},$$

$$\gamma \leq \frac{1}{4\sqrt{2eL(\tau-1)\left(\widetilde{A}(\tau-1) + \widehat{A} + \frac{2C(\widetilde{B}(\tau-1)+\widehat{B})}{\rho(1-\rho)}\right)}}$$

Then (11) holds with

$$H = \frac{4e(\tau-1)(\widetilde{B}(\tau-1)+\widehat{B})(2+\rho)\gamma^2}{\rho}, \quad D_3 = 2e(\tau-1)\left(\widetilde{D}_1(\tau-1) + \widehat{D}_1 + \frac{2D_2(\widetilde{B}(\tau-1)+\widehat{B})}{\rho}\right). \tag{28}$$

Proof. Consider some integer $k \geq 0$. There exists such integer $t \geq 0$ that $\tau t \leq k \leq \tau(t+1) - 1$. Using this and

[15] When $\rho = 1$ one can always set the parameters in such a way that $\widetilde{B} = \widehat{B} = C = G = 0$, $D_2 = 0$. In this case we assume that $\frac{2\widetilde{B}C}{\rho(1-\rho)} = \frac{2\widehat{B}C}{\rho(1-\rho)} = \frac{2\widetilde{B}G}{\rho(1-\rho)} = \frac{2\widehat{B}G}{\rho(1-\rho)} = 0$.

Lemma I.1 we get

$$\begin{aligned}
\mathbf{E}[V_k] &\overset{(4),(20)}{=} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|x_i^{\tau t} - \gamma\sum_{l=\tau t}^{k-1} g_i^l - x^{\tau t} + \gamma\sum_{l=\tau t}^{k-1} g^l\right\|^2\right] \\
&= \frac{\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\sum_{l=\tau t}^{k-1}\left(g_i^l - g^l\right)\right\|^2\right] \\
&\overset{(141)}{\leq} \frac{e\gamma^2(k-\tau t)}{n}\sum_{i=1}^n\sum_{l=\tau t}^{k-1}\mathbf{E}\left[\left\|\bar{g}_i^l - \bar{g}^l\right\|^2\right] + \frac{e\gamma^2}{n}\sum_{i=1}^n\sum_{l=\tau t}^{k-1}\mathbf{E}\left[\left\|g_i^l - \bar{g}_i^l - \left(g^l - \bar{g}^l\right)\right\|^2\right] \\
&\overset{(139)}{\leq} \frac{e\gamma^2(\tau-1)}{n}\sum_{i=1}^n\sum_{l=\tau t}^{k-1}\mathbf{E}\left[\left\|\bar{g}_i^l\right\|^2\right] + \frac{e\gamma^2}{n}\sum_{i=1}^n\sum_{l=\tau t}^{k-1}\mathbf{E}\left[\left\|g_i^l - \bar{g}_i^l\right\|^2\right],
\end{aligned}$$

where $\bar{g}^k = \frac{1}{n}\sum\limits_{i=1}^n \bar{g}_i^k$. Applying Assumption E.1, we obtain

$$\begin{aligned}
\mathbf{E}V_k \overset{(26),(27)}{\leq}\ & 2e\left(\widetilde{A}(\tau-1)+\widehat{A}\right)\gamma^2\sum_{l=\tau t}^{k-1}\mathbf{E}\left[f(x^l)-f(x^*)\right] + e\left(\widetilde{B}(\tau-1)+\widehat{B}\right)\gamma^2\sum_{l=\tau t}^{k-1}\mathbf{E}\sigma_l^2 \\
&+e\left(\widetilde{F}(\tau-1)+\widehat{F}\right)\gamma^2\sum_{l=\tau t}^{k-1}\mathbf{E}V_l + e(\tau-1)\left(\widetilde{D}_1(\tau-1)+\widehat{D}_1\right)\gamma^2,
\end{aligned}$$

hence

$$\begin{aligned}
\sum_{j=\tau t}^{k} w_j\mathbf{E}V_j \leq\ & 2e\left(\widetilde{A}(\tau-1)+\widehat{A}\right)\gamma^2\sum_{j=\tau t}^{k}\sum_{l=\tau t}^{j-1} w_j\mathbf{E}\left[f(x^l)-f(x^*)\right] + e\left(\widetilde{B}(\tau-1)+\widehat{B}\right)\gamma^2\sum_{j=\tau t}^{k}\sum_{l=\tau t}^{j-1} w_j\mathbf{E}\sigma_l^2 \\
&+e\left(\widetilde{F}(\tau-1)+\widehat{F}\right)\gamma^2\sum_{j=\tau t}^{k}\sum_{l=\tau t}^{j-1} w_j\mathbf{E}V_l + e(\tau-1)\left(\widetilde{D}_1(\tau-1)+\widehat{D}_1\right)\gamma^2\sum_{j=\tau t}^{k} w_j. \qquad (29)
\end{aligned}$$

Recall that $w_k = (1-\eta)^{-(k+1)}$ and $\eta = \min\left\{\gamma\mu, \frac{\rho}{4}\right\}$. Together with our assumption on γ it implies that for all $0 \leq i < k$, $0 \leq j \leq \tau - 1$ we have

$$\begin{aligned}
w_k &= (1-\eta)^{-(k-j+1)}(1-\eta)^{-j} \overset{(137)}{\leq} w_{k-j}(1+2\eta)^j \\
&\leq w_{k-j}(1+2\gamma\mu)^j \leq w_{k-j}\left(1+\frac{1}{2(\tau-1)}\right)^j \leq w_{k-j}\exp\left(\frac{j}{2(\tau-1)}\right) \\
&\leq w_{k-j}\exp\left(\frac{1}{2}\right) \leq 2w_{k-j}, \qquad (30)
\end{aligned}$$

$$w_k = (1-\eta)^{-(k-i+1)}(1-\eta)^{-i} \overset{(137)}{\leq} w_{k-i}(1+2\eta)^i \leq w_{k-i}\left(1+\frac{\rho}{2}\right)^i, \qquad (31)$$

$$w_k \overset{(137)}{\leq} (1+2\eta)^{k+1} \leq \left(1+\frac{\rho}{2}\right)^{k+1}. \qquad (32)$$

For simplicity, we introduce new notation: $r_k \overset{\text{def}}{=} \mathbf{E}\left[f(x^k)-f(x^*)\right]$. Using this we get

$$\begin{aligned}
\sum_{j=\tau t}^{k}\sum_{l=\tau t}^{j-1} w_j r_l &\overset{(30)}{\leq} \sum_{j=\tau t}^{k}\sum_{l=\tau t}^{j-1} 2w_l r_l \leq 2(k-\tau t)\sum_{j=\tau t}^{k} w_j r_j \leq 2(\tau-1)\sum_{j=\tau t}^{k} w_j r_j, \\
\sum_{j=\tau t}^{k}\sum_{l=\tau t}^{j-1} w_j\mathbf{E}\sigma_l^2 &\overset{(30)}{\leq} \sum_{j=\tau t}^{k}\sum_{l=\tau t}^{j-1} 2w_l\mathbf{E}\sigma_l^2 \leq 2(k-\tau t)\sum_{j=\tau t}^{k} w_j\mathbf{E}\sigma_j^2 \leq 2(\tau-1)\sum_{j=\tau t}^{k} w_j\mathbf{E}\sigma_j^2, \\
\sum_{j=\tau t}^{k}\sum_{l=\tau t}^{j-1} w_j\mathbf{E}V_l &\overset{(30)}{\leq} \sum_{j=\tau t}^{k}\sum_{l=\tau t}^{j-1} 2w_l\mathbf{E}V_l \leq 2(k-\tau t)\sum_{j=\tau t}^{k} w_j\mathbf{E}V_j \leq 2(\tau-1)\sum_{j=\tau t}^{k} w_j\mathbf{E}V_j.
\end{aligned}$$

Plugging these inequalities in (29) we derive

$$\begin{aligned}\sum_{j=\tau t}^{k} w_j \mathbf{E}V_j &\leq 4e(\tau-1)(\widetilde{A}(\tau-1)+\widehat{A})\gamma^2 \sum_{j=\tau t}^{k} w_j r_j + 2e(\tau-1)(\widetilde{B}(\tau-1)+\widehat{B})\gamma^2 \sum_{j=\tau t}^{k} w_j \mathbf{E}\sigma_j^2 \\ &\quad +2e(\tau-1)(\widetilde{F}(\tau-1)+\widehat{F})\gamma^2 \sum_{j=\tau t}^{k} w_j \mathbf{E}V_j + e\left(\widetilde{D}_1(\tau-1)+\widehat{D}_1\right)\gamma^2 \sum_{j=\tau t}^{k} w_j.\end{aligned}$$

Since $V_{\tau t}=0$ for all integer $t \geq 0$ we obtain

$$\begin{aligned}\sum_{k=0}^{K} w_k \mathbf{E}V_k &\leq 4e(\tau-1)(\widetilde{A}(\tau-1)+\widehat{A})\gamma^2 \sum_{k=0}^{K} w_k r_k + 2e(\tau-1)(\widetilde{B}(\tau-1)+\widehat{B})\gamma^2 \sum_{k=0}^{K} w_k \mathbf{E}\sigma_k^2 \\ &\quad +2e(\tau-1)(\widetilde{F}(\tau-1)+\widehat{F})\gamma^2 \sum_{k=0}^{K} w_k \mathbf{E}V_k + e\left(\widetilde{D}_1(\tau-1)+\widehat{D}_1\right)\gamma^2 \sum_{k=0}^{K} w_k \qquad (33)\end{aligned}$$

It remains to estimate the second term in the right-hand side of the previous inequality. First of all,

$$\begin{aligned}\mathbf{E}\sigma_{k+1}^2 &\overset{(10)}{\leq} (1-\rho)\mathbf{E}\sigma_k^2 + 2C\underbrace{\mathbf{E}\left[f(x^k)-f(x^*)\right]}_{r_k} + G\mathbf{E}V_k + D_2 \\ &\leq (1-\rho)^{k+1}\mathbf{E}\sigma_0^2 + 2C\sum_{l=0}^{k}(1-\rho)^{k-l}r_l + G\sum_{l=0}^{k}(1-\rho)^{k-l}\mathbf{E}V_l + D_2\sum_{l=0}^{k}(1-\rho)^l \\ &\leq (1-\rho)^{k+1}\mathbf{E}\sigma_0^2 + 2C\sum_{l=0}^{k}(1-\rho)^{k-l}r_l + G\sum_{l=0}^{k}(1-\rho)^{k-l}\mathbf{E}V_l + D_2\sum_{l=0}^{\infty}(1-\rho)^l \\ &= (1-\rho)^{k+1}\mathbf{E}\sigma_0^2 + 2C\sum_{l=0}^{k}(1-\rho)^{k-l}r_l + G\sum_{l=0}^{k}(1-\rho)^{k-l}\mathbf{E}V_l + \frac{D_2}{\rho}. \qquad (34)\end{aligned}$$

It implies that

$$\begin{aligned}\sum_{k=0}^{K} w_k \mathbf{E}\sigma_k^2 &\overset{(34)}{\leq} \mathbf{E}\sigma_0^2\sum_{k=0}^{K} w_k(1-\rho)^k + \frac{2C}{1-\rho}\sum_{k=0}^{K}\sum_{l=0}^{k} w_k(1-\rho)^{k-l}r_l \\ &\quad + \frac{G}{1-\rho}\sum_{k=0}^{K}\sum_{l=0}^{k} w_k(1-\rho)^{k-l}\mathbf{E}V_l + \frac{D_2 W_K}{\rho} \\ &\overset{(31),(32)}{\leq} \mathbf{E}\sigma_0^2\left(1+\frac{\rho}{2}\right)\sum_{k=0}^{K}\left(1+\frac{\rho}{2}\right)^k(1-\rho)^k + \frac{2C}{1-\rho}\sum_{k=0}^{K}\sum_{l=0}^{k} w_l\left(1+\frac{\rho}{2}\right)^{k-l}(1-\rho)^{k-l}r_l \\ &\quad + \frac{G}{1-\rho}\sum_{k=0}^{K}\sum_{l=0}^{k} w_l\left(1+\frac{\rho}{2}\right)^{k-l}(1-\rho)^{k-l}\mathbf{E}V_l + \frac{D_2 W_K}{\rho} \\ &\overset{(138)}{\leq} \mathbf{E}\sigma_0^2\left(1+\frac{\rho}{2}\right)\sum_{k=0}^{K}\left(1-\frac{\rho}{2}\right)^k + \frac{2C}{1-\rho}\sum_{k=0}^{K}\sum_{l=0}^{k} w_l r_l\left(1-\frac{\rho}{2}\right)^{k-l} \\ &\quad + \frac{G}{1-\rho}\sum_{k=0}^{K}\sum_{l=0}^{k} w_l\mathbf{E}V_l\left(1-\frac{\rho}{2}\right)^{k-l} + \frac{D_2 W_K}{\rho} \\ &\leq \mathbf{E}\sigma_0^2\left(1+\frac{\rho}{2}\right)\sum_{k=0}^{\infty}\left(1-\frac{\rho}{2}\right)^k + \frac{2C}{1-\rho}\left(\sum_{k=0}^{K} w_k r_k\right)\left(\sum_{l=0}^{\infty}\left(1-\frac{\rho}{2}\right)^l\right) \\ &\quad + \frac{G}{1-\rho}\left(\sum_{k=0}^{K} w_k\mathbf{E}V_k\right)\left(\sum_{l=0}^{\infty}\left(1-\frac{\rho}{2}\right)^l\right) + \frac{D_2 W_K}{\rho} \\ &= \frac{\mathbf{E}\sigma_0^2(2+\rho)}{\rho} + \frac{4C}{\rho(1-\rho)}\sum_{k=0}^{K} w_k r_k + \frac{2G}{\rho(1-\rho)}\sum_{k=0}^{K} w_k\mathbf{E}V_k + \frac{D_2 W_K}{\rho}. \qquad (35)\end{aligned}$$

Plugging this inequality in (33) we get

$$\begin{aligned}\sum_{k=0}^{K} w_k \mathbf{E}V_k &\leq 4e(\tau-1)\gamma^2\left(\widetilde{A}(\tau-1)+\widehat{A}+\frac{2C(\widetilde{B}(\tau-1)+\widehat{B})}{\rho(1-\rho)}\right)\sum_{k=0}^{K} w_k r_k\\ &\quad+\frac{2e(\tau-1)(\widetilde{B}(\tau-1)+\widehat{B})\mathbf{E}\sigma_0^2(2+\rho)\gamma^2}{\rho}\\ &\quad+2e(\tau-1)\gamma^2\left(\widetilde{F}(\tau-1)+\widehat{F}+\frac{2G(\widetilde{B}(\tau-1)+\widehat{B})}{\rho(1-\rho)}\right)\sum_{k=0}^{K} w_k \mathbf{E}V_k\\ &\quad+e(\tau-1)\gamma^2\left(\widetilde{D}_1(\tau-1)+\widehat{D}_1+\frac{2D_2(\widetilde{B}(\tau-1)+\widehat{B})}{\rho}\right)W_K.\end{aligned}$$

Our choice of γ implies

$$4e(\tau-1)\gamma^2\left(\widetilde{A}(\tau-1)+\widehat{A}+\frac{2C(\widetilde{B}(\tau-1)+\widehat{B})}{\rho(1-\rho)}\right)\leq\frac{1}{8L}$$

and

$$2e(\tau-1)\gamma^2\left(\widetilde{F}(\tau-1)+\widehat{F}+\frac{2G(\widetilde{B}(\tau-1)+\widehat{B})}{\rho(1-\rho)}\right)\leq\frac{1}{2}.$$

Using these inequalities we continue our derivations

$$\begin{aligned}\frac{1}{2}\sum_{k=0}^{K} w_k \mathbf{E}V_k &\leq \frac{1}{8L}\sum_{k=0}^{K} w_k r_k+\frac{2e(\tau-1)(\widetilde{B}(\tau-1)+\widehat{B})\mathbf{E}\sigma_0^2(2+\rho)\gamma^2}{\rho}\\ &\quad+e(\tau-1)\gamma^2\left(\widetilde{D}_1(\tau-1)+\widehat{D}_1+\frac{2D_2(\widetilde{B}(\tau-1)+\widehat{B})}{\rho}\right)W_K.\end{aligned}$$

Multiplying both sides by $4L$ we get the result. □

Clearly, this lemma and Theorem 2.1 imply the following result.

Corollary E.1. *Let the assumptions of Lemma E.1 are satisfied. Then Assumption 2.3 holds and, in particular, if*

$$\begin{aligned}\gamma &\leq \min\left\{\frac{1}{2\left(A'+\frac{4B'C}{3\rho}\right)},\frac{L}{F'+\frac{4B'G}{3\rho}}\right\},\\ \gamma &\leq \min\left\{\frac{1}{4(\tau-1)\mu},\frac{1}{2\sqrt{e(\tau-1)\left(\widetilde{F}(\tau-1)+\widehat{F}+\frac{2G(\widetilde{B}(\tau-1)+\widehat{B})}{\rho(1-\rho)}\right)}}\right\},\\ \gamma &\leq \frac{1}{4\sqrt{2eL(\tau-1)\left(\widetilde{A}(\tau-1)+\widehat{A}+\frac{2C(\widetilde{B}(\tau-1)+\widehat{B})}{\rho(1-\rho)}\right)}},\end{aligned}$$

then for all $K\geq 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K)-f(x^*)\right]\leq\frac{2\|x^0-x^*\|^2+\frac{8B'}{3\rho}\gamma^2\mathbf{E}\sigma_0^2+4LH\gamma\mathbf{E}\sigma_0^2}{\gamma W_K}+2\gamma\left(D_1'+\frac{4B'D_2}{3\rho}+2L\gamma D_3\right),\tag{36}$$

where $\overline{x}^K\overset{def}{=}\frac{1}{W_K}\sum_{k=0}^{K} w_k x^k$ and

$$H=\frac{4e(\tau-1)(\widetilde{B}(\tau-1)+\widehat{B})(2+\rho)\gamma^2}{\rho},\quad D_3=2e(\tau-1)\left(\widetilde{D}_1(\tau-1)+\widehat{D}_1+\frac{2D_2(\widetilde{B}(\tau-1)+\widehat{B})}{\rho}\right).$$

Moreover, if $\mu > 0$, then

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \left(1 - \min\left\{\gamma\mu, \frac{\rho}{4}\right\}\right)^K \frac{2\|x^0 - x^*\|^2 + \frac{8B'}{3\rho}\gamma^2\mathbf{E}\sigma_0^2 + 4LH\gamma\mathbf{E}\sigma_0^2}{\gamma} + 2\gamma\left(D_1' + \frac{4B'D_2}{3\rho} + 2L\gamma D_3\right), \tag{37}$$

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

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{2\|x^0 - x^*\|^2 + \frac{8B'}{3\rho}\gamma^2\mathbf{E}\sigma_0^2 + 4LH\gamma\mathbf{E}\sigma_0^2}{\gamma K} + 2\gamma\left(D_1' + \frac{4B'D_2}{3\rho} + 2L\gamma D_3\right). \tag{38}$$

E.1.2 ζ-Heterogeneous Data

In this section we assume that $f_1, f_2, \ldots, f_n$ are ζ-heterogeneous (see Definition 3.1). Moreover, we additionally assume that $\mathbf{E}\left[g_i^k \mid x_i^k\right] = \nabla f_i(x_i^k)$ and that the functions f_i for $i \in [n]$ are μ-strongly convex,

$$f_i(x) \geq f_i(y) + \langle \nabla f_i(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2 \qquad \forall x, y \in \mathbb{R}^d \tag{39}$$

which implies (e.g., see (Nesterov, 2018))

$$\langle \nabla f_i(x) - \nabla f_i(y), x - y\rangle \geq \mu\|x - y\|^2 \qquad \forall x, y \in \mathbb{R}^d. \tag{40}$$

Lemma E.2. *Let Assumption 2.2 be satisfied, inequalities (7)-(10) hold and*[16]

$$\gamma \leq \min\left\{\frac{1}{4(\tau-1)\mu}, \frac{1}{2\sqrt{(\tau-1)\left(F + \frac{2BG}{\rho(1-\rho)}\right)}}, \frac{1}{4\sqrt{2L(\tau-1)\left(A + \frac{2BC}{\rho(1-\rho)}\right)}}\right\}.$$

Moreover, assume that $f_1, f_2, \ldots, f_n$ are ζ-heterogeneous and μ-strongly convex, and $\mathbf{E}\left[g_i^k \mid x_i^k\right] = \nabla f_i(x_i^k)$ for all $i \in [n]$. Then (11) holds with

$$H = \frac{4B(\tau-1)\gamma^2(2+\rho)}{\rho}, \quad D_3 = 2(\tau-1)\left(D_1 + \frac{\zeta^2}{\gamma\mu} + \frac{2BD_2}{\rho}\right). \tag{41}$$

Proof. First of all, if $k \mod \tau = 0$, then $V_k = 0$ by definition. Otherwise, we have

$$\begin{aligned}
V_k &\overset{(4),(20)}{=} \frac{1}{n}\sum_{i=1}^n \left\|x_i^{k-1} - x^{k-1} - \gamma g_i^{k-1} + \gamma g^{k-1}\right\|^2 \\
&= \frac{1}{n}\sum_{i=1}^n \|x_i^{k-1} - x^{k-1}\|^2 + \frac{2\gamma}{n}\sum_{i=1}^n \left\langle x_i^{k-1} - x^{k-1}, g^{k-1} - g_i^{k-1}\right\rangle + \frac{\gamma^2}{n}\sum_{i=1}^n \|g_i^{k-1} - g^{k-1}\|^2 \\
&= V_{k-1} + 2\gamma\left\langle \frac{1}{n}\sum_{i=1}^n x_i^{k-1} - x^{k-1}, g^{k-1}\right\rangle + \frac{2\gamma}{n}\sum_{i=1}^n \left\langle x^{k-1} - x_i^{k-1}, g_i^{k-1}\right\rangle \\
&\quad + \frac{\gamma^2}{n}\sum_{i=1}^n \|g_i^{k-1} - g^{k-1}\|^2 \\
&= V_{k-1} + \frac{2\gamma}{n}\sum_{i=1}^n \left\langle x^{k-1} - x_i^{k-1}, g_i^{k-1}\right\rangle + \frac{\gamma^2}{n}\sum_{i=1}^n \|g_i^{k-1} - g^{k-1}\|^2.
\end{aligned}$$

[16] When $\rho = 1$ one can always set the parameters in such a way that $B = C = G = 0$, $D_2 = 0$. In this case we assume that $\frac{2BC}{\rho(1-\rho)} = \frac{2BG}{\rho(1-\rho)} = 0$.

Next, we take the conditional expectation $\mathbf{E}\left[\cdot \mid x^{k-1}\right] \stackrel{\text{def}}{=} \mathbf{E}\left[\cdot \mid x_1^{k-1}, \ldots, x_n^{k-1}\right]$ on both sides of the obtained inequality and get

$$\begin{aligned}
\mathbf{E}\left[V_k \mid x^{k-1}\right] &= V_{k-1} + \frac{2\gamma}{n}\sum_{i=1}^n \left\langle x^{k-1} - x_i^{k-1}, \nabla f_i(x_i^{k-1})\right\rangle + \frac{\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^{k-1} - g^{k-1}\|^2 \mid x^{k-1}\right] \\
&\overset{(139)}{\leq} V_{k-1} + \frac{2\gamma}{n}\sum_{i=1}^n \left\langle x^{k-1} - x_i^{k-1}, \nabla f_i(x_i^{k-1}) - \nabla f_i(x^{k-1})\right\rangle \\
&\quad + \frac{2\gamma}{n}\sum_{i=1}^n \left\langle x^{k-1} - x_i^{k-1}, \nabla f_i(x^{k-1})\right\rangle + \frac{\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^{k-1}\|^2 \mid x^{k-1}\right].
\end{aligned}$$

Since $\frac{1}{n}\sum_{i=1}^n \langle x^{k-1} - x_i^{k-1}, \nabla f(x^{k-1})\rangle = 0$, we can continue as follows:

$$\begin{aligned}
\mathbf{E}\left[V_k \mid x^{k-1}\right] &\overset{(40)}{\leq} V_{k-1} - \frac{2\gamma\mu}{n}\sum_{i=1}^n \|x^{k-1} - x_i^{k-1}\|^2 + \frac{\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^{k-1}\|^2 \mid x^{k-1}\right] \\
&\quad + \frac{2\gamma}{n}\sum_{i=1}^n \left\langle x^{k-1} - x_i^{k-1}, \nabla f_i(x^{k-1}) - \nabla f(x^{k-1})\right\rangle \\
&\overset{(132)}{\leq} (1 - 2\gamma\mu)V_{k-1} + \frac{\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^{k-1}\|^2 \mid x^{k-1}\right] \\
&\quad + \frac{2\gamma}{n}\sum_{i=1}^n \left(\frac{\mu}{2}\|x^{k-1} - x_i^{k-1}\|^2 + \frac{1}{2\mu}\|\nabla f_i(x^{k-1}) - \nabla f(x^{k-1})\|^2\right) \\
&\overset{(15)}{\leq} (1 - \gamma\mu)V_{k-1} + \frac{\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^{k-1}\|^2 \mid x^{k-1}\right] + \frac{\gamma\zeta^2}{\mu}.
\end{aligned}$$

Taking full expectation on both sides of previous inequality, we obtain

$$\mathbf{E}V_k \overset{(140)}{\leq} \mathbf{E}\left[V_{k-1}\right] + \frac{\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^{k-1}\|^2\right] + \frac{\gamma\zeta^2}{\mu}.$$

Let t be a non-negative integer for which $\tau t \leq k < \tau(t+1)$. Using this and $V_{\tau t} = 0$, we unroll the recurrence and derive

$$\begin{aligned}
\mathbf{E}[V_k] &\leq \frac{\gamma^2}{n}\sum_{l=\tau t}^{k-1}\sum_{i=1}^n \mathbf{E}\left[\|g_i^l\|^2\right] + \frac{\gamma\zeta^2(k-\tau t)}{\mu} \\
&\overset{(8)}{\leq} \gamma^2 \sum_{l=\tau t}^{k-1}\left(2A\mathbf{E}\left[f(x^l) - f(x^*)\right] + B\mathbf{E}[\sigma_l^2] + F\mathbf{E}[V_l] + D_1\right) + \frac{\gamma\zeta^2(k-\tau t)}{\mu},
\end{aligned}$$

whence

$$\begin{aligned}
\sum_{j=\tau t}^{k} w_j \mathbf{E}V_j &\leq 2A\gamma^2 \sum_{j=\tau t}^{k}\sum_{l=\tau t}^{j-1} w_j \mathbf{E}\left[f(x^l) - f(x^*)\right] + B\gamma^2 \sum_{j=\tau t}^{k}\sum_{l=\tau t}^{j-1} w_j \mathbf{E}\sigma_l^2 \\
&\quad + F\gamma^2 \sum_{j=\tau t}^{k}\sum_{l=\tau t}^{j-1} w_j \mathbf{E}V_l + (\tau - 1)\left(\gamma^2 D_1 + \frac{\gamma\zeta^2}{\mu}\right)\sum_{j=\tau t}^{k} w_j.
\end{aligned}$$

If we substitute A with $e(\widetilde{A}(\tau-1) + \widehat{A})$, B with $e(\widetilde{B}(\tau-1) + \widehat{B})$, F with $e(\widetilde{F}(\tau-1) + \widehat{F})$, and $\left(\gamma^2 D_1 + \frac{\gamma\zeta^2}{\mu}\right)$ with $e\gamma^2(\widetilde{D}_1(\tau-1) + \widehat{D}_1)$ in the inequality above, we will get inequality (29). Following the same steps as in the proof of Lemma E.1, we get

$$\begin{aligned}
\sum_{k=0}^{K} w_k \mathbf{E}V_k &\leq 4(\tau-1)\gamma^2\left(A + \frac{2BC}{\rho(1-\rho)}\right)\sum_{k=0}^{K} w_k r_k + \frac{2B\mathbf{E}\sigma_0^2(2+\rho)(\tau-1)\gamma^2}{\rho} \\
&\quad + 2(\tau-1)\gamma^2\left(F + \frac{2BG}{\rho(1-\rho)}\right)\sum_{k=0}^{K} w_k \mathbf{E}V_k + (\tau-1)\gamma^2\left(D_1 + \frac{\zeta^2}{\gamma\mu} + \frac{2BD_2}{\rho}\right)W_K.
\end{aligned}$$

Our choice of γ implies that

$$4(\tau-1)\gamma^2\left(A+\frac{2BC}{\rho(1-\rho)}\right)\leq\frac{1}{8L}\quad\text{and}\quad 2(\tau-1)\gamma^2\left(F+\frac{2BG}{\rho(1-\rho)}\right)\leq\frac{1}{2}.$$

Using these inequalities we continue our derivations

$$\begin{aligned}\frac{1}{2}\sum_{k=0}^{K}w_k\mathbf{E}V_k &\leq \frac{1}{8L}\sum_{k=0}^{K}w_k r_k+\frac{2B\mathbf{E}\sigma_0^2(2+\rho)(\tau-1)\gamma^2}{\rho}\\ &\quad+(\tau-1)\gamma^2\left(D_1+\frac{\zeta^2}{\gamma\mu}+\frac{2BD_2}{\rho}\right)W_K.\end{aligned}$$

Multiplying both sides by $4L$ we get the result. □

Clearly, this lemma and Theorem 2.1 imply the following result.

Corollary E.2. *Let the assumptions of Lemma E.2 be satisfied. Then Assumption 2.3 holds and, in particular, if*

$$\begin{aligned}\gamma &\leq \min\left\{\frac{1}{2(A'+CM)},\frac{L}{F'+GM}\right\},\quad M=\frac{4B'}{3\rho},\\ \gamma &\leq \min\left\{\frac{1}{4(\tau-1)\mu},\frac{1}{2\sqrt{(\tau-1)\left(F+\frac{2BG}{\rho(1-\rho)}\right)}},\frac{1}{4\sqrt{2L(\tau-1)\left(A+\frac{2BC}{\rho(1-\rho)}\right)}}\right\},\end{aligned}$$

then for all $K\geq 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K)-f(x^*)\right]\leq\frac{2T^0+4LH\gamma\mathbf{E}\sigma_0^2}{\gamma W_K}+2\gamma\left(D_1'+MD_2+2L\gamma D_3\right),\tag{42}$$

where $\overline{x}^K\overset{def}{=}\frac{1}{W_K}\sum_{k=0}^{K}w_k x^k$ and

$$H=\frac{4B(\tau-1)\gamma^2(2+\rho)}{\rho},\quad D_3=2(\tau-1)\left(D_1+\frac{\zeta^2}{\gamma\mu}+\frac{2BD_2}{\rho}\right).$$

Moreover, if $\mu>0$, then

$$\mathbf{E}\left[f(\overline{x}^K)-f(x^*)\right]\leq\left(1-\min\left\{\gamma\mu,\frac{\rho}{4}\right\}\right)^K\frac{2T^0+4LH\gamma\mathbf{E}\sigma_0^2}{\gamma}+2\gamma\left(D_1'+MD_2+2L\gamma D_3\right),\tag{43}$$

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

$$\mathbf{E}\left[f(\overline{x}^K)-f(x^*)\right]\leq\frac{2T^0+4LH\gamma\mathbf{E}\sigma_0^2}{\gamma K}+2\gamma\left(D_1'+MD_2+2L\gamma D_3\right).\tag{44}$$

E.2 Random Local Loop

In this section we show how our results can be applied to analyze (4) in the case when

$$c_k=\begin{cases}1, & \text{with probability } p,\\ 0, & \text{with probability } 1-p,\end{cases}$$

where p encodes the probability of initiating communication. This choice in effect leads to a method using a random-length local loop on all devices.

E.2.1 Heterogeneous Data

As in Section E.1.1, our analysis of (4) with random length of the local loop relies on Assumption E.1. Next lemma provides an upper bound for the weighted sum of $\mathbf{E}[V_k]$ in this case.

Lemma E.3. *Let Assumptions 2.1, 2.2 and E.1 be satisfied and*[17]

$$\gamma \leq \min\left\{\frac{p}{16\mu}, \frac{p}{2\sqrt{(1-p)((2+p)\widetilde{F}+p\widehat{F})}}\right\},$$

$$\gamma \leq \min\left\{\frac{p\sqrt{3\rho(1-\rho)}}{8\sqrt{2G(1-p)\left((p+2)\widetilde{B}+p\widehat{B}\right)}}, \frac{p\sqrt{3}}{16\sqrt{2L(1-p)\left((2+p)\widetilde{A}+p\widehat{A}+\frac{2C\left((p+2)\widetilde{B}+p\widehat{B}\right)}{\rho(1-\rho)}\right)}}\right\}.$$

Then (11) holds with

$$H = \frac{64(1-p)\left((p+2)\widetilde{B}+p\widehat{B}\right)(2+\rho)\gamma^2}{3p^2\rho}, \quad D_3 = \frac{8(1-p)}{p^2}\left((p+2)\widetilde{D}_1+p\widehat{D}_1+\frac{8D_2\left((p+2)\widetilde{B}+p\widehat{B}\right)}{3\rho}\right). \tag{45}$$

Proof. First of all, we introduce new notation: $\mathbf{E}[\cdot \mid x^k, g^k] \overset{\text{def}}{=} \mathbf{E}[\cdot \mid x_1^k, \ldots, x_n^k, g_1^k, \ldots, g_n^k]$, $\mathbf{E}[\cdot \mid x^k] \overset{\text{def}}{=} \mathbf{E}[\cdot \mid x_1^k, \ldots, x_n^k]$. By definition of V_k, we have

$$\begin{aligned}
\mathbf{E}\left[V_{k+1} \mid x^k\right] &\overset{(140)}{=} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\mathbf{E}\left[\|x_i^{k+1}-x^{k+1}\|^2 \mid x^k, g^k\right] \mid x^k\right] \\
&= \frac{1-p}{n}\sum_{i=1}^n \mathbf{E}\left[\|x_i^k - x^k - \gamma g_i^k + \gamma g^k\|^2 \mid x^k\right] \\
&\overset{(139)}{=} \frac{1-p}{n}\sum_{i=1}^n \|x_i^k - x^k - \gamma \bar{g}_i^k + \gamma \bar{g}^k\|^2 + \frac{(1-p)\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k - (g^k - \bar{g}^k)\|^2 \mid x^k\right] \\
&\overset{(135),(139)}{\leq} \frac{(1-p)\left(1+\frac{p}{2}\right)}{n}\sum_{i=1}^n \|x_i^k - x^k\|^2 + \frac{(1-p)\left(1+\frac{2}{p}\right)\gamma^2}{n}\sum_{i=1}^n \|\bar{g}_i^k - \bar{g}^k\|^2 \\
&\quad + \frac{(1-p)\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] \\
&\overset{(138),(139)}{\leq} \left(1-\frac{p}{2}\right)V_k + \frac{(1-p)(2+p)\gamma^2}{pn}\sum_{i=1}^n \left\|\bar{g}_i^k\right\|^2 + \frac{(1-p)\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right],
\end{aligned}$$

where $\bar{g}^k = \mathbf{E}[g^k \mid x^k]$. Taking the full expectation we derive

$$\begin{aligned}
\mathbf{E}\left[V_{k+1}\right] &\leq \left(1-\frac{p}{2}\right)\mathbf{E}\left[V_k\right] + \frac{(1-p)(2+p)\gamma^2}{pn}\sum_{i=1}^n \mathbf{E}\left[\left\|\bar{g}_i^k\right\|^2\right] + \frac{(1-p)\gamma^2}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2\right] \\
&\overset{(26),(27)}{\leq} \left(1-\frac{p}{2}\right)\mathbf{E}\left[V_k\right] + 2(1-p)\gamma^2\left(\frac{2+p}{p}\widetilde{A}+\widehat{A}\right)\mathbf{E}\left[f(x^k)-f(x^*)\right] \\
&\quad + (1-p)\gamma^2\left(\left(\frac{2+p}{p}\widetilde{B}+\widehat{B}\right)\mathbf{E}\sigma_k^2 + \left(\frac{2+p}{p}\widetilde{F}+\widehat{F}\right)\mathbf{E}V_k\right) \\
&\quad + (1-p)\gamma^2\left(\frac{2+p}{p}\widetilde{D}_1+\widehat{D}_1\right).
\end{aligned}$$

[17] When $\rho = 1$ one can always set the parameters in such a way that $\widetilde{B} = \widehat{B} = C = G = 0$, $D_2 = 0$. In this case we assume that $\frac{2\widetilde{B}C}{\rho(1-\rho)} = \frac{2\widehat{B}C}{\rho(1-\rho)} = \frac{2\widetilde{B}G}{\rho(1-\rho)} = \frac{2\widehat{B}G}{\rho(1-\rho)} = 0$.

This inequality together with $\gamma \leq \frac{p}{2\sqrt{(1-p)((2+p)\widetilde{F}+p\widehat{F})}}$ imply

$$\begin{aligned}\mathbf{E}\left[V_{k+1}\right] &\leq \left(1-\frac{p}{4}\right)\mathbf{E}\left[V_k\right] + 2(1-p)\gamma^2\left(\frac{2+p}{p}\widetilde{A}+\widehat{A}\right)\mathbf{E}\left[f(x^k)-f(x^*)\right] \\ &\quad +(1-p)\gamma^2\left(\frac{2+p}{p}\widetilde{B}+\widehat{B}\right)\mathbf{E}\sigma_k^2 + (1-p)\gamma^2\left(\frac{2+p}{p}\widetilde{D}_1+\widehat{D}_1\right).\end{aligned}$$

Unrolling the recurrence, we obtain

$$\begin{aligned}\mathbf{E}\left[V_{k+1}\right] &\leq 2(1-p)\gamma^2\left(\frac{2+p}{p}\widetilde{A}+\widehat{A}\right)\sum_{l=0}^{k}\left(1-\frac{p}{4}\right)^{k-l}\mathbf{E}\left[f(x^l)-f(x^*)\right] \\ &\quad +(1-p)\gamma^2\left(\frac{2+p}{p}\widetilde{B}+\widehat{B}\right)\sum_{l=0}^{k}\left(1-\frac{p}{4}\right)^{k-l}\mathbf{E}\sigma_l^2 \\ &\quad +(1-p)\gamma^2\left(\frac{2+p}{p}\widetilde{D}_1+\widehat{D}_1\right)\sum_{l=0}^{k}\left(1-\frac{p}{4}\right)^{k-l}.\end{aligned}$$

As a consequence, we derive

$$\begin{aligned}\sum_{k=0}^{K} w_k\mathbf{E}\left[V_k\right] &\leq \frac{2(1-p)\left((2+p)\widetilde{A}+p\widehat{A}\right)\gamma^2}{p\left(1-\frac{p}{4}\right)}\sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{p}{4}\right)^{k-l}w_k r_l \\ &\quad +\frac{(1-p)\left((2+p)\widetilde{B}+p\widehat{B}\right)\gamma^2}{p\left(1-\frac{p}{4}\right)}\sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{p}{4}\right)^{k-l}w_k\mathbf{E}\left[\sigma_l^2\right] \\ &\quad +\frac{(1-p)\left((2+p)\widetilde{D}_1+p\widehat{D}_1\right)\gamma^2}{p}\sum_{k=0}^{K}\sum_{l=0}^{k-1}\left(1-\frac{p}{4}\right)^{k-1-l}w_k, \qquad (46)\end{aligned}$$

where we use new notation: $r_l = \mathbf{E}\left[f(x^l)-f(x^*)\right]$. Recall that $w_k = (1-\eta)^{-(k+1)}$ and $\eta = \min\left\{\gamma\mu, \frac{\rho}{4}\right\}$. Together with our assumption on γ it implies that for all $0 \leq i < k$ we have

$$\begin{aligned}w_k &= (1-\eta)^{-(k-i+1)}(1-\eta)^{-i} \overset{(137)}{\leq} w_{k-i}(1+2\eta)^i \\ &\leq w_{k-i}(1+2\gamma\mu)^i \leq w_{k-i}\left(1+\frac{p}{8}\right)^i, \qquad (47)\\ w_k &= (1-\eta)^{-(k-i+1)}(1-\eta)^{-i} \overset{(137)}{\leq} w_{k-i}(1+2\eta)^i \leq w_{k-i}\left(1+\frac{\rho}{2}\right)^i, \qquad (48)\\ w_k &\overset{(137)}{\leq} (1+2\eta)^{k+1} \leq \left(1+\frac{\rho}{2}\right)^{k+1}. \qquad (49)\end{aligned}$$

Having these inequalities in hand we obtain

$$\begin{aligned}\sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{p}{4}\right)^{k-l}w_k r_l &\overset{(47)}{\leq} \sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{p}{4}\right)^{k-l}\left(1+\frac{p}{8}\right)^{k-l}w_l r_l \\ &\overset{(138)}{\leq} \sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{p}{8}\right)^{k-l}w_l r_l \leq \left(\sum_{k=0}^{K}w_k r_k\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{p}{8}\right)^k\right) \\ &= \frac{8}{p}\sum_{k=0}^{K}w_k r_k,\end{aligned}$$

$$\sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{p}{4}\right)^{k-l}w_k\mathbf{E}\left[\sigma_l^2\right] \overset{(47)}{\leq} \sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{p}{4}\right)^{k-l}\left(1+\frac{p}{8}\right)^{k-l}w_l\mathbf{E}\left[\sigma_l^2\right]$$
$$\overset{(138)}{\leq} \sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{p}{8}\right)^{k-l}w_l\mathbf{E}\left[\sigma_l^2\right] \leq \left(\sum_{k=0}^{K}w_k\mathbf{E}\left[\sigma_k^2\right]\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{p}{8}\right)^k\right)$$
$$= \frac{8}{p}\sum_{k=0}^{K}w_k\mathbf{E}\left[\sigma_k^2\right],$$

and

$$\sum_{k=0}^{K}\sum_{l=0}^{k-1}\left(1-\frac{p}{4}\right)^{k-1-l}w_k \leq \left(\sum_{k=0}^{K}w_k\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{p}{4}\right)^k\right) = \frac{4W_K}{p}.$$

Plugging these inequalities together with $1-\frac{p}{4}\geq\frac{3}{4}$ in (46), we derive

$$\sum_{k=0}^{K}w_k\mathbf{E}\left[V_k\right] \leq \frac{64(1-p)\left((2+p)\widetilde{A}+p\widehat{A}\right)\gamma^2}{3p^2}\sum_{k=0}^{K}w_kr_k + \frac{32(1-p)\left((2+p)\widetilde{B}+p\widehat{B}\right)\gamma^2}{3p^2}\sum_{k=0}^{K}w_k\mathbf{E}\left[\sigma_k^2\right]$$
$$+\frac{4(1-p)\left((2+p)\widetilde{D}_1+p\widehat{D}_1\right)\gamma^2}{p^2}W_K. \tag{50}$$

It remains to estimate the second term on the right-hand side of this inequality. We notice that an analogous term appears in the proof of Lemma E.1. In particular, in that proof inequality (35) was shown via inequalities (10), (48), (49) and (138) which hold in this case too. Therefore, we get that

$$\sum_{k=0}^{K}w_k\mathbf{E}\left[\sigma_k^2\right] \overset{(35)}{\leq} \frac{\mathbf{E}\sigma_0^2(2+\rho)}{\rho} + \frac{4C}{\rho(1-\rho)}\sum_{k=0}^{K}w_kr_k + \frac{2G}{\rho(1-\rho)}\sum_{k=0}^{K}w_k\mathbf{E}V_k + \frac{D_2W_K}{\rho},$$

whence

$$\sum_{k=0}^{K}w_k\mathbf{E}\left[V_k\right] \overset{(50)}{\leq} \frac{64(1-p)\gamma^2\left((2+p)\widetilde{A}+p\widehat{A}+\frac{2C\left((p+2)\widetilde{B}+p\widehat{B}\right)}{\rho(1-\rho)}\right)}{3p^2}\sum_{k=0}^{K}w_kr_k$$
$$+\frac{32(1-p)\left((p+2)\widetilde{B}+p\widehat{B}\right)(2+\rho)\gamma^2\mathbf{E}\sigma_0^2}{3p^2\rho}$$
$$+\frac{64G(1-p)\left((p+2)\widetilde{B}+p\widehat{B}\right)\gamma^2}{3p^2\rho(1-\rho)}\sum_{k=0}^{K}w_k\mathbf{E}\left[V_k\right]$$
$$+\frac{4(1-p)\gamma^2}{p^2}\left((p+2)\widetilde{D}_1+p\widehat{D}_1+\frac{8D_2\left((p+2)\widetilde{B}+p\widehat{B}\right)}{3\rho}\right)W_K.$$

Our assumptions on γ imply

$$\frac{64(1-p)\gamma^2\left((2+p)\widetilde{A}+p\widehat{A}+\frac{2C\left((p+2)\widetilde{B}+p\widehat{B}\right)}{\rho(1-\rho)}\right)}{3p^2}\leq\frac{1}{8L},\quad \frac{64G(1-p)\left((p+2)\widetilde{B}+p\widehat{B}\right)\gamma^2}{3p^2\rho(1-\rho)}\leq\frac{1}{2}.$$

Next, we introduce new notation as follows:

$$H=\frac{64(1-p)\left((p+2)\widetilde{B}+p\widehat{B}\right)(2+\rho)\gamma^2}{3p^2\rho},\quad D_3=\frac{8(1-p)}{p^2}\left((p+2)\widetilde{D}_1+p\widehat{D}_1+\frac{8D_2\left((p+2)\widetilde{B}+p\widehat{B}\right)}{3\rho}\right).$$

Putting all together, we get

$$\frac{1}{2}\sum_{k=0}^{K} w_k \mathbf{E}\left[V_k\right] \leq \frac{1}{8L}\sum_{k=0}^{K} w_k r_k + \frac{H}{2}\mathbf{E}\sigma_0^2 + \frac{D_3}{2}\gamma^2 W_K,$$

which concludes the proof. □

This lemma and Theorem 2.1 imply the following result.

Corollary E.3. *Let the assumptions of Lemma E.3 be satisfied. Then Assumption 2.3 holds and, in particular, if*

$$\gamma \leq \min\left\{\frac{1}{2\left(A' + \frac{4B'C}{3\rho}\right)}, \frac{L}{F' + \frac{4B'G}{3\rho}}, \frac{p}{16\mu}, \frac{p}{2\sqrt{(1-p)((2+p)\widetilde{F} + p\widehat{F})}}\right\},$$

$$\gamma \leq \min\left\{\frac{p\sqrt{3\rho(1-\rho)}}{8\sqrt{2G(1-p)\left((p+2)\widetilde{B} + p\widehat{B}\right)}}, \frac{p\sqrt{3}}{16\sqrt{2L(1-p)\left((2+p)\widetilde{A} + p\widehat{A} + \frac{2C\left((p+2)\widetilde{B}+p\widehat{B}\right)}{\rho(1-\rho)}\right)}}\right\},$$

then for all $K \geq 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{2\|x^0 - x^*\|^2 + \frac{8B'}{3\rho}\gamma^2\mathbf{E}\sigma_0^2 + 4LH\gamma\mathbf{E}\sigma_0^2}{\gamma W_K} + 2\gamma\left(D_1' + \frac{4B'D_2}{3\rho} + 2L\gamma D_3\right), \tag{51}$$

where $\overline{x}^K \overset{def}{=} \frac{1}{W_K}\sum_{k=0}^{K} w_k x^k$ and

$$H = \frac{64(1-p)\left((p+2)\widetilde{B} + p\widehat{B}\right)(2+\rho)\gamma^2}{3p^2\rho}, \quad D_3 = \frac{8(1-p)}{p^2}\left((p+2)\widetilde{D}_1 + p\widehat{D}_1 + \frac{8D_2\left((p+2)\widetilde{B} + p\widehat{B}\right)}{3\rho}\right).$$

Moreover, if $\mu > 0$, then

$$\begin{aligned}\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq{}& \left(1 - \min\left\{\gamma\mu, \frac{\rho}{4}\right\}\right)^K \frac{2\|x^0 - x^*\|^2 + \frac{8B'}{3\rho}\gamma^2\mathbf{E}\sigma_0^2 + 4LH\gamma\mathbf{E}\sigma_0^2}{\gamma} \\ &+2\gamma\left(D_1' + \frac{4B'D_2}{3\rho} + 2L\gamma D_3\right),\end{aligned} \tag{52}$$

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

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{2\|x^0 - x^*\|^2 + \frac{8B'}{3\rho}\gamma^2\mathbf{E}\sigma_0^2 + 4LH\gamma\mathbf{E}\sigma_0^2}{\gamma K} + 2\gamma\left(D_1' + \frac{4B'D_2}{3\rho} + 2L\gamma D_3\right). \tag{53}$$

E.2.2 ζ-Heterogeneous Data

In this section we assume that $f_1, f_2, \ldots, f_n$ are ζ-heterogeneous (see Definition 3.1). Moreover, we additionally assume that $\mathbf{E}\left[g_i^k \mid x_i^k\right] = \nabla f_i(x_i^k)$ and we also assume μ-strong convexity of the functions f_i for $i \in [n]$.

Lemma E.4. *Let Assumption 2.2 be satisfied, inequalities (7)-(10) hold and*[18]

$$\gamma \leq \min\left\{\frac{p}{8\mu}, \sqrt{\frac{p}{2F(1-p)}}, \sqrt{\frac{p\rho(1-\rho)}{32BG(1-p)}}, \sqrt{\frac{p}{128L(1-p)\left(A + \frac{2BC}{\rho(1-\rho)}\right)}}\right\}.$$

Moreover, assume that $f_1, f_2, \ldots, f_n$ are ζ-heterogeneous and μ-strongly convex, and $\mathbf{E}\left[g_i^k \mid x_i^k\right] = \nabla f_i(x_i^k)$ for all $i \in [n]$. Then (11) holds with

$$H = \frac{16B(1-p)(2+\rho)\gamma^2}{p\rho}, \quad D_3 = \frac{4(1-p)}{p}\left(D_1 + \frac{\zeta^2}{\gamma\mu} + \frac{4BD_2}{\rho}\right). \tag{54}$$

[18] When $\rho = 1$ one can always set the parameters in such a way that $B = C = G = 0$, $D_2 = 0$. In this case we assume that $\frac{2BC}{\rho(1-\rho)} = \frac{2BG}{\rho(1-\rho)} = 0$.

Proof. First of all, we introduce new notation: $\mathbf{E}[\cdot \mid x^k, g^k] \stackrel{\text{def}}{=} \mathbf{E}[\cdot \mid x_1^k, \ldots, x_n^k, g_1^k, \ldots, g_n^k]$. By definition of V_k for all $k \geq 1$ we have

$$
\begin{aligned}
\mathbf{E}[V_k \mid x^{k-1}, g^{k-1}] &\overset{(4),(20)}{=} \frac{1-p}{n}\sum_{i=1}^n \left\|x_i^{k-1} - x^{k-1} - \gamma g_i^{k-1} + \gamma g^{k-1}\right\|^2 \\
&= \frac{1-p}{n}\sum_{i=1}^n \|x_i^{k-1} - x^{k-1}\|^2 + \frac{2\gamma(1-p)}{n}\sum_{i=1}^n \left\langle x_i^{k-1} - x^{k-1}, g^{k-1} - g_i^{k-1}\right\rangle \\
&\quad + \frac{\gamma^2(1-p)}{n}\sum_{i=1}^n \|g_i^{k-1} - g^{k-1}\|^2 \\
&= (1-p)V_{k-1} + 2\gamma(1-p)\left\langle \frac{1}{n}\sum_{i=1}^n x_i^{k-1} - x^{k-1}, g^{k-1}\right\rangle \\
&\quad + \frac{2\gamma(1-p)}{n}\sum_{i=1}^n \left\langle x^{k-1} - x_i^{k-1}, g_i^{k-1}\right\rangle + \frac{\gamma^2(1-p)}{n}\sum_{i=1}^n \|g_i^{k-1} - g^{k-1}\|^2 \\
&= (1-p)V_{k-1} + \frac{2\gamma(1-p)}{n}\sum_{i=1}^n \left\langle x^{k-1} - x_i^{k-1}, g_i^{k-1}\right\rangle \\
&\quad + \frac{\gamma^2(1-p)}{n}\sum_{i=1}^n \|g_i^{k-1} - g^{k-1}\|^2.
\end{aligned}
$$

Next, we take the conditional expectation $\mathbf{E}\left[\cdot \mid x^{k-1}\right] \stackrel{\text{def}}{=} \mathbf{E}\left[\cdot \mid x_1^{k-1}, \ldots, x_n^{k-1}\right]$ on both sides of the obtained inequality and get

$$
\begin{aligned}
\mathbf{E}\left[V_k \mid x^{k-1}\right] &= (1-p)V_{k-1} + \frac{2\gamma(1-p)}{n}\sum_{i=1}^n \left\langle x^{k-1} - x_i^{k-1}, \nabla f_i(x_i^{k-1})\right\rangle \\
&\quad + \frac{\gamma^2(1-p)}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^{k-1} - g^{k-1}\|^2 \mid x^{k-1}\right] \\
&\overset{(139)}{\leq} (1-p)V_{k-1} + \frac{2\gamma(1-p)}{n}\sum_{i=1}^n \left\langle x^{k-1} - x_i^{k-1}, \nabla f_i(x_i^{k-1}) - \nabla f_i(x^{k-1})\right\rangle \\
&\quad + \frac{2\gamma(1-p)}{n}\sum_{i=1}^n \left\langle x^{k-1} - x_i^{k-1}, \nabla f_i(x^{k-1})\right\rangle \\
&\quad + \frac{\gamma^2(1-p)}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^{k-1}\|^2 \mid x^{k-1}\right].
\end{aligned}
$$

Since $\frac{1}{n}\sum_{i=1}^n \langle x^{k-1} - x_i^{k-1}, \nabla f(x^{k-1})\rangle = 0$, we can continue as follows:

$$\begin{aligned}
\mathbf{E}\left[V_k \mid x^{k-1}\right] &\overset{(40)}{\leq} (1-p)V_{k-1} - \frac{2\gamma\mu(1-p)}{n}\sum_{i=1}^n \|x^{k-1} - x_i^{k-1}\|^2 \\
&\quad + \frac{2\gamma(1-p)}{n}\sum_{i=1}^n \left\langle x^{k-1} - x_i^{k-1}, \nabla f_i(x^{k-1}) - \nabla f(x^{k-1})\right\rangle \\
&\quad + \frac{\gamma^2(1-p)}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^{k-1}\|^2 \mid x^{k-1}\right] \\
&\overset{(132)}{\leq} (1-p)(1-2\gamma\mu)V_{k-1} + \frac{\gamma^2(1-p)}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^{k-1}\|^2 \mid x^{k-1}\right] \\
&\quad + \frac{2\gamma(1-p)}{n}\sum_{i=1}^n \left(\frac{\mu}{2}\|x^{k-1} - x_i^{k-1}\|^2 + \frac{1}{2\mu}\|\nabla f_i(x^{k-1}) - \nabla f(x^{k-1})\|^2\right) \\
&\overset{(15)}{\leq} (1-p)(1-\gamma\mu)V_{k-1} + \frac{\gamma^2(1-p)}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^{k-1}\|^2 \mid x^{k-1}\right] + \frac{(1-p)\gamma\zeta^2}{\mu}.
\end{aligned}$$

Taking full mathematical expectation on both sides of previous inequality and using $1-\gamma\mu \leq 1$ we obtain

$$\begin{aligned}
\mathbf{E}V_k &\overset{(140)}{\leq} (1-p)\mathbf{E}\left[V_{k-1}\right] + \frac{\gamma^2(1-p)}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^{k-1}\|^2\right] + \frac{(1-p)\gamma\zeta^2}{\mu} \\
&\overset{(8)}{\leq} (1-p)\mathbf{E}[V_{k-1}] + (1-p)\gamma^2\left(2A\mathbf{E}[f(x^{k-1}) - f(x^*)] + B\mathbf{E}[\sigma_k^2] + F\mathbf{E}[V_{k-1}] + D_1\right) \\
&\quad + \frac{(1-p)\gamma\zeta^2}{\mu}.
\end{aligned}$$

Since $\gamma \leq \sqrt{\frac{p}{2F(1-p)}}$ we have $(1-p)\gamma^2 F \leq \frac{p}{2}$ and

$$\mathbf{E}V_k \leq \left(1 - \frac{p}{2}\right)\mathbf{E}[V_{k-1}] + (1-p)\gamma^2\left(2A\mathbf{E}[f(x^{k-1}) - f(x^*)] + B\mathbf{E}[\sigma_k^2] + D_1 + \frac{\zeta^2}{\gamma\mu}\right).$$

Unrolling the recurrence we obtain

$$\mathbf{E}\left[V_k\right] \leq (1-p)\gamma^2\sum_{l=0}^{k-1}\left(1 - \frac{p}{2}\right)^{k-1-l}\left(2A\mathbf{E}\left[f(x^l) - f(x^*)\right] + B\mathbf{E}\left[\sigma_l^2\right] + D_1 + \frac{\zeta^2}{\gamma\mu}\right).$$

As a consequence, we derive

$$\begin{aligned}
\sum_{k=0}^K w_k\mathbf{E}\left[V_k\right] &\leq \frac{2A(1-p)\gamma^2}{1-\frac{p}{2}}\sum_{k=0}^K\sum_{l=0}^k\left(1-\frac{p}{2}\right)^{k-l} w_k r_l \\
&\quad + \frac{B(1-p)\gamma^2}{1-\frac{p}{2}}\sum_{k=0}^K\sum_{l=0}^k\left(1-\frac{p}{2}\right)^{k-l} w_k\mathbf{E}\left[\sigma_l^2\right] \\
&\quad + \left(D_1 + \frac{\zeta^2}{\gamma\mu}\right)(1-p)\gamma^2\sum_{k=0}^K\sum_{l=0}^{k-1}\left(1-\frac{p}{2}\right)^{k-1-l} w_k,
\end{aligned} \tag{55}$$

where we use new notation: $r_l = \mathbf{E}\left[f(x^l) - f(x^*)\right]$. Recall that $w_k = (1-\eta)^{-(k+1)}$ and $\eta = \min\left\{\gamma\mu, \frac{\rho}{4}\right\}$.

Together with our assumption on γ it implies that for all $0 \leq i < k$ we have

$$\begin{aligned} w_k &= (1-\eta)^{-(k-i+1)}(1-\eta)^{-i} \overset{(137)}{\leq} w_{k-i}(1+2\eta)^i \\ &\leq w_{k-i}(1+2\gamma\mu)^i \leq w_{k-i}\left(1+\frac{p}{4}\right)^i, \end{aligned} \tag{56}$$

$$w_k = (1-\eta)^{-(k-i+1)}(1-\eta)^{-i} \overset{(137)}{\leq} w_{k-i}(1+2\eta)^i \leq w_{k-i}\left(1+\frac{\rho}{2}\right)^i, \tag{57}$$

$$w_k \overset{(137)}{\leq} (1+2\eta)^{k+1} \leq \left(1+\frac{\rho}{2}\right)^{k+1}. \tag{58}$$

Having these inequalities in hand we obtain

$$\begin{aligned} \sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{p}{2}\right)^{k-l} w_k r_l &\overset{(56)}{\leq} \sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{p}{2}\right)^{k-l}\left(1+\frac{p}{4}\right)^{k-l} w_l r_l \\ &\overset{(138)}{\leq} \sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{p}{4}\right)^{k-l} w_l r_l \leq \left(\sum_{k=0}^{K} w_k r_k\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{p}{4}\right)^k\right) \\ &= \frac{4}{p}\sum_{k=0}^{K} w_k r_k, \end{aligned}$$

$$\begin{aligned} \sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{p}{2}\right)^{k-l} w_k \mathbf{E}\left[\sigma_l^2\right] &\overset{(56)}{\leq} \sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{p}{2}\right)^{k-l}\left(1+\frac{p}{4}\right)^{k-l} w_l \mathbf{E}\left[\sigma_l^2\right] \\ &\overset{(138)}{\leq} \sum_{k=0}^{K}\sum_{l=0}^{k}\left(1-\frac{p}{4}\right)^{k-l} w_l \mathbf{E}\left[\sigma_l^2\right] \leq \left(\sum_{k=0}^{K} w_k \mathbf{E}\left[\sigma_k^2\right]\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{p}{4}\right)^k\right) \\ &= \frac{4}{p}\sum_{k=0}^{K} w_k \mathbf{E}\left[\sigma_k^2\right], \end{aligned}$$

and

$$\sum_{k=0}^{K}\sum_{l=0}^{k-1}\left(1-\frac{p}{2}\right)^{k-1-l} w_k \leq \left(\sum_{k=0}^{K} w_k\right)\left(\sum_{k=0}^{\infty}\left(1-\frac{p}{2}\right)^k\right) = \frac{2W_K}{p}.$$

Plugging these inequalities together with $1-\frac{p}{2} \geq \frac{1}{2}$ in (55) we derive

$$\begin{aligned} \sum_{k=0}^{K} w_k \mathbf{E}\left[V_k\right] &\leq \frac{16A(1-p)\gamma^2}{p}\sum_{k=0}^{K} w_k r_k + \frac{8B(1-p)\gamma^2}{p}\sum_{k=0}^{K} w_k \mathbf{E}\left[\sigma_k^2\right] \\ &\quad + \frac{2\left(D_1 + \frac{\zeta^2}{\gamma\mu}\right)(1-p)\gamma^2}{p} W_K. \end{aligned} \tag{59}$$

It remains to estimate the second term in the right-hand side of this inequality. We notice that an analogous term appear in the proof of Lemma E.1. In particular, in that proof inequality (35) was shown via inequalities (10), (48), (49) and (138) which hold in this case too. Therefore, we get that

$$\sum_{k=0}^{K} w_k \mathbf{E}\left[\sigma_k^2\right] \overset{(35)}{\leq} \frac{\mathbf{E}\sigma_0^2(2+\rho)}{\rho} + \frac{4C}{\rho(1-\rho)}\sum_{k=0}^{K} w_k r_k + \frac{2G}{\rho(1-\rho)}\sum_{k=0}^{K} w_k \mathbf{E} V_k + \frac{D_2 W_K}{\rho},$$

hence

$$\sum_{k=0}^{K} w_k \mathbf{E}\left[V_k\right] \overset{(50)}{\leq} \frac{16(1-p)\gamma^2\left(A+\frac{2BC}{\rho(1-\rho)}\right)}{p}\sum_{k=0}^{K} w_k r_k + \frac{8B(1-p)(2+\rho)\gamma^2\mathbf{E}\sigma_0^2}{p\rho} + \frac{16BG(1-p)\gamma^2}{p\rho(1-\rho)}\sum_{k=0}^{K} w_k \mathbf{E}\left[V_k\right] + \frac{2(1-p)\gamma^2}{p}\left(D_1+\frac{\zeta^2}{\gamma\mu}+\frac{4BD_2}{\rho}\right)W_K.$$

Our assumption on γ imply

$$\frac{16(1-p)\gamma^2\left(A+\frac{2BC}{\rho(1-\rho)}\right)}{p} \leq \frac{1}{8L}, \quad \frac{16BG(1-p)\gamma^2}{p\rho(1-\rho)} \leq \frac{1}{2}.$$

Next, we introduce new notation as follows:

$$H = \frac{16B(1-p)(2+\rho)\gamma^2}{p\rho}, \quad D_3 = \frac{4(1-p)}{p}\left(D_1+\frac{\zeta^2}{\gamma\mu}+\frac{4BD_2}{\rho}\right).$$

Putting all together we get

$$\frac{1}{2}\sum_{k=0}^{K} w_k \mathbf{E}\left[V_k\right] \leq \frac{1}{8L}\sum_{k=0}^{K} w_k r_k + \frac{H}{2}\mathbf{E}\sigma_0^2 + \frac{D_3}{2}\gamma^2 W_K$$

which concludes the proof. □

This lemma and Theorem 2.1 imply the following result.

Corollary E.4. *Let the assumptions of Lemma E.4 are satisfied. Then Assumption 2.3 holds and, in particular, if*

$$\gamma \leq \min\left\{\frac{1}{2(A'+CM)}, \frac{L}{F'+GM}, \frac{p}{8\mu}\right\}, \quad M = \frac{4B'}{3\rho},$$

$$\gamma \leq \min\left\{\sqrt{\frac{p}{2F(1-p)}}, \sqrt{\frac{p\rho(1-\rho)}{32BG(1-p)}}, \sqrt{\frac{p}{128L(1-p)\left(A+\frac{2BC}{\rho(1-\rho)}\right)}}\right\},$$

then for all $K \geq 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{2T^0 + 4LH\gamma\mathbf{E}\sigma_0^2}{\gamma W_K} + 2\gamma\left(D_1' + MD_2 + 2L\gamma D_3\right), \tag{60}$$

where $\overline{x}^K \overset{def}{=} \frac{1}{W_K}\sum_{k=0}^{K} w_k x^k$ and

$$H = \frac{16B(1-p)(2+\rho)\gamma^2}{p\rho}, \quad D_3 = \frac{4(1-p)}{p}\left(D_1+\frac{\zeta^2}{\gamma\mu}+\frac{4BD_2}{\rho}\right).$$

Moreover, if $\mu > 0$, then

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \left(1-\min\left\{\gamma\mu, \frac{\rho}{4}\right\}\right)^K \frac{2T^0 + 4LH\gamma\mathbf{E}\sigma_0^2}{\gamma} + 2\gamma\left(D_1' + MD_2 + 2L\gamma D_3\right), \tag{61}$$

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

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{2T^0 + 4LH\gamma\mathbf{E}\sigma_0^2}{\gamma K} + 2\gamma\left(D_1' + MD_2 + 2L\gamma D_3\right). \tag{62}$$

F Missing Parts from Section 4

Let us start with an useful Lemma that bounds the Bregman distance between the local iterate x_i^k and the optimum x^* by the Bregman distance between the virtual iterate x^k and the optimum.

Lemma F.1. *Assume f_i is L-smooth for all $i \in [n]$. Then*

$$D_{f_i}(x_i^k, x^*) \leq 2D_{f_i}(x^k, x^*) + L\|x_i^k - x^k\|^2 \quad \forall i \in [n]. \tag{63}$$

Proof. Using corollaries of L-smoothness and Young's inequality, we derive

$$\begin{aligned} D_{f_i}(x_i^k, x^*) &\overset{(18)}{\leq} D_{f_i}(x^k, x^*) + \langle \nabla f_i(x^k) - \nabla f_i(x^*), x_i^k - x^k \rangle + \frac{L}{2}\|x_i^k - x^k\|^2 \\ &\overset{(132)}{\leq} D_{f_i}(x^k, x^*) + \frac{1}{2L}\|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 + L\|x_i^k - x^k\|^2 \\ &\overset{(6)}{\leq} 2D_{f_i}(x^k, x^*) + L\|x_i^k - x^k\|^2. \end{aligned}$$

□

F.1 Proof of Lemma 4.1

Let us bound $\frac{1}{n}\sum\limits_{i=1}^n \mathbf{E}_k\left[\|g_i^k\|^2\right]$ first:

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n \mathbf{E}_k\left[\|g_i^k\|^2\right] &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}_k\left[\|a_i^k - b_i^k\|^2\right] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}_k\left[\|a_i^k - \nabla f_i(x^*) - (b_i^k - \nabla f_i(x^*))\|^2\right] \\ &\leq \frac{2}{n}\sum_{i=1}^n \mathbf{E}_k\left[\|a_i^k - \nabla f_i(x^*)\|^2 + \|b_i^k - \nabla f_i(x^*)\|^2\right] \\ &\leq \frac{2}{n}\sum_{i=1}^n \left(2A_i D_{f_i}(x_i^k, x^*) + B_i\sigma_{i,k}^2 + D_{1,i} + \mathbf{E}_k\left[\|b_i^k - \nabla f_i(x^*)\|^2\right]\right) \\ &\overset{(63)}{\leq} \frac{2}{n}\sum_{i=1}^n \left(4A_i D_{f_i}(x^k, x^*) + 2A_i L\|x_i^k - x^k\|^2 + B_i\sigma_{i,k}^2 + D_{1,i} + \mathbf{E}_k\left[\|b_i^k - \nabla f_i(x^*)\|^2\right]\right) \\ &\leq 8\max_i\{A_i\}(f(x^k) - f(x^*)) + 4\max_i\{A_i\}LV_k + \frac{2}{n}\sum_{i=1}^n \left(B_i\sigma_{i,k}^2 + D_{1,i} + \mathbf{E}_k\left[\|b_i^k - \nabla f_i(x^*)\|^2\right]\right). \end{aligned}$$

Taking the full expectation, we arrive at

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k\|^2\right] \leq 8\max_i\{A_i\}\mathbf{E}(f(x^k) - f(x^*)) + 4\max_i\{A_i\}L\mathbf{E}V_k + \frac{2}{n}\sum_{i=1}^n \left(B_i\mathbf{E}\sigma_{i,k}^2 + D_{1,i} + \mathbf{E}\|b_i^k - \nabla f_i(x^*)\|^2\right). \tag{64}$$

Next, we have

$$
\begin{aligned}
\mathbf{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k\right\|^2\right] &= \mathbf{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n a_i^k - b_i^k\right\|^2\right] \\
&= \mathbf{E}_k\left[\left\|\frac{1}{n}\sum_{i=1}^n a_i^k - \nabla f_i(x^*)\right\|^2\right] \\
&= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n a_i^k - \nabla f_i(x^*)\right] + \left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k) - \nabla f_i(x^*)\right\|^2 \\
&\le \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n a_i^k - \nabla f_i(x^*)\right] + \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(x_i^k) - \nabla f_i(x^*)\right\|^2 \\
&\le \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n a_i^k - \nabla f_i(x^*)\right] + \frac{2L}{n}\sum_{i=1}^n D_{f_i}(x_i^k, x^*) \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}\left[a_i^k - \nabla f_i(x^*)\right] + \frac{2L}{n}\sum_{i=1}^n D_{f_i}(x_i^k, x^*) \\
&\le \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}_k\left[\left\|a_i^k - \nabla f_i(x^*)\right\|^2\right] + \frac{2L}{n}\sum_{i=1}^n D_{f_i}(x_i^k, x^*) \\
&\le \frac{1}{n^2}\sum_{i=1}^n \left(2A_i D_{f_i}(x_i^k, x^*) + B_i\sigma_{i,k}^2 + D_{1,i}\right) + \frac{2L}{n}\sum_{i=1}^n D_{f_i}(x_i^k, x^*) \\
&\le \frac{1}{n^2}\sum_{i=1}^n \left(2\left(\max_i\{A_i\} + nL\right) D_{f_i}(x_i^k, x^*) + B_i\sigma_{i,k}^2 + D_{1,i}\right) \\
&\overset{(63)}{\le} \left(\frac{4\max_i\{A_i\}}{n} + 2L\right) D_f(x^k, x^*) + \frac{1}{n^2}\sum_{i=1}^n \left(2(\max_i\{A_i\}L + nL^2)\|x_i^k - x^*\|^2 + B_i\sigma_{i,k}^2 + D_{1,i}\right) \\
&= \left(\frac{4\max_i\{A_i\}}{n} + 2L\right)\left(f(x^k) - f(x^*)\right) + 2\left(\frac{\max_i\{A_i\}L}{n} + L^2\right) V_k + \frac{1}{n^2}\sum_{i=1}^n \left(B_i\sigma_{i,k}^2 + D_{1,i}\right).
\end{aligned}
$$

Further, we define

$$
\omega_k^2 \overset{\text{def}}{=} \frac{2}{n}\sum_{i=1}^n B_i\sigma_{i,k}^2 \tag{65}
$$

and consequently, we get

$$
\begin{aligned}
\mathbf{E}\left[\omega_{k+1}^2\right] &= \frac{2}{n}\sum_{i=1}^n B_i\mathbf{E}\left[\sigma_{i,k+1}^2\right] \\
&\le (1-\rho)\omega_k^2 + \frac{2}{n}\sum_{i=1}^n B_i C_i D_{f_i}(x_i^k, x^*) + \frac{2}{n}\sum_{i=1}^n B_i D_{2,i} \\
&\overset{(63)}{\le} (1-\rho)\omega_k^2 + \frac{4}{n}\sum_{i=1}^n B_i C_i D_{f_i}(x^k, x^*) + \frac{2}{n}\sum_{i=1}^n B_i C_i L\|x_i^k - x^k\|^2 + \frac{2}{n}\sum_{i=1}^n B_i D_{2,i} \\
&\le (1-\rho)\omega_k^2 + 4\max_i\{B_i C_i\} D_f(x^k, x^*) + 2\max_i\{B_i C_i\} L V_k + \frac{2}{n}\sum_{i=1}^n B_i D_{2,i}.
\end{aligned}
$$

We will provide a bound on $\mathbf{E}\|b_i^k - \nabla f_i(x^*)\|^2$ based on the choices of b_i^k:

Case I. The choice $b_i^k = 0$ yields $\mathbf{E}\|b_i^k - \nabla f_i(x^*)\|^2 = \|\nabla f_i(x^*)\|^2$.

Case II. The choice $b_i^k = \nabla f_i(x^*)$ yields $\mathbf{E}\|b_i^k - \nabla f_i(x^*)\|^2 = 0$. Overall, for both Case I and II we have

$$\mathbf{E}\sigma_{k+1}^2 \leq (1-\rho)\mathbf{E}\sigma_k^2 + 4\max_i\{B_iC_i\}D_f(x^k,x^*) + 2\max_i\{B_iC_i\}LV_k + \frac{2}{n}\sum_{i=1}^n B_iD_{2,i}$$

as desired, where $\sigma_k = \omega_k$.

Case III. The choice $b_i^k = h_i^k - \frac{1}{n}\sum_{i=1}^n h_i^k$ yields

$$\frac{1}{n}\sum_{i=1}^n \|b_i^k - \nabla f_i(x^*)\|^2 = \frac{1}{n}\sum_{i=1}^n \left\|h_i^k - \frac{1}{n}\sum_{i=1}^n h_i^k - \nabla f_i(x^*)\right\|^2 \leq \frac{1}{n}\sum_{i=1}^n \|h_i^k - \nabla f_i(x^*)\|^2$$

where

$$\begin{aligned}\mathbf{E}_k\left[\|h_i^{k+1} - \nabla f_i(x^*)\|^2\right] &= (1-\rho_i')\|h_i^k - \nabla f_i(x^*)\|^2 + \rho_i'\mathbf{E}_k\|l_i^k - \nabla f_i(x^*)\|^2 \\ &\overset{(16)}{\leq} (1-\rho_i')\|h_i^k - \nabla f_i(x^*)\|^2 + 2\rho_i'A_i'D_{f_i}(x_i^k,x^*) + \rho_i'D_{3,i}.\end{aligned}$$

Next, set $\sigma_k^2 \overset{\text{def}}{=} \omega_k^2 + \|h_i^k - \nabla f_i(x^*)\|^2$ for this case. Consequently, we have

$$\begin{aligned}\mathbf{E}_k\sigma_{k+1}^2 \leq\ & (1-\rho)\sigma_k^2 + 4(\max_i\{B_iC_i\} + \max_i\{\rho_i'A_i'\})D_f(x^k,x^*) + 2(\max_i\{B_iC_i\} + \max_i\{\rho_i'A_i'\})LV_k \\ &+ \frac{1}{n}\sum_{i=1}^n (2B_iD_{2,i} + \rho_i'D_{3,i}),\end{aligned}$$

where $\rho = \min_i \min\{\rho_i, \rho_i'\}$.

It remains to plug everything back to (8), (9) and (10).

G Special Cases: Technical details

G.1 Local-SGD

We start with the analysis of Local-SGD (see Algorithm 1) under different assumptions of stochastic gradients and data similarity.

Algorithm 1 Local-SGD

Require: learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$, communication period $\tau \geq 1$

for $k = 0, 1, \ldots$ **do**
  for $i = 1, \ldots, n$ in parallel **do**
    Sample $g_i^k = \nabla f_{\xi_i^k}(x_i^k)$ independently from other nodes
    if $k + 1 \mod \tau = 0$ **then**
      $x_i^{k+1} = x^{k+1} = \frac{1}{n} \sum_{i=1}^n \left(x_i^k - \gamma g_i^k\right)$ ▷ averaging
    else
      $x_i^{k+1} = x_i^k - \gamma g_i^k$ ▷ local update
    end if
  end for
end for

G.1.1 Uniformly Bounded Variance

In this section we assume that f_i has a form of expectation (see (2)) and stochastic gradients $\nabla f_{\xi_i}(x)$ satisfy

$$\mathbf{E}_{\xi_i}\left[\|\nabla f_{\xi_i}(x) - \nabla f_i(x)\|^2\right] \leq D_{1,i}, \quad \forall\, x \in \mathbb{R}^d,\ \forall\, i \in [n]. \tag{66}$$

We also introduce the average variance σ^2 and the parameter of heterogeneity at the solution ζ_*^2 in the following way:

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n D_{1,i}, \quad \zeta_*^2 = \frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x^*)\|^2.$$

Lemma G.1. *Assume that functions f_i are convex and L-smooth for all $i \in [n]$. Then*

$$\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x_i^k)\|^2 \leq 6L\left(f(x^k) - f(x^*)\right) + 3L^2 V_k + 3\zeta_*^2 \tag{67}$$

and

$$\left\|\frac{1}{n} \sum_{i=1}^n \nabla f_i(x_i^k)\right\|^2 \leq 4L\left(f(x^k) - f(x^*)\right) + 2L^2 V_k. \tag{68}$$

Proof. First, to show (67) we shall have

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \|\nabla f_i(x_i^k)\|^2 &\overset{(136)}{\leq} \frac{3}{n} \sum_{i=1}^n \|\nabla f_i(x_i^k) - \nabla f_i(x^k)\|^2 + \frac{3}{n} \sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 \\
&\quad + \frac{3}{n} \sum_{i=1}^n \|\nabla f_i(x^*)\|^2 \\
&\overset{(6),(19)}{\leq} \frac{3L^2}{n} \sum_{i=1}^n \|x_i^k - x^k\|^2 + \frac{6L}{n} \sum_{i=1}^n D_{f_i}(x^k, x^*) + 3\zeta_*^2 \\
&= 6L\left(f(x^k) - f(x^*)\right) + 3L^2 V_k + 3\zeta_*^2.
\end{aligned}$$

Next, to establish (68), we have

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)\right\|^2 &= \left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(x_i^k) - \nabla f_i(x^*)\right)\right\|^2 \\ &\overset{(136)}{\leq} \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x_i^k) - \nabla f(x^k)\|^2 + \frac{2}{n}\sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f(x^*)\|^2 \\ &\overset{(6),(19)}{\leq} \frac{2L^2}{n}\sum_{i=1}^n \|x_i^k - x^k\|^2 + \frac{4L}{n}\sum_{i=1}^n D_{f_i}(x^k, x^*) \\ &= 4L\left(f(x^k) - f(x^*)\right) + 2L^2 V_k.\end{aligned}$$

□

Lemma G.2. *Let f_i be convex and L-smooth for all $i \in [n]$. Then for all $k \geq 0$*

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k\|^2 \mid x^k\right] \leq 6L\left(f(x^k) - f(x^*)\right) + 3L^2 V_k + \sigma^2 + 3\zeta_*^2, \tag{69}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] \leq \sigma^2, \tag{70}$$

$$\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k\right\|^2 \mid x^k\right] \leq 4L\left(f(x^k) - f(x^*)\right) + 2L^2 V_k + \frac{\sigma^2}{n}, \tag{71}$$

where $\mathbf{E}[\cdot \mid x^k] \overset{def}{=} \mathbf{E}[\cdot \mid x_1^k, \ldots, x_n^k]$.

Proof. First of all, we notice that $\bar{g}_i^k = \mathbf{E}\left[g_i^k \mid x^k\right] = \nabla f_i(x_i^k)$. Using this we get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x_i^k\right] &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i^k}\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x_i^k)\|^2 \overset{(66)}{\leq} \frac{1}{n}\sum_{i=1}^n D_{1,i}, \\ \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k\|^2 \mid x_i^k\right] &\overset{(139)}{=} \frac{1}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i^k}\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x_i^k)\|^2 + \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_i^k)\|^2 \\ &\overset{(66),(67)}{\leq} 6L\left(f(x^k) - f(x^*)\right) + 3L^2 V_k + \frac{1}{n}\sum_{i=1}^n \left(D_{1,i} + 3\|\nabla f_i(x^*)\|^2\right).\end{aligned}$$

Finally, using independence of $g_1^k, g_2^k, \ldots, g_n^k$ we obtain

$$\begin{aligned}\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k\right\|^2 \mid x^k\right] &\overset{(139)}{\leq} \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(g_i^k - \nabla f_i(x_i^k)\right)\right\|^2 \mid x^k\right] + \left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)\right\|^2 \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \nabla f_i(x_i^k)\|^2 \mid x_i^k\right] + \left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)\right\|^2 \\ &\overset{(66),(68)}{\leq} 4L\left(f(x^k) - f(x^*)\right) + 2L^2 V_k + \frac{1}{n^2}\sum_{i=1}^n D_{1,i}.\end{aligned}$$

□

Heterogeneous Data

Applying Corollary E.1 and Lemmas G.1 and G.2 we get the following result.

Theorem G.1. *Assume that $f_i(x)$ is μ-strongly convex and L-smooth for every $i \in [n]$. Then* `Local-SGD` *satisfies Assumption E.1 with*

$$\widetilde{A} = 3L, \quad \widehat{A} = 0, \quad \widetilde{B} = \widehat{B} = 0, \quad \widetilde{F} = 3L^2, \quad \widehat{F} = 0, \quad \widetilde{D}_1 = 3\zeta_*^2, \quad \widehat{D} = \sigma^2,$$
$$A' = 2L, \quad B' = 0, \quad F' = 2L^2, \quad D_1' = \frac{\sigma^2}{n}, \quad \sigma_k^2 \equiv 0, \quad \rho = 1, \quad C = 0, \quad G = 0, \quad D_2 = 0,$$
$$H = 0, \quad D_3 = 2e(\tau-1)\left(3(\tau-1)\zeta_*^2 + \sigma^2\right)$$

with γ satisfying

$$\gamma \quad \leq \quad \min\left\{\frac{1}{4L}, \frac{1}{4\sqrt{6e}(\tau-1)L}\right\}.$$

and for all $K \geq 0$

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \frac{2\|x^0 - x^*\|^2}{\gamma W_K} + 2\gamma\left({}^{\sigma^2}/_n + 4eL(\tau-1)\gamma\left(\sigma^2 + 3(\tau-1)\zeta_*^2\right)\right).$$

In particular, if $\mu > 0$ then

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad (1-\gamma\mu)^K \frac{2\|x^0 - x^*\|^2}{\gamma} + 2\gamma\left({}^{\sigma^2}/_n + 4eL(\tau-1)\gamma\left(\sigma^2 + 3(\tau-1)\zeta_*^2\right)\right) \tag{72}$$

and when $\mu = 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \frac{2\|x^0 - x^*\|^2}{\gamma K} + 2\gamma\left({}^{\sigma^2}/_n + 4eL(\tau-1)\gamma\left(\sigma^2 + 3(\tau-1)\zeta_*^2\right)\right). \tag{73}$$

The theorem above together with Lemma I.2 implies the following result.

Corollary G.1. *Let assumptions of Theorem G.1 hold with $\mu > 0$. Then for*

$$\gamma = \min\left\{\frac{1}{4L}, \frac{1}{4\sqrt{6e}(\tau-1)L}, \frac{\ln\left(\max\left\{2, \min\left\{{}^{\|x^0-x^*\|^2 n\mu^2 K^2}/_{\sigma^2}, {}^{\|x^0-x^*\|^2\mu^3K^3}/_{4eL(\tau-1)\left(\sigma^2+3(\tau-1)\zeta_*^2\right)}\right\}\right\}\right)}{\mu K}\right\}$$

for all K such that

$$\textit{either} \quad \frac{\ln\left(\max\left\{2, \min\left\{{}^{\|x^0-x^*\|^2 n\mu^2 K^2}/_{\sigma^2}, {}^{\|x^0-x^*\|^2\mu^3K^3}/_{4eL(\tau-1)\left(\sigma^2+3(\tau-1)\zeta_*^2\right)}\right\}\right\}\right)}{K} \leq 1$$
$$\textit{or} \quad \min\left\{\frac{1}{4L}, \frac{1}{4\sqrt{6e}(\tau-1)L}\right\} \leq \frac{\ln\left(\max\left\{2, \min\left\{{}^{\|x^0-x^*\|^2 n\mu^2 K^2}/_{\sigma^2}, {}^{\|x^0-x^*\|^2\mu^3K^3}/_{4eL(\tau-1)\left(\sigma^2+3(\tau-1)\zeta_*^2\right)}\right\}\right\}\right)}{\mu K}$$

we have that

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] = \widetilde{\mathcal{O}}\left(\tau L\|x^0 - x^*\|^2 \exp\left(-\frac{\mu}{\tau L}K\right) + \frac{\sigma^2}{n\mu K} + \frac{L(\tau-1)\left(\sigma^2 + (\tau-1)\zeta_*^2\right)}{\mu^2 K^2}\right). \tag{74}$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ in this case `Local-SGD` *requires*

$$\widetilde{\mathcal{O}}\left(\frac{\tau L}{\mu} + \frac{\sigma^2}{n\mu\varepsilon} + \sqrt{\frac{L(\tau-1)\left(\sigma^2 + (\tau-1)\zeta_*^2\right)}{\mu^2\varepsilon}}\right)$$

iterations/oracle calls per node and τ times less communication rounds.

Now we consider some special cases. First of all, if $D_{1,i} = 0$ for all $i \in [n]$, i.e. $g_i^k = \nabla f_i(x_i^k)$ almost surely, then our result implies that for `Local-SGD` it is enough to perform

$$\widetilde{\mathcal{O}}\left(\frac{\tau L}{\mu} + \sqrt{\frac{L(\tau-1)^2\zeta_*^2}{\mu^2\varepsilon}}\right)$$

iterations in order to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$. It is clear that for this scenario the optimal choice for τ is $\tau = 1$ which recovers[19] the rate of Gradient Descent.

Secondly, if $\tau = 1$ then we recover the rate of parallel SGD:

$$\widetilde{\mathcal{O}}\left(\frac{L}{\mu} + \frac{\sigma^2}{n\mu\varepsilon}\right) \qquad \text{communication rounds/oracle calls per node}$$

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

Finally, our result gives a negative answer to the following question: is Local-SGD always worse then Parallel Minibatch SGD (PMSGD) for heterogeneous data? To achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ Local-SGD requires

$$\widetilde{\mathcal{O}}\left(\frac{\tau L}{\mu} + \frac{\sigma^2}{n\mu\varepsilon} + \sqrt{\frac{L(\tau-1)\left(\sigma^2 + (\tau-1)\zeta_*^2\right)}{\mu^2\varepsilon}}\right) \quad \text{oracle calls per node.}$$

It means that if $\frac{\sigma^2}{n\sqrt{L(\tau-1)(\sigma^2+(\tau-1)\zeta_*^2)\varepsilon}} \geq 1$ for given $\tau > 1$ and ε and σ^2 are such that the first term in the complexity bound is dominated by other terms, then the second term corresponding to the complexity of PMSGD dominates the third term. Informally speaking, if the variance is large or ε is small then Local-SGD with $\tau > 1$ has the same complexity bounds as PMSGD.

Combining Theorem G.1 and Lemma I.3 we derive the following result for the convergence of Local-SGD in the case when $\mu = 0$.

Corollary G.2. *Let assumptions of Theorem G.1 hold with $\mu = 0$. Then for*

$$\gamma = \min\left\{\frac{1}{4L}, \frac{1}{4\sqrt{6e}(\tau-1)L}, \sqrt{\frac{nR_0^2}{\sigma^2 K}}, \sqrt[3]{\frac{R_0^2}{4eL(\tau-1)\left(\sigma^2 + (\tau-1)\zeta_*^2\right)K}}\right\},$$

where $R_0 = \|x^0 - x^*\|$, we have that

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] = \mathcal{O}\left(\frac{\tau L R_0^2}{K} + \sqrt{\frac{R_0^2\sigma^2}{nK}} + \frac{\sqrt[3]{LR_0^4(\tau-1)\left(\sigma^2 + (\tau-1)\zeta_*^2\right)}}{K^{2/3}}\right). \tag{75}$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ in this case Local-SGD requires

$$\mathcal{O}\left(\frac{\tau L R_0^2}{\varepsilon} + \frac{R_0^2\sigma^2}{n\varepsilon^2} + \frac{R_0^2\sqrt{L(\tau-1)\left(\sigma^2 + (\tau-1)\zeta_*^2\right)}}{\varepsilon^{3/2}}\right)$$

iterations/oracle calls per node and τ times less communication rounds.

Homogeneous Data

In this case we modify the approach a little bit and apply the following result.

Lemma G.3 (Lemma 1 from (Khaled et al., 2020)). *Under the homogeneous data assumption for Local-SGD we have*

$$\mathbf{E}\left[V_k\right] \leq (\tau-1)\gamma^2\sigma^2 \tag{76}$$

for all $k \geq 0$.

Using this we derive the following inequality for the weighted sum of V_k:

$$2L\sum_{k=0}^{K} w_k \mathbf{E}[V_k] \leq 2L(\tau-1)\gamma^2\sigma^2\sum_{k=0}^{K} w_k = 2L(\tau-1)\gamma^2\sigma^2 W_K.$$

Together with Lemmas G.1 and G.2 and Theorem 2.1 it gives the following result.

[19] We notice that for this particular case our analysis doesn't give extra logarithmical factors if we apply (72) instead of (74).

Theorem G.2. *Assume that $f(x)$ is μ-strongly convex and L-smooth and $f_1 = \ldots = f_n = f$. Then* `Local-SGD` *satisfies Assumption 2.3 with*

$$A = 3L, \quad B = 0, \quad F = 3L^2, \quad D_1 = \sigma^2, \quad A' = 2L, \quad B' = 0, \quad F' = 2L^2, \quad D_1' = \frac{\sigma^2}{n},$$
$$\sigma_k^2 \equiv 0, \quad \rho = 1, \quad C = 0, \quad G = 0, \quad D_2 = 0, \quad H = 0, \quad D_3 = (\tau - 1)\sigma^2$$

with γ satisfying

$$\gamma \quad \leq \quad \frac{1}{4L}.$$

and for all $K \geq 0$

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \frac{2\|x^0 - x^*\|^2}{\gamma W_K} + 2\gamma\left(\sigma^2/n + 2L(\tau - 1)\gamma\sigma^2\right).$$

In particular, if $\mu > 0$ then

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad (1 - \gamma\mu)^K \frac{2\|x^0 - x^*\|^2}{\gamma} + 2\gamma\left(\sigma^2/n + 2L(\tau - 1)\gamma\sigma^2\right) \tag{77}$$

and when $\mu = 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \frac{2\|x^0 - x^*\|^2}{\gamma K} + 2\gamma\left(\sigma^2/n + 2L(\tau - 1)\gamma\sigma^2\right). \tag{78}$$

The theorem above together with Lemma I.2 implies the following result.

Corollary G.3. *Let assumptions of Theorem G.2 hold with $\mu > 0$. Then for*

$$\gamma = \min\left\{\frac{1}{4L}, \frac{\ln\left(\max\left\{2, \min\left\{\|x^0 - x^*\|^2 n\mu^2 K^2/\sigma^2, \|x^0 - x^*\|^2\mu^3 K^3/2L(\tau-1)\sigma^2\right\}\right\}\right)}{\mu K}\right\}$$

for all K such that

$$\begin{aligned} &\textit{either} \quad && \frac{\ln\left(\max\left\{2, \min\left\{\|x^0 - x^*\|^2 n\mu^2 K^2/\sigma^2, \|x^0 - x^*\|^2\mu^3 K^3/2L(\tau-1)\sigma^2\right\}\right\}\right)}{K} \leq 1 \\ &\textit{or} \quad && \frac{1}{4L} \leq \frac{\ln\left(\max\left\{2, \min\left\{\|x^0 - x^*\|^2 n\mu^2 K^2/\sigma^2, \|x^0 - x^*\|^2\mu^3 K^3/2L(\tau-1)\sigma^2\right\}\right\}\right)}{\mu K} \end{aligned}$$

we have that

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] = \widetilde{\mathcal{O}}\left(L\|x^0 - x^*\|^2 \exp\left(-\frac{\mu}{L}K\right) + \frac{\sigma^2}{n\mu K} + \frac{L(\tau - 1)\sigma^2}{\mu^2 K^2}\right). \tag{79}$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ in this case `Local-SGD` *requires*

$$\widetilde{\mathcal{O}}\left(\frac{L}{\mu}\ln\left(\frac{L\|x^0 - x^*\|^2}{\varepsilon}\right) + \frac{\sigma^2}{n\mu\varepsilon} + \sqrt{\frac{L(\tau - 1)\sigma^2}{\mu^2\varepsilon}}\right)$$

iterations/oracle calls per node and τ times less communication rounds.

It means that if $\frac{\sigma^2}{n^2 L\varepsilon} \geq 1$, $\tau \leq 1 + \frac{\sigma^2}{n^2 L\varepsilon}$ and ε and σ^2 are such that the first term in the complexity bound is dominated by other terms, then the second term corresponding to the complexity of `PMSGD` dominates the third term. Informally speaking, if the variance is large or ε is small then `Local-SGD` with $\tau > 1$ has the same complexity bounds as `PMSGD`.

Combining Theorem G.2 and Lemma I.3 we derive the following result for the convergence of `Local-SGD` in the case when $\mu = 0$.

Corollary G.4. *Let assumptions of Theorem G.2 hold with $\mu = 0$. Then for*

$$\gamma = \min\left\{\frac{1}{4L}, \sqrt{\frac{nR_0^2}{\sigma^2 K}}, \sqrt[3]{\frac{R_0^2}{2L(\tau-1)\sigma^2 K}}\right\},$$

where $R_0 = \|x^0 - x^*\|$, we have that

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] = \mathcal{O}\left(\frac{LR_0^2}{K} + \sqrt{\frac{R_0^2\sigma^2}{nK}} + \frac{\sqrt[3]{LR_0^4(\tau-1)\sigma^2}}{K^{2/3}}\right). \tag{80}$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ in this case `Local-SGD` *requires*

$$\mathcal{O}\left(\frac{LR_0^2}{\varepsilon} + \frac{R_0^2\sigma^2}{n\varepsilon^2} + \frac{R_0^2\sqrt{L(\tau-1)\sigma^2}}{\varepsilon^{3/2}}\right)$$

iterations/oracle calls per node and τ times less communication rounds.

ζ-Heterogeneous Data

In this setup we also use an external result to bound $\mathbf{E}[V_k]$.

Lemma G.4 (Lemma 8 from (Woodworth et al., 2020a)). *If $f_1, f_2, \ldots, f_n$ are ζ-heterogeneous then for* `Local-SGD` *we have*

$$\mathbf{E}\left[V_k\right] \leq 3\tau\gamma^2\sigma^2 + 6\tau^2\gamma^2\zeta^2 \tag{81}$$

for all $k \geq 0$.

Using this we derive the following inequality for the weighted sum of V_k:

$$2L\sum_{k=0}^{K} w_k \mathbf{E}[V_k] \leq 6\tau L\gamma^2\left(\sigma^2 + 2\tau\zeta^2\right)\sum_{k=0}^{K} w_k = 6\tau L\gamma^2\left(\sigma^2 + 2\tau\zeta^2\right)W_K.$$

Together with Lemmas G.1 and G.2 and Theorem 2.1 it gives the following result.

Theorem G.3. *Assume that $f_1, \ldots, f_n$ are ζ-heterogeneous, μ-strongly convex and L-smooth functions. Then* `Local-SGD` *satisfies Assumption 2.3 with*

$$A = 3L, \quad B = 0, \quad F = 3L^2, \quad D_1 = \sigma^2 + 3\zeta_*^2, \quad A' = 2L, \quad B' = 0, \quad F' = 2L^2, \quad D_1' = \frac{\sigma^2}{n},$$
$$\sigma_k^2 \equiv 0, \quad \rho = 1, \quad C = 0, \quad G = 0, \quad D_2 = 0, \quad H = 0, \quad D_3 = 3\tau\left(\sigma^2 + 2\tau\zeta^2\right)$$

with γ satisfying

$$\gamma \quad \leq \quad \frac{1}{4L}.$$

and for all $K \geq 0$

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \frac{2\|x^0 - x^*\|^2}{\gamma W_K} + 2\gamma\left(\sigma^2/n + 6L\tau\gamma\left(\sigma^2 + 2\tau\zeta^2\right)\right).$$

In particular, if $\mu > 0$ then

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad (1-\gamma\mu)^K\frac{2\|x^0 - x^*\|^2}{\gamma} + 2\gamma\left(\sigma^2/n + 6L\tau\gamma\left(\sigma^2 + 2\tau\zeta^2\right)\right) \tag{82}$$

and when $\mu = 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \frac{2\|x^0 - x^*\|^2}{\gamma K} + 2\gamma\left(\sigma^2/n + 6L\tau\gamma\left(\sigma^2 + 2\tau\zeta^2\right)\right). \tag{83}$$

The theorem above together with Lemma I.2 implies the following result.

Corollary G.5. *Let assumptions of Theorem G.3 hold with $\mu > 0$. Then for*

$$\gamma = \min\left\{\frac{1}{4L}, \frac{\ln\left(\max\left\{2, \min\left\{\|x^0-x^*\|^2 n\mu^2K^2/\sigma^2, \|x^0-x^*\|^2\mu^3K^3/6L\tau(\sigma^2+2\tau\zeta^2)\right\}\right\}\right)}{\mu K}\right\}$$

for all K such that

$$\text{either} \quad \frac{\ln\left(\max\left\{2, \min\left\{\|x^0-x^*\|^2 n\mu^2K^2/\sigma^2, \|x^0-x^*\|^2\mu^3K^3/6L\tau(\sigma^2+2\tau\zeta^2)\right\}\right\}\right)}{K} \leq 1$$

$$\text{or} \quad \frac{1}{4L} \leq \frac{\ln\left(\max\left\{2, \min\left\{\|x^0-x^*\|^2 n\mu^2K^2/\sigma^2, \|x^0-x^*\|^2\mu^3K^3/6L\tau(\sigma^2+2\tau\zeta^2)\right\}\right\}\right)}{\mu K}$$

we have that

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] = \widetilde{\mathcal{O}}\left(L\|x^0-x^*\|^2\exp\left(-\frac{\mu}{L}K\right) + \frac{\sigma^2}{n\mu K} + \frac{L\tau(\sigma^2+\tau\zeta^2)}{\mu^2K^2}\right). \tag{84}$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ in this case `Local-SGD` *requires*

$$\widetilde{\mathcal{O}}\left(\frac{L}{\mu}\ln\left(\frac{L\|x^0-x^*\|^2}{\varepsilon}\right) + \frac{\sigma^2}{n\mu\varepsilon} + \sqrt{\frac{L\tau(\sigma^2+\tau\zeta^2)}{\mu^2\varepsilon}}\right)$$

iterations/oracle calls per node and τ times less communication rounds.

Combining Theorem G.3 and Lemma I.3 we derive the following result for the convergence of `Local-SGD` in the case when $\mu = 0$.

Corollary G.6. *Let assumptions of Theorem G.3 hold with $\mu = 0$. Then for*

$$\gamma = \min\left\{\frac{1}{4L}, \sqrt{\frac{nR_0^2}{\sigma^2K}}, \sqrt[3]{\frac{R_0^2}{6L\tau(\sigma^2+2\tau\zeta^2)K}}\right\},$$

where $R_0 = \|x^0-x^*\|$, we have that

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] = \mathcal{O}\left(\frac{LR_0^2}{K} + \sqrt{\frac{R_0^2\sigma^2}{nK}} + \frac{\sqrt[3]{LR_0^4\tau(\sigma^2+\tau\zeta^2)}}{K^{2/3}}\right). \tag{85}$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ in this case `Local-SGD` *requires*

$$\mathcal{O}\left(\frac{LR_0^2}{\varepsilon} + \frac{R_0^2\sigma^2}{n\varepsilon^2} + \frac{R_0^2\sqrt{L\tau(\sigma^2+\tau\zeta^2)}}{\varepsilon^{3/2}}\right)$$

iterations/oracle calls per node and τ times less communication rounds.

G.1.2 Expected Smoothness and Arbitrary Sampling

In this section we continue our consideration of `Local-SGD` but now we make another assumption on stochastic gradients $\nabla f_{\xi_i}(x)$.

Assumption G.1 (Expected Smoothness)**.** *We assume that for all $i \in [n]$ stochastic gradients $\nabla f_{\xi_i}(x)$ are unbiased estimators of $\nabla f_i(x)$ and there exists such constant $\mathcal{L} > 0$ that $\forall x, y \in \mathbb{R}^d$*

$$\mathbf{E}_{\xi_i\sim\mathcal{D}_i}\left[\|\nabla f_{\xi_i}(x) - \nabla f_{\xi_i}(x^*)\|^2\right] \leq 2\mathcal{L}D_{f_i}(x, x^*) \tag{86}$$

where $D_{f_i}(x,y) \overset{def}{=} f_i(x) - f_i(y) - \langle\nabla f_i(y), x-y\rangle$.

In particular, let us consider the following special case. Assume that $f_i(x)$ has a form of finite sum (see (3)) and consider the following stochastic reformulation:

$$f_i(x) = \mathbf{E}_{\xi_i}\left[f_{\xi_i}(x)\right], \quad f_{\xi_i}(x) = \frac{1}{m}\sum_{j=1}^{m}\xi_{i,j}f_{i,j}(x), \tag{87}$$

where $\mathbf{E}[\xi_{i,j}] = 1$ and $\mathbf{E}[\xi_{i,j}^2] < \infty$. In this case, $\mathbf{E}_{\xi_i}[\nabla f_{\xi_i}] = \nabla f_i(x)$. If each $f_{i,j}(x)$ is $L_{i,j}$-smooth then there exists such $\mathcal{L} \leq \max_{j\in[m]} L_{i,j}$ that Assumption G.1 holds. Clearly, $\mathcal{L}$ depends on the sampling strategy and in some cases one can make $\mathcal{L}$ much smaller than $\max_{j\in[m]} L_{i,j}$ via good choice of this strategy. Our analysis works for an arbitrary sampling strategy that satisfies Assumption G.1.

Lemma G.5. *Let f_i be convex and L-smooth for all $i \in [n]$. Then for all $k \geq 0$*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\|g_i^k\|^2 \mid x^k\right] \leq 8\mathcal{L}\left(f(x^k) - f(x^*)\right) + 4\mathcal{L}LV_k + 2\sigma_*^2 + 2\zeta_*^2, \tag{88}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] \leq 8\mathcal{L}\left(f(x^k) - f(x^*)\right) + 4\mathcal{L}LV_k + 2\sigma_*^2, \tag{89}$$

$$\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}g_i^k\right\|^2 \mid x^k\right] \leq 4\left(2\mathcal{L}/n + L\right)\left(f(x^k) - f(x^*)\right) + 2L\left(2\mathcal{L}/n + L\right)V_k + \frac{2\sigma_*^2}{n}, \tag{90}$$

where $\sigma_*^2 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2$, $\zeta_*^2 = \frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x^*)\|^2$ and $\mathbf{E}[\cdot \mid x^k] \overset{def}{=} \mathbf{E}[\cdot \mid x_1^k, \ldots, x_n^k]$.

Proof. First of all, we notice that $\bar{g}_i^k = \mathbf{E}\left[g_i^k \mid x^k\right] = \nabla f_i(x_i^k)$. Using this we get

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\|g_i^k\|^2 \mid x^k\right] &\overset{(136)}{\leq} \frac{2}{n}\sum_{i=1}^{n}\mathbf{E}_{\xi_i^k}\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_{\xi_i^k}(x^*)\|^2 + \frac{2}{n}\sum_{i=1}^{n}\mathbf{E}_{\xi_i^k}\|\nabla f_{\xi_i^k}(x^*)\|^2 \\
&\overset{(86),(139)}{\leq} \frac{4\mathcal{L}}{n}\sum_{i=1}^{n}D_{f_i}(x_i^k, x^*) + \frac{2}{n}\sum_{i=1}^{n}\mathbf{E}_{\xi_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right] + \frac{2}{n}\sum_{i=1}^{n}\|\nabla f_i(x^*)\|^2 \\
&\overset{(63)}{\leq} 8\mathcal{L}\left(f(x^k) - f(x^*)\right) + 4\mathcal{L}LV_k + 2\sigma_*^2 + 2\zeta_*^2
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x^k\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_{\xi_i^k}\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x_i^k)\|^2 \\
&\overset{(139)}{\leq} \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_{\xi_i^k}\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x^*)\|^2 \\
&\overset{(136)}{\leq} \frac{2}{n}\sum_{i=1}^{n}\mathbf{E}_{\xi_i^k}\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_{\xi_i^k}(x^*)\|^2 + \frac{2}{n}\sum_{i=1}^{n}\mathbf{E}_{\xi_i^k}\|\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\|^2 \\
&\overset{(86)}{\leq} \frac{4\mathcal{L}}{n}\sum_{i=1}^{n}D_{f_i}(x_i^k, x^*) + 2\sigma_*^2 \\
&\overset{(63)}{\leq} 8\mathcal{L}\left(f(x^k) - f(x^*)\right) + 4\mathcal{L}LV_k + 2\sigma_*^2.
\end{aligned} \tag{91}$$

Finally, using independence of $\xi_1^k, \xi_2^k, \ldots, \xi_n^k$ we obtain

$$\begin{aligned}
\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}g_i^k\right\|^2 \mid x^k\right] &\overset{(139)}{=} \mathbf{E}_{\xi_i^k}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}(\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x_i^k))\right\|^2\right] + \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^k)\right\|^2 \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{E}_{\xi_i^k}\left[\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x_i^k)\|^2\right] + \left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^k)\right\|^2 \\
&\overset{(91),(68)}{\leq} 4\left(2\mathcal{L}/n + L\right)\left(f(x^k) - f(x^*)\right) + 2L\left(2\mathcal{L}/n + L\right)V_k + \frac{2\sigma_*^2}{n}.
\end{aligned}$$

□

Heterogeneous Data

Applying Corollary E.1 and Lemmas G.1 and G.5 we get the following result.

Theorem G.4. *Assume that $f_i(x)$ is μ-strongly convex and L-smooth for $i \in [n]$. Let Assumption G.1 holds. Then* `Local-SGD` *satisfies Assumption E.1 with*

$$\widetilde{A} = 3L, \quad \widehat{A} = 4\mathcal{L}, \quad \widetilde{B} = \widehat{B} = 0, \quad \widetilde{F} = 3L^2, \quad \widehat{F} = 4\mathcal{L}L, \quad \widetilde{D}_1 = 3\zeta_*^2, \quad \widehat{D}_1 = 2\sigma_*^2$$
$$A' = \frac{4\mathcal{L}}{n} + 2L, \quad B' = 0, \quad F' = \frac{4\mathcal{L}L}{n} + 2L^2, \quad D_1' = \frac{2\sigma_*^2}{n},$$
$$\sigma_k^2 \equiv 0, \quad \rho = 1, \quad C = 0, \quad G = 0, \quad D_2 = 0,$$
$$H = 0, \quad D_3 = 2e(\tau-1)\left(2\sigma_*^2 + 3(\tau-1)\zeta_*^2\right)$$

with γ satisfying

$$\gamma \quad \leq \quad \min\left\{\frac{1}{8\mathcal{L}/n + 4L}, \frac{1}{4\sqrt{2eL(\tau-1)\left(3L(\tau-1)+4\mathcal{L}\right)}}\right\}.$$

and for all $K \geq 0$

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \frac{2\|x^0 - x^*\|^2}{\gamma W_K} + 2\gamma\left(2\sigma_*^2/n + 4eL(\tau-1)\gamma\left(2\sigma_*^2 + 3(\tau-1)\zeta_*^2\right)\right).$$

In particular, if $\mu > 0$ then

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad (1-\gamma\mu)^K \frac{2\|x^0 - x^*\|^2}{\gamma} + 2\gamma\left(2\sigma_*^2/n + 4eL(\tau-1)\gamma\left(2\sigma_*^2 + 3(\tau-1)\zeta_*^2\right)\right) \tag{92}$$

and when $\mu = 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \frac{2\|x^0 - x^*\|^2}{\gamma K} + 2\gamma\left(2\sigma_*^2/n + 4eL(\tau-1)\gamma\left(2\sigma_*^2 + 3(\tau-1)\zeta_*^2\right)\right). \tag{93}$$

The theorem above together with Lemma I.2 implies the following result.

Corollary G.7. *Let assumptions of Theorem G.4 hold with $\mu > 0$. Then for*

$$\gamma_0 \quad = \quad \min\left\{\frac{1}{8\mathcal{L}/n + 4L}, \frac{1}{4\sqrt{2eL(\tau-1)\left(3L(\tau-1)+4\mathcal{L}\right)}}\right\},$$
$$\gamma \quad = \quad \min\left\{\gamma_0, \frac{\ln\left(\max\left\{2, \min\left\{n\|x^0-x^*\|^2\mu^2K^2/2\sigma_*^2, \|x^0-x^*\|^2\mu^3K^3/4eL(\tau-1)\gamma\left(2\sigma_*^2+3(\tau-1)\zeta_*^2\right)\right\}\right\}\right)}{\mu K}\right\}$$

for all K such that

$$\textit{either} \qquad \frac{\ln\left(\max\left\{2, \min\left\{n\|x^0-x^*\|^2\mu^2K^2/2\sigma_*^2, \|x^0-x^*\|^2\mu^3K^3/4eL(\tau-1)\gamma\left(2\sigma_*^2+3(\tau-1)\zeta_*^2\right)\right\}\right\}\right)}{K} \leq 1$$
$$\textit{or} \qquad \gamma_0 \leq \frac{\ln\left(\max\left\{2, \min\left\{n\|x^0-x^*\|^2\mu^2K^2/2\sigma_*^2, \|x^0-x^*\|^2\mu^3K^3/4eL(\tau-1)\gamma\left(2\sigma_*^2+3(\tau-1)\zeta_*^2\right)\right\}\right\}\right)}{\mu K}$$

we have that $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right]$ is of the order

$$\widetilde{\mathcal{O}}\left(\left(L\tau + \mathcal{L}/n + \sqrt{(\tau-1)\mathcal{L}L}\right)R_0^2 \exp\left(-\frac{\mu}{L\tau + \mathcal{L}/n + \sqrt{(\tau-1)\mathcal{L}L}}K\right) + \frac{\sigma_*^2}{n\mu K} + \frac{L(\tau-1)\left(\sigma_*^2 + (\tau-1)\zeta_*^2\right)}{\mu^2K^2}\right),$$

where $R_0 = \|x^0 - x^*\|$. That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ in this case `Local-SGD` *requires*

$$\widetilde{\mathcal{O}}\left(\frac{L\tau}{\mu} + \frac{\mathcal{L}}{n\mu} + \frac{\sqrt{(\tau-1)\mathcal{L}L}}{\mu} + \frac{\sigma_*^2}{n\mu\varepsilon} + \sqrt{\frac{L(\tau-1)\left(\sigma_*^2 + (\tau-1)\zeta_*^2\right)}{\mu^2\varepsilon}}\right)$$

iterations/oracle calls per node and τ times less communication rounds.

Combining Theorem G.4 and Lemma I.3 we derive the following result for the convergence of `Local-SGD` in the case when $\mu = 0$.

Corollary G.8. *Let assumptions of Theorem G.4 hold with $\mu = 0$. Then for*

$$\begin{aligned}\gamma_0 &= \min\left\{\frac{1}{8\mathcal{L}/n + 4L}, \frac{1}{4\sqrt{2eL(\tau-1)\left(3L(\tau-1)+4\mathcal{L}\right)}}\right\},\\ \gamma &= \min\left\{\gamma_0, \sqrt{\frac{nR_0^2}{2\sigma_*^2 K}}, \sqrt[3]{\frac{R_0^2}{4eL(\tau-1)\left(2\sigma_*^2+3(\tau-1)\zeta_*^2\right)K}}\right\},\end{aligned}$$

where $R_0 = \|x^0 - x^*\|$, we have that

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] = \mathcal{O}\left(\frac{\left(L\tau + \mathcal{L}/n + \sqrt{(\tau-1)\mathcal{L}L}\right)R_0^2}{K} + \sqrt{\frac{R_0^2\sigma_*^2}{nK}} + \frac{\sqrt[3]{LR_0^4(\tau-1)\left(\sigma_*^2+(\tau-1)\zeta_*^2\right)}}{K^{2/3}}\right). \quad (94)$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ in this case `Local-SGD` *requires*

$$\mathcal{O}\left(\frac{\left(L\tau + \mathcal{L}/n + \sqrt{(\tau-1)\mathcal{L}L}\right)R_0^2}{\varepsilon} + \frac{R_0^2\sigma_*^2}{n\varepsilon^2} + \frac{R_0^2\sqrt{L(\tau-1)\left(\sigma_*^2+(\tau-1)\zeta_*^2\right)}}{\varepsilon^{3/2}}\right)$$

iterations/oracle calls per node and τ times less communication rounds.

ζ-Heterogeneous Data

Applying Corollary E.2 and Lemma G.5 we get the following result.

Theorem G.5. *Assume that $f_i(x)$ is L-smooth for $i \in [n]$ and $f_1, \ldots, f_n$ are ζ-heterogeneous and μ-strongly convex. Let Assumption G.1 holds. Then* `Local-SGD` *satisfies Assumption 2.3 with*

$$\begin{gathered}A = 4\mathcal{L}, \quad B = 0, \quad F = 4\mathcal{L}L, \quad D_1 = 2\sigma_*^2 + 2\zeta_*^2,\\ A' = \frac{4\mathcal{L}}{n} + 2L, \quad B' = 0, \quad F' = \frac{4\mathcal{L}L}{n} + 2L^2, \quad D_1' = \frac{2\sigma_*^2}{n},\\ \sigma_k^2 \equiv 0, \quad \rho = 1, \quad C = 0, \quad G = 0, \quad D_2 = 0,\\ H = 0, \quad D_3 = 2(\tau-1)\left(2\sigma_*^2 + 2\zeta_*^2 + \frac{\zeta^2}{\gamma\mu}\right)\end{gathered}$$

with γ satisfying

$$\gamma \leq \min\left\{\frac{1}{8\mathcal{L}/n + 4L}, \frac{1}{8\sqrt{2L\mathcal{L}(\tau-1)}}\right\}.$$

and for all $K \geq 0$

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{2\|x^0 - x^*\|^2}{\gamma W_K} + 2\gamma\left(\frac{2\sigma_*^2}{n} + \frac{4L\zeta^2(\tau-1)}{\mu} + 8L(\tau-1)\gamma\left(\sigma_*^2 + \zeta_*^2\right)\right).$$

In particular, if $\mu > 0$ then

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq (1-\gamma\mu)^K\frac{2\|x^0 - x^*\|^2}{\gamma} + 2\gamma\left(\frac{2\sigma_*^2}{n} + \frac{4L\zeta^2(\tau-1)}{\mu} + 8L(\tau-1)\gamma\left(\sigma_*^2 + \zeta_*^2\right)\right) \quad (95)$$

and when $\mu = 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{2\|x^0 - x^*\|^2}{\gamma K} + 2\gamma\left(\frac{2\sigma_*^2}{n} + \frac{4L\zeta^2(\tau-1)}{\mu} + 8L(\tau-1)\gamma\left(\sigma_*^2 + \zeta_*^2\right)\right). \quad (96)$$

The theorem above together with Lemma I.2 implies the following result.

Corollary G.9. *Let assumptions of Theorem G.5 hold with $\mu > 0$. Then for*

$$\begin{aligned}\gamma_0 &= \min\left\{\frac{1}{8\mathcal{L}/n + 4L}, \frac{1}{8\sqrt{2L\mathcal{L}(\tau-1)}}\right\},\\ \gamma &= \min\left\{\gamma_0, \frac{\ln\left(\max\left\{2, \min\left\{\|x^0-x^*\|^2\mu^2K^2/\left(2\sigma_*^2/n+4L\zeta^2(\tau-1)/\mu\right), \|x^0-x^*\|^2\mu^3K^3/8L(\tau-1)\left(\sigma_*^2+\zeta_*^2\right)\right\}\right\}\right)}{\mu K}\right\}\end{aligned}$$

for all K such that

$$\begin{aligned}&\textit{either} \quad \frac{\ln\left(\max\left\{2, \min\left\{\|x^0-x^*\|^2\mu^2K^2/\left(2\sigma_*^2/n+4L\zeta^2(\tau-1)/\mu\right), \|x^0-x^*\|^2\mu^3K^3/8L(\tau-1)\left(\sigma_*^2+\zeta_*^2\right)\right\}\right\}\right)}{K} \leq 1\\ &\textit{or} \quad \gamma_0 \leq \frac{\ln\left(\max\left\{2, \min\left\{\|x^0-x^*\|^2\mu^2K^2/\left(2\sigma_*^2/n+4L\zeta^2(\tau-1)/\mu\right), \|x^0-x^*\|^2\mu^3K^3/8L(\tau-1)\left(\sigma_*^2+\zeta_*^2\right)\right\}\right\}\right)}{\mu K}\end{aligned}$$

we have that $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right]$ is of the order

$$\widetilde{\mathcal{O}}\left(\left(L + \mathcal{L}/n + \sqrt{(\tau-1)\mathcal{L}L}\right)R_0^2 \exp\left(-\frac{\mu}{L + \mathcal{L}/n + \sqrt{(\tau-1)\mathcal{L}L}}K\right) + \frac{\sigma_*^2}{n\mu K} + \frac{L\zeta^2(\tau-1)}{\mu^2 K} + \frac{L(\tau-1)\left(\sigma_*^2+\zeta_*^2\right)}{\mu^2K^2}\right),$$

where $R_0 = \|x^0 - x^*\|$. That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ in this case `Local-SGD` *requires*

$$\widetilde{\mathcal{O}}\left(\frac{L}{\mu} + \frac{\mathcal{L}}{n\mu} + \frac{\sqrt{(\tau-1)\mathcal{L}L}}{\mu} + \frac{\sigma_*^2}{n\mu\varepsilon} + \frac{L\zeta^2(\tau-1)}{\mu^2\varepsilon} + \sqrt{\frac{L(\tau-1)\left(\sigma_*^2+\zeta_*^2\right)}{\mu^2\varepsilon}}\right)$$

iterations/oracle calls per node and τ times less communication rounds.

Combining Theorem G.5 and Lemma I.3 we derive the following result for the convergence of `Local-SGD` in the case when $\mu = 0$.

Corollary G.10. *Let assumptions of Theorem G.5 hold with $\mu = 0$. Then for*

$$\begin{aligned}\gamma_0 &= \min\left\{\frac{1}{8\mathcal{L}/n + 4L}, \frac{1}{8\sqrt{2L\mathcal{L}(\tau-1)}}\right\},\\ \gamma &= \min\left\{\gamma_0, \sqrt{\frac{R_0^2}{\left(2\sigma_*^2/n + 4L\zeta^2(\tau-1)/\mu\right)K}}, \sqrt[3]{\frac{R_0^2}{8L(\tau-1)\left(\sigma_*^2+\zeta_*^2\right)K}}\right\},\end{aligned}$$

where $R_0 = \|x^0 - x^*\|$, we have that

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] = \mathcal{O}\left(\frac{\left(L + \mathcal{L}/n + \sqrt{(\tau-1)\mathcal{L}L}\right)R_0^2}{K} + \sqrt{\frac{R_0^2\left(\sigma_*^2/n + L\zeta^2(\tau-1)/\mu\right)}{K}} + \frac{\sqrt[3]{LR_0^4(\tau-1)\left(\sigma_*^2+\zeta_*^2\right)}}{K^{2/3}}\right).$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ in this case `Local-SGD` *requires*

$$\mathcal{O}\left(\frac{\left(L + \mathcal{L}/n + \sqrt{(\tau-1)\mathcal{L}L}\right)R_0^2}{\varepsilon} + \frac{\left(\sigma_*^2/n + L\zeta^2(\tau-1)/\mu\right)R_0^2}{\varepsilon^2} + \frac{R_0^2\sqrt{L(\tau-1)\left(\sigma_*^2+\zeta_*^2\right)}}{\varepsilon^{3/2}}\right)$$

iterations/oracle calls per node and τ times less communication rounds.

G.2 `Local-SVRG`

As an alternative to `Local-SGD` when the local objective is of a finite sum structure (3), we propose `L-SVRG` (Hofmann et al., 2015; Kovalev et al., 2019) stochastic gradient as a local direction instead of the plain stochastic gradient. Specifically, we consider

$$a_i^k \overset{\text{def}}{=} \nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(w_i^k) + \nabla f_i(w_i^k), \qquad b_i^k = 0,$$

where index $1 \leq j_i \leq m$ is selected uniformly at random and w_i^k is a particular iterate from the local history updated as follows:

$$w_i^{k+1} = \begin{cases} x_i^k & \text{w.p. } q \\ w_i^k & \text{w.p. } 1-q. \end{cases}$$

Next, we will assume that the local functions $f_{i,j}$ are $\max L_{ij}$-smooth.[20] Lastly, we will equip the mentioned method with the fixed local loop. The formal statement of the described instance of (4) is given as Algorithm 2.

Algorithm 2 `Local-SVRG`

Require: learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$, communication period $\tau \geq 1$

for $k = 0, 1, \ldots$ **do**
  for $i = 1, \ldots, n$ in parallel **do**
    Choose j_i uniformly at random, independently across nodes
    $g_i^k = \nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(w_i^k) + \nabla f_i(w_i^k)$
    $w_i^{k+1} = \begin{cases} x_i^k & \text{w.p. } q \\ w_i^k & \text{w.p. } 1-q \end{cases}$
    if $k+1 \mod \tau = 0$ **then**
      $x_i^{k+1} = x^{k+1} = \frac{1}{n}\sum\limits_{i=1}^n \left(x_i^k - \gamma g_i^k\right)$ ▷ averaging
    else
      $x_i^{k+1} = x_i^k - \gamma g_i^k$ ▷ local update
    end if
  end for
end for

Let us next provide the details on the convergence rate. In order to do so, let us identify the parameters of Assumption 4.1.

Proposition G.1 (see (Gorbunov et al., 2020a)). *Gradient estimator a_i^k satisfies Assumption 4.1 with parameters $A_i = 2\max L_{ij}, B_i = 2, D_{1,i} = 0, \rho_i = q, C_i = \max L_{ij} q, D_{2,i} = 0$, and $\sigma_{i,k}^2 = \frac{1}{m}\sum\limits_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2$.*

G.2.1 ζ-Heterogeneous Data

It remains to use Lemma 4.1 along with Corollary E.2 to recover all parameters of Assumption 2.3 and obtain a convergence rate of Algorithm 2 in ζ-heterogeneous case.

Theorem G.6. *Assume that $f_i(x)$ is μ-strongly convex and L-smooth for $i \in [n]$ and $f_1, \ldots, f_n$ are ζ-heterogeneous, convex and $\max L_{ij}$-smooth. Then* `Local-SVRG` *satisfies Assumption 2.3 with*

$$A = 8\max L_{ij}, \quad B = 2, \quad F = 8L\max L_{ij}, \quad D_1 = 2\zeta_*^2,$$
$$A' = \frac{4\max L_{ij}}{n} + L, \quad B' = \frac{1}{n}, \quad F' = \frac{4L\max L_{ij}}{n} + 2L^2, \quad D_1' = 0,$$
$$\sigma_k^2 = \frac{4}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2, \quad \rho = q, \quad C = 8q\max L_{ij}, \quad G = 4qL\max L_{ij}, \quad D_2 = 0,$$
$$H = \frac{8(\tau-1)(2+q)\gamma^2}{q}, \quad D_3 = 2(\tau-1)\left(2\zeta_*^2 + \frac{\zeta^2}{\gamma\mu}\right)$$

with γ satisfying

$$\gamma \quad \leq \quad \min\left\{\frac{1}{2\left(44\max L_{ij}/n + L\right)}, \frac{1}{16\sqrt{L\max L_{ij}(\tau-1)\left(1 + 4/(1-q)\right)}}\right\}.$$

[20] It is easy to see that we must have $\max L_{ij} \geq L \geq \frac{1}{m}\max L_{ij}$.

and for all $K \geq 0$

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \frac{\Phi^0}{\gamma W_K} + 8L(\tau-1)\gamma\left(\frac{\zeta^2}{\mu} + 2\gamma\zeta_*^2\right),$$

where $\Phi^0 = 2\|x^0 - x^*\|^2 + \frac{8}{3nq}\gamma^2\sigma_0^2 + \frac{32L(\tau-1)(2+q)\gamma^3}{q}\sigma_0^2$. *In particular, if* $\mu > 0$ *then*

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \left(1 - \min\left\{\gamma\mu, \frac{q}{4}\right\}\right)^K \frac{\Phi^0}{\gamma} + 8L(\tau-1)\gamma\left(\frac{\zeta^2}{\mu} + 2\gamma\zeta_*^2\right) \tag{97}$$

and when $\mu = 0$ *we have*

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \frac{\Phi^0}{\gamma K} + 8L(\tau-1)\gamma\left(\frac{\zeta^2}{\mu} + 2\gamma\zeta_*^2\right). \tag{98}$$

The theorem above together with Lemma I.2 implies the following result.

Corollary G.11. *Let assumptions of Theorem G.6 hold with* $\mu > 0$. *Then for*

$$\begin{aligned}
\gamma_0 &= \min\left\{\frac{1}{2\left(44\max L_{ij}/n + L\right)}, \frac{1}{16\sqrt{L\max L_{ij}(\tau-1)\left(1 + 4/(1-q)\right)}}\right\}, \quad q = \frac{1}{m}, \quad m > 1,\\
\widetilde{\Phi}^0 &= 2\|x^0 - x^*\|^2 + \frac{8}{3nq}\gamma_0^2\sigma_0^2 + \frac{32L(\tau-1)(2+q)\gamma_0^3}{q}\sigma_0^2,\\
\gamma &= \min\left\{\gamma_0, \frac{\ln\left(\max\left\{2, \min\left\{\widetilde{\Phi}^0\mu^3K^2/8L\zeta^2(\tau-1), \widetilde{\Phi}^0\mu^3K^3/16L(\tau-1)\zeta_*^2\right\}\right\}\right)}{\mu K}\right\},
\end{aligned}$$

for all K *such that*

$$\begin{aligned}
\textit{either} \quad & \frac{\ln\left(\max\left\{2, \min\left\{\widetilde{\Phi}^0\mu^3K^2/8L\zeta^2(\tau-1), \widetilde{\Phi}^0\mu^3K^3/16L(\tau-1)\zeta_*^2\right\}\right\}\right)}{K} \leq \frac{1}{m}\\
\textit{or} \quad & \gamma_0 \leq \frac{\ln\left(\max\left\{2, \min\left\{\widetilde{\Phi}^0\mu^3K^2/8L\zeta^2(\tau-1), \widetilde{\Phi}^0\mu^3K^3/16L(\tau-1)\zeta_*^2\right\}\right\}\right)}{\mu K}
\end{aligned}$$

we have that $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right]$ *is of the order*

$$\widetilde{\mathcal{O}}\left(\frac{\widetilde{\Phi}^0}{\gamma_0}\exp\left(-\min\left\{m^{-1}, \gamma_0\mu\right\}K\right) + \frac{\zeta^2L(\tau-1)}{\mu^2K} + \frac{L(\tau-1)\zeta_*^2}{\mu^2K^2}\right).$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ *in this case* `Local-SVRG` *requires*

$$\widetilde{\mathcal{O}}\left(m + \frac{L}{\mu} + \frac{\max L_{ij}}{n\mu} + \frac{\sqrt{(\tau-1)L\max L_{ij}}}{\mu} + \frac{L\zeta^2(\tau-1)}{\mu^2\varepsilon} + \sqrt{\frac{L(\tau-1)\zeta_*^2}{\mu^2\varepsilon}}\right)$$

iterations/oracle calls per node and τ *times less communication rounds.*

Combining Theorem G.6 and Lemma I.3 we derive the following result for the convergence of `Local-SVRG` in the case when $\mu = 0$.

Corollary G.12. *Let assumptions of Theorem G.6 hold with* $\mu = 0$. *Then for*

$$\begin{aligned}
\gamma_0 &= \min\left\{\frac{1}{2\left(44\max L_{ij}/n + L\right)}, \frac{1}{16\sqrt{L\max L_{ij}(\tau-1)\left(1 + 4/(1-q)\right)}}\right\}, \quad q = \frac{1}{m}, \quad m > 1,\\
\gamma &= \min\left\{\gamma_0, \sqrt{\frac{3nR_0^2}{4m\sigma_0^2}}, \sqrt[3]{\frac{R_0^2}{16Lm(\tau-1)(2+1/m)\sigma_0^2}}, \sqrt{\frac{\mu R_0^2}{4L\zeta^2(\tau-1)K}}, \sqrt[3]{\frac{R_0^2}{8L(\tau-1)\zeta_*^2K}}\right\},
\end{aligned}$$

where $R_0 = \|x^0 - x^*\|$, *we have that* $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right]$ *is of the order*

$$\mathcal{O}\left(\frac{(L + \max L_{ij}/n + \sqrt{(\tau-1)L\max L_{ij}})R_0^2 + \sqrt{m\sigma_0^2 R_0^2/n} + \sqrt[3]{Lm(\tau-1)\sigma_0^2 R_0^4}}{K} + \sqrt{\frac{LR_0^2\zeta^2(\tau-1)}{\mu K}} + \frac{\sqrt[3]{LR_0^4(\tau-1)\zeta_*^2}}{K^{2/3}}\right).$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ *in this case* `Local-SVRG` *requires*

$$\mathcal{O}\left(\frac{(L + \max L_{ij}/n + \sqrt{(\tau-1)L\max L_{ij}})R_0^2 + \sqrt{m\sigma_0^2 R_0^2/n} + \sqrt[3]{Lm(\tau-1)\sigma_0^2 R_0^4}}{\varepsilon} + \frac{L\zeta^2(\tau-1)R_0^2}{\mu\varepsilon^2} + \frac{R_0^2\sqrt{L(\tau-1)\zeta_*^2}}{\varepsilon^{3/2}}\right)$$

iterations/oracle calls per node and τ *times less communication rounds.*

Remark G.1. *To get the rate from Tbl. 4 it remains to apply the following inequality:*

$$\sigma_0^2 = \frac{4}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{ij}(x^0) - \nabla f_{ij}(x^*)\|^2 \overset{(6)}{\leq} 4\max L_{ij}^2 \|x^0 - x^*\|^2.$$

G.2.2 Heterogeneous Data

First of all, we need the following lemma.

Lemma G.6. *Assume that* $f_i(x)$ *is* L*-smooth for* $i \in [n]$ *and* f_{ij} *is convex and* $\max L_{ij}$*-smooth for* $i \in [n], j \in [m]$. *Then for* `Local-SVRG` *we have*

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\bar{g}_i^k\|^2\right] \leq 6L\mathbf{E}\left[f(x^k) - f(x^*)\right] + 3L^2\mathbf{E}[V_k] + 3\zeta_*^2, \tag{99}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|g_i^k - \bar{g}_i^k\right\|^2\right] \leq 8\max L_{ij}\mathbf{E}\left[f(x^k) - f(x^*)\right] + \frac{1}{2}\mathbf{E}[\sigma_k^2] + 4L\max L_{ij}\mathbf{E}[V_k], \tag{100}$$

where $\sigma_k^2 = \frac{4}{nm}\sum\limits_{i=1}^n\sum\limits_{j=1}^m \|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2$.

Proof. Inequality (99) follows from $\bar{g}_i^k = \mathbf{E}\left[g_i^k \mid x^k\right] = \nabla f_i(x_i^k)$ and inequality (67). Next, using Young's inequality we derive

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|g_i^k - \bar{g}_i^k\right\|^2\right] &\overset{(139)}{\leq} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|g_i^k - \nabla f_i(x^*)\right\|^2\right] \\
&\overset{(136)}{\leq} \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{ij_i}(x_i^k) - \nabla f_{ij_i}(x^*)\|^2\right] \\
&\quad + \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{ij_i}(w_i^k) - \nabla f_{ij_i}(x^*) - (\nabla f_i(w_i^k) - \nabla f_i(x^*))\|^2\right] \\
&\overset{(140)}{=} \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \mathbf{E}\left[\|\nabla f_{ij}(x_i^k) - \nabla f_{ij}(x^*)\|^2\right] \\
&\quad + \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \mathbf{E}\left[\|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*) - (\nabla f_i(w_i^k) - \nabla f_i(x^*))\|^2\right] \\
&\overset{(19),(139)}{\leq} \frac{4\max L_{ij}}{n}\sum_{i=1}^n \mathbf{E}\left[D_{f_i}(x_i^k, x^*)\right] + \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \mathbf{E}\left[\|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2\right] \\
&\overset{(63)}{\leq} 8\max L_{ij}\mathbf{E}\left[f(x^k) - f(x^*)\right] + \frac{1}{2}\mathbf{E}[\sigma_k^2] + 4L\max L_{ij}\mathbf{E}[V_k].
\end{aligned}$$

□

Applying Corollary E.1, Lemma G.6, Proposition G.1 and Lemma 4.1 we get the following result.

Theorem G.7. *Assume that $f_i(x)$ is μ-strongly convex and L-smooth for $i \in [n]$ and f_{ij} is convex and $\max L_{ij}$-smooth for $i \in [n], j \in [m]$. Then* `Local-SVRG` *satisfies Assumption E.1 with*

$$\widetilde{A} = 3L, \quad \widehat{A} = 4\max L_{ij}, \quad \widetilde{B} = 0, \quad \widehat{B} = \frac{1}{2}, \quad \widetilde{F} = 3L^2, \quad \widehat{F} = 4L\max L_{ij}, \quad \widetilde{D}_1 = 3\zeta_*^2, \quad \widehat{D}_1 = 0$$

$$A' = \frac{4\max L_{ij}}{n} + L, \quad B' = \frac{1}{n}, \quad F' = \frac{4L\max L_{ij}}{n} + 2L^2, \quad D_1' = 0,$$

$$\sigma_k^2 = \frac{4}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\|\nabla f_{ij}(w_i^k) - \nabla f_{ij}(x^*)\|^2, \quad \rho = q, \quad C = 8q\max L_{ij}, \quad G = 4qL\max L_{ij}, \quad D_2 = 0,$$

$$H = \frac{2e(\tau-1)(2+q)\gamma^2}{q}, \quad D_3 = 6e(\tau-1)^2\zeta_*^2$$

with γ satisfying

$$\gamma \quad \leq \quad \min\left\{\frac{1}{2\left(44\max L_{ij}/n + L\right)}, \frac{1}{4\sqrt{2eL(\tau-1)\left(3L(\tau-1) + 4\max L_{ij} + 8\max L_{ij}/(1-q)\right)}}\right\}.$$

and for all $K \geq 0$

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \frac{\Phi^0}{\gamma W_K} + 24eL(\tau-1)^2\zeta_*^2\gamma^2,$$

where $\Phi^0 = 2\|x^0 - x^*\|^2 + \frac{8}{3nq}\gamma^2\sigma_0^2 + \frac{8eL(\tau-1)(2+q)\gamma^3}{q}\sigma_0^2$ In particular, if $\mu > 0$ then

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \left(1 - \min\left\{\gamma\mu, \frac{q}{4}\right\}\right)^K \frac{\Phi^0}{\gamma} + 24eL(\tau-1)^2\zeta_*^2\gamma^2 \tag{101}$$

and when $\mu = 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \frac{\Phi^0}{\gamma K} + 24eL(\tau-1)^2\zeta_*^2\gamma^2. \tag{102}$$

The theorem above together with Lemma I.2 implies the following result.

Corollary G.13. *Let assumptions of Theorem G.7 hold with $\mu > 0$. Then for*

$$\begin{aligned}
\gamma_0 \quad &= \quad \min\left\{\frac{1}{2\left(44\max L_{ij}/n + L\right)}, \frac{1}{4\sqrt{2eL(\tau-1)\left(3L(\tau-1) + 4\max L_{ij} + 8\max L_{ij}/(1-q)\right)}}\right\}, \\
\widetilde{\Phi}^0 \quad &= \quad 2\|x^0 - x^*\|^2 + \frac{8}{3nq}\gamma_0^2\sigma_0^2 + \frac{8eL(\tau-1)(2+q)\gamma_0^3}{q}\sigma_0^2, \quad q = \frac{1}{m}, \quad m > 1, \\
\gamma \quad &= \quad \min\left\{\gamma_0, \frac{\ln\left(\max\left\{2, \widetilde{\Phi}^0\mu^3K^3/24eL(\tau-1)^2\zeta_*^2\right\}\right)}{\mu K}\right\},
\end{aligned}$$

for all K such that

$$\textit{either} \qquad \frac{\ln\left(\max\left\{2, \widetilde{\Phi}^0\mu^3K^3/24eL(\tau-1)^2\zeta_*^2\right\}\right)}{K} \leq \frac{1}{m} \qquad \textit{or} \qquad \gamma_0 \leq \frac{\ln\left(\max\left\{2, \widetilde{\Phi}^0\mu^3K^3/24eL(\tau-1)^2\zeta_*^2\right\}\right)}{\mu K}$$

we have that $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right]$ is of the order

$$\widetilde{\mathcal{O}}\left(\frac{\widetilde{\Phi}^0}{\gamma_0}\exp\left(-\min\left\{m^{-1}, \gamma_0\mu\right\}K\right) + \frac{L(\tau-1)^2\zeta_*^2}{\mu^2K^2}\right).$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ *in this case* `Local-SVRG` *requires*

$$\widetilde{\mathcal{O}}\left(m + \frac{L\tau}{\mu} + \frac{\max L_{ij}}{n\mu} + \frac{\sqrt{(\tau-1)L\max L_{ij}}}{\mu} + \sqrt{\frac{L(\tau-1)^2\zeta_*^2}{\mu^2\varepsilon}}\right)$$

iterations/oracle calls per node and τ *times less communication rounds.*

Combining Theorem G.7 and Lemma I.3 we derive the following result for the convergence of `Local-SVRG` in the case when $\mu = 0$.

Corollary G.14. *Let assumptions of Theorem G.7 hold with* $\mu = 0$. *Then for* $q = \frac{1}{m}$, $m > 1$ *and*

$$\begin{aligned}\gamma_0 &= \min\left\{\frac{1}{2\left(44\max L_{ij}/n + L\right)}, \frac{1}{4\sqrt{2eL(\tau-1)\left(3L(\tau-1) + 4\max L_{ij} + 8\max L_{ij}/(1-q)\right)}}\right\},\\ \gamma &= \min\left\{\gamma_0, \sqrt{\frac{3nR_0^2}{4m\sigma_0^2}}, \sqrt[3]{\frac{R_0^2}{4eLm(\tau-1)(2+1/m)\sigma_0^2}}, \sqrt[3]{\frac{R_0^2}{12eL(\tau-1)^2\zeta_*^2K}}\right\},\end{aligned}$$

where $R_0 = \|x^0 - x^*\|$, *we have that* $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right]$ *is of the order*

$$\mathcal{O}\left(\frac{(L\tau + \max L_{ij}/n + \sqrt{(\tau-1)L\max L_{ij}})R_0^2 + \sqrt{m\sigma_0^2R_0^2/n} + \sqrt[3]{Lm(\tau-1)\sigma_0^2R_0^4}}{K} + \frac{\sqrt[3]{LR_0^4(\tau-1)^2\zeta_*^2}}{K^{2/3}}\right).$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ *in this case* `Local-SVRG` *requires*

$$\mathcal{O}\left(\frac{(L\tau + \max L_{ij}/n + \sqrt{(\tau-1)L\max L_{ij}})R_0^2 + \sqrt{m\sigma_0^2R_0^2/n} + \sqrt[3]{Lm(\tau-1)\sigma_0^2R_0^4}}{\varepsilon} + \frac{R_0^2\sqrt{L(\tau-1)^2\zeta_*^2}}{\varepsilon^{3/2}}\right)$$

iterations/oracle calls per node and τ *times less communication rounds.*

Remark G.2. *To get the rate from Tbl. 4 it remains to apply the following inequality:*

$$\sigma_0^2 = \frac{4}{nm}\sum_{i=1}^n\sum_{j=1}^m\|\nabla f_{ij}(x^0) - \nabla f_{ij}(x^*)\|^2 \overset{(6)}{\leq} 4\max L_{ij}^2\|x^0 - x^*\|^2.$$

G.3 `S*-Local-SGD`

In this section we consider the same settings as in Section G.1.1 and our goal is to remove one of the main drawbacks of `Local-SGD` in heterogeneous case which in the case of μ-strongly convex f_i with $\mu > 0$ converges with linear rate only to the neighbourhood of the solution even in the full-gradients case, i.e. when $D_{1,i} = 0$ for all $i \in [n]$. However, we start with unrealistic assumption that i-th node has an access to $\nabla f_i(x^*)$ for all $i \in [n]$. Under this assumption we present a new method called Star-Shifted Local-SGD (`S*-Local-SGD`, see Algorithm 3).

Lemma G.7. *Let* f_i *be convex and* L*-smooth for all* $i \in [n]$. *Then for all* $k \geq 0$

$$\frac{1}{n}\sum_{i=1}^n\mathbf{E}\left[g_i^k \mid x_i^k\right] = \frac{1}{n}\sum_{i=1}^n\nabla f_i(x_i^k), \tag{103}$$

$$\frac{1}{n}\sum_{i=1}^n\|\bar{g}_i^k\|^2 \leq 4L\left(f(x^k) - f(x^*)\right) + 2L^2V_k, \tag{104}$$

$$\frac{1}{n}\sum_{i=1}^n\mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x_i^k\right] \leq \sigma^2, \tag{105}$$

$$\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k\right\|^2 \mid x^k\right] \leq 4L\left(f(x^k) - f(x^*)\right) + 2L^2V_k + \frac{\sigma^2}{n}, \tag{106}$$

where $\sigma^2 \overset{def}{=} \frac{1}{n}\sum_{i=1}^n D_{1,i}$ *and* $\mathbf{E}[\cdot \mid x^k] \overset{def}{=} \mathbf{E}[\cdot \mid x_1^k, \ldots, x_n^k]$.

Algorithm 3 `S*-Local-SGD`

Require: learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$, communication period $\tau \geq 1$

for $k = 0, 1, \ldots$ **do**
  for $i = 1, \ldots, n$ in parallel **do**
    Sample $\hat{g}_i^k = \nabla f_{\xi_i^k}(x_i^k)$ independently from other nodes
    $g_i^k = \hat{g}_i^k - \nabla f_i(x^*)$
    if $k+1 \mod \tau = 0$ **then**
      $x_i^{k+1} = x^{k+1} = \frac{1}{n}\sum_{i=1}^n \left(x_i^k - \gamma g_i^k\right)$ ▷ averaging
    else
      $x_i^{k+1} = x_i^k - \gamma g_i^k$ ▷ local update
    end if
  end for
end for

Proof. First of all, we notice that $\mathbf{E}\left[g_i^k \mid x_i^k\right] = \nabla f_i(x_i^k) - \nabla f_i(x^*)$ and

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[g_i^k \mid x_i^k\right] = \frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(x_i^k) - \nabla f_i(x^*)\right) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k).$$

Using this we get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \|\bar{g}_i^k\|^2 &= \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x_i^k) - \nabla f_i(x^*)\|^2 \overset{(19)}{\leq} \frac{2L}{n}\sum_{i=1}^n D_{f_i}(x_i^k, x^*) \\ &\overset{(63)}{\leq} 4L\left(f(x^k) - f(x^*)\right) + 2L^2 V_k\end{aligned}$$

and

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2 \mid x_i^k\right] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x_i^k)\|^2\right] \overset{(66)}{\leq} \frac{1}{n}\sum_{i=1}^n D_{1,i} =: \sigma^2.$$

Finally, using independence of $g_1^k, g_2^k, \ldots, g_n^k$ and $\frac{1}{n}\sum_{i=1}^n \nabla f_i(x^*) = \nabla f(x^*) = 0$ we obtain

$$\begin{aligned}\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k\right\|^2 \mid x^k\right] &\overset{(139),(103)}{=} \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(g_i^k - \nabla f_i(x_i^k)\right)\right\|^2 \mid x^k\right] + \left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)\right\|^2 \\ &= \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x_i^k)\right)\right\|^2 \mid x^k\right] + \left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)\right\|^2 \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}_{\xi_i^k}\left[\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x_i^k)\|^2\right] + \left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)\right\|^2 \\ &\overset{(66),(68)}{\leq} 4L\left(f(x^k) - f(x^*)\right) + 2L^2 V_k + \frac{\sigma^2}{n}.\end{aligned}$$

□

Applying Corollary E.1 and Lemma G.7 we get the following result.

Theorem G.8. *Assume that $f_i(x)$ is μ-strongly convex and L-smooth for every $i \in [n]$. Then* `S*-Local-SGD` *satisfies Assumption E.1 with*

$$\widetilde{A} = 2L, \quad \widehat{A} = 0, \quad \widetilde{B} = \widehat{B} = 0, \quad \widetilde{F} = 2L^2, \quad \widehat{F} = 0, \quad \widetilde{D}_1 = 0, \quad \widehat{D}_1 = \sigma^2 := \frac{1}{n}\sum_{i=1}^n D_{1,i}$$

$$A' = 2L, \quad B' = 0, \quad F' = 2L^2, \quad D_1' = \frac{\sigma^2}{n}, \quad \sigma_k^2 \equiv 0, \quad \rho = 1, \quad C = 0, \quad G = 0, \quad D_2 = 0,$$
$$H = 0, \quad D_3 = 2e(\tau - 1)\sigma^2.$$

Consequently, if

$$\gamma \quad \leq \quad \min\left\{\frac{1}{4L}, \frac{1}{8\sqrt{e}(\tau-1)L}\right\}.$$

we have for $\mu > 0$

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad (1-\gamma\mu)^K \frac{2\|x^0 - x^*\|^2}{\gamma} + 2\gamma\left(\frac{\sigma^2}{n} + 4eL(\tau-1)\gamma\sigma^2\right)$$

and when $\mu = 0$ *we have*

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \frac{2\|x^0 - x^*\|^2}{\gamma K} + 2\gamma\left(\frac{\sigma^2}{n} + 4eL(\tau-1)\gamma\sigma^2\right).$$

In the special case when $\nabla f_{\xi_i^k}(x_i^k) = \nabla f_i(x_i^k)$ for all $i \in [n]$ and $k \geq 0$ we obtain `S*-Local-GD` which converges with $\mathcal{O}\left(\tau\kappa \ln\frac{1}{\varepsilon}\right)$ rate when $\mu > 0$ and with $\mathcal{O}\left(\frac{L\tau\|x^0-x^*\|^2}{\varepsilon}\right)$ rate when $\mu = 0$ to the exact solution asymptotically.

The theorem above together with Lemma I.2 implies the following result.

Corollary G.15. *Let assumptions of Theorem G.8 hold with* $\mu > 0$*. Then for*

$$\gamma = \min\left\{\frac{1}{4L}, \frac{1}{8\sqrt{e}(\tau-1)L}, \frac{\ln\left(\max\left\{2, \min\left\{\|x^0-x^*\|^2 n\mu^2K^2/\sigma^2, \|x^0-x^*\|^2\mu^3K^3/4eL(\tau-1)\sigma^2\right\}\right\}\right)}{\mu K}\right\}$$

for all K *such that*

$$\textit{either} \quad \frac{\ln\left(\max\left\{2, \min\left\{\|x^0-x^*\|^2 n\mu^2K^2/\sigma^2, \|x^0-x^*\|^2\mu^3K^3/4eL(\tau-1)\sigma^2\right\}\right\}\right)}{K} \leq 1$$

$$\textit{or} \quad \min\left\{\frac{1}{4L}, \frac{1}{8\sqrt{e}(\tau-1)L}\right\} \leq \frac{\ln\left(\max\left\{2, \min\left\{\|x^0-x^*\|^2 n\mu^2K^2/\sigma^2, \|x^0-x^*\|^2\mu^3K^3/4eL(\tau-1)\sigma^2\right\}\right\}\right)}{\mu K}$$

we have that

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] = \widetilde{\mathcal{O}}\left(\tau L\|x^0 - x^*\|^2 \exp\left(-\frac{\mu}{\tau L}K\right) + \frac{\sigma^2}{n\mu K} + \frac{L(\tau-1)\sigma^2}{\mu^2K^2}\right).$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ *in this case* `S*-Local-SGD` *requires*

$$\widetilde{\mathcal{O}}\left(\frac{\tau L}{\mu} + \frac{\sigma^2}{n\mu\varepsilon} + \sqrt{\frac{L(\tau-1)\sigma^2}{\mu^2\varepsilon}}\right)$$

iterations/oracle calls per node and τ *times less communication rounds.*

Combining Theorem G.8 and Lemma I.3 we derive the following result for the convergence of `S*-Local-SGD` in the case when $\mu = 0$.

Corollary G.16. *Let assumptions of Theorem G.8 hold with* $\mu = 0$*. Then for*

$$\gamma = \min\left\{\frac{1}{4L}, \frac{1}{8\sqrt{e}(\tau-1)L}, \sqrt{\frac{nR_0^2}{\sigma^2K}}, \sqrt[3]{\frac{R_0^2}{4eL(\tau-1)\sigma^2K}}\right\},$$

where $R_0 = \|x^0 - x^*\|$*, we have that*

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] = \mathcal{O}\left(\frac{\tau LR_0^2}{K} + \sqrt{\frac{R_0^2\sigma^2}{nK}} + \frac{\sqrt[3]{LR_0^4(\tau-1)\sigma^2}}{K^{2/3}}\right).$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ *in this case* `S*-Local-SGD` *requires*

$$\mathcal{O}\left(\frac{\tau LR_0^2}{\varepsilon} + \frac{R_0^2\sigma^2}{n\varepsilon^2} + \frac{R_0^2\sqrt{L(\tau-1)\sigma^2}}{\varepsilon^{3/2}}\right)$$

iterations/oracle calls per node and τ *times less communication rounds.*

G.4 SS-Local-SGD

G.4.1 Uniformly Bounded Variance

In this section we consider the same settings as in Section G.1.1

Algorithm 4 Stochastically Shifted Local-SGD (SS-Local-SGD)

Require: learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$, probability of communication $p \in (0,1]$, probability of the shift's update $q \in (0,1]$, batchsize r for computing shifts

1: $y^0 = x^0$

2: For $i \in [n]$ compute r independent samples $\nabla f_{\bar{\xi}_{i,1}^0}(y^0), \nabla f_{\bar{\xi}_{i,2}^0}(y^0), \ldots, \nabla f_{\bar{\xi}_{i,r}^0}(y^0)$, set $\nabla f_{\bar{\xi}_i^0}(y^0) = \frac{1}{r}\sum_{j=1}^r \nabla f_{\bar{\xi}_{i,j}^0}(y^0)$ and $\nabla f_{\bar{\xi}^0}(y^0) = \frac{1}{n}\sum_{i=1}^n \nabla f_{\bar{\xi}_i^0}(y^0)$

3: **for** $k = 0, 1, \ldots$ **do**

4: **for** $i = 1, \ldots, n$ in parallel **do**

5: Sample $\nabla f_{\xi_i^k}(x_i^k)$ independently from other nodes

6: $g_i^k = \nabla f_{\xi_i^k}(x_i^k) - \nabla f_{\widetilde{\xi}_i^k}(y^k) + \nabla f_{\widetilde{\xi}^k}(y^k)$, where $\nabla f_{\bar{\xi}_i^k}(y^k) = \frac{1}{r}\sum_{j=1}^r \nabla f_{\bar{\xi}_{i,j}^k}(y^k)$ and $\nabla f_{\bar{\xi}^k}(y^k) = \frac{1}{n}\sum_{i=1}^n \nabla f_{\bar{\xi}_i^k}(y^k)$

7: $x_i^{k+1} = \begin{cases} x^{k+1}, & \text{w.p. } p, \\ x_i^k - \gamma g_i^k, & \text{w.p. } 1-p, \end{cases}$ where $x^{k+1} = \frac{1}{n}\sum\limits_{i=1}^n (x_i^k - \gamma g_i^k)$

8: $y^{k+1} = \begin{cases} x^k, & \text{w.p. } q, \\ y^k, & \text{w.p. } 1-q, \end{cases}$ and for all $i \in [n]$, $j \in [r]$ $\bar{\xi}_{i,j}^{k+1}$ is $\begin{cases} \text{a fresh sample}, & \text{if } y^{k+1} \neq y^k, \\ \text{equal to } \bar{\xi}_{i,j}^k, & \text{otherwise.} \end{cases}$

9: **end for**

10: **end for**

The main algorithm in this section is Stochastically Shifted Local-SGD (SS-Local-SVRG, see Algorithm 4). We notice that the updates for x_i^{k+1} and y^{k+1} can be dependent, e.g., one can take $p = q$ and update y^{k+1} as x^k every time x_i^{k+1} is updated by x^{k+1}. Moreover, with probability q line 8 implies a round of communication and computation of new stochastic gradient by each worker.

We emphasize that in expectation y^k is updated only once per $\lceil 1/q \rceil$ iterations. Therefore, if $r = O(1/q)$ and $q \leq p$, then up to a constant numerical factor the overall expected number of oracle calls and communication rounds are the same as for Local-SGD with either the same probability p of communication or with constant local loop length $\tau = \lceil 1/p \rceil$.

Finally, we notice that due to independence of $\bar{\xi}_{i,1}^k, \bar{\xi}_{i,2}^k, \ldots, \bar{\xi}_{i,r}^k$ we have

$$\mathbf{E}\|\nabla f_{\bar{\xi}_i^k}(y^k) - \nabla f_i(y^k)\|^2 \overset{(66)}{\leq} \frac{D_{1,i}}{r}. \tag{107}$$

Lemma G.8. *Let f_i be convex and L-smooth for all $i \in [n]$. Then for all $k \geq 0$*

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}_k\left[g_i^k\right] = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k), \tag{108}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\bar{g}_i^k\|^2\right] \leq 8L\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2\mathbf{E}[\sigma_k^2] + 4L^2\mathbf{E}[V_k] + \frac{2\sigma^2}{r}, \tag{109}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2\right] \leq \sigma^2, \tag{110}$$

$$\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k\right\|^2\right] \leq 4L\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2L^2\mathbf{E}\left[V_k\right] + \frac{\sigma^2}{n}, \tag{111}$$

where $\sigma_k^2 \overset{def}{=} \frac{1}{n}\sum\limits_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^*)\right\|^2$ and $\sigma^2 \overset{def}{=} \frac{1}{n}\sum_{i=1}^n D_{1,i}$.

Proof. We start with unbiasedness:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbf{E}_k\left[g_i^k\right] &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}_k\left[\nabla f_{\xi_i^k}(x_i^k) - \nabla f_{\bar{\xi}_i^k}(y^k) + \nabla f_{\bar{\xi}^k}(y^k)\right]\\ &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}_k\left[\nabla f_{\xi_i^k}(x_i^k)\right] + \mathbf{E}_k\left[\nabla f_{\bar{\xi}^k}(y^k) - \frac{1}{n}\sum_{i=1}^n \nabla f_{\bar{\xi}_i^k}(y^k)\right] = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k).\end{aligned}$$

Using this we get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\bar{g}_i^k\|^2\right] &\overset{(136)}{\leq} \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_i(x_i^k) - \nabla f_i(x^*)\|^2\right]\\ &\quad + \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_{\bar{\xi}_i^k}(y^k) - \nabla f_i(x^*) - \left(\nabla f_{\bar{\xi}^k}(y^k) - \nabla f(x^*)\right)\right\|^2\right]\\ &\overset{(19),(139)}{\leq} \frac{4L}{n}\sum_{i=1}^n \mathbf{E}\left[D_{f_i}(x_i^k, x^*)\right] + \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_{\bar{\xi}_i^k}(y^k) - \nabla f_i(x^*)\right\|^2\right]\\ &\overset{(63),(139)}{\leq} 8L\mathbf{E}\left[f(x^k) - f(x^*)\right] + 4L^2\mathbf{E}[V_k] + \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_i(y^k) - \nabla f_i(x^*)\right\|^2\right]\\ &\quad + \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_{\bar{\xi}_i^k}(y^k) - \nabla f_i(y^k)\right\|^2\right]\\ &\overset{(107)}{\leq} 8L\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2\mathbf{E}[\sigma_k^2] + 4L^2\mathbf{E}[V_k] + \frac{2\sigma^2}{r}\end{aligned}$$

and

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2\right] = \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x_i^k)\|^2\right] \overset{(66)}{\leq} \sigma^2.$$

Finally, we use independence of $\nabla f_{\xi_1^k}(x_1^k), \ldots, \nabla f_{\xi_n^k}(x_n^k)$ and derive

$$\begin{aligned}\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k\right\|^2\right] &= \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_{\xi_i^k}(x_i^k)\right\|^2\right]\\ &\overset{(139)}{=} \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)\right\|^2\right] + \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x_i^k)\right)\right\|^2\right]\\ &\overset{(68)}{\leq} 4L\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2L^2\mathbf{E}[V_k] + \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x_i^k)\|^2\right]\\ &\overset{(66)}{\leq} 4L\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2L^2\mathbf{E}\left[V_k\right] + \frac{\sigma^2}{n}\end{aligned}$$

which finishes the proof. □

Lemma G.9. *Let f_i be convex and L-smooth for all $i \in [n]$. Then for all $k \geq 0$*

$$\mathbf{E}\left[\sigma_{k+1}^2\right] \leq (1-q)\mathbf{E}\left[\sigma_k^2\right] + 2Lq\mathbf{E}\left[f(x^k) - f(x^*)\right] \tag{112}$$

where $\sigma_k^2 \overset{def}{=} \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^*)\right\|^2$.

Proof. By definition of y^{k+1} we have

$$\begin{aligned}\mathbf{E}\left[\sigma_{k+1}^2 \mid x_1^k, \ldots, x_n^k\right] &= \frac{1-q}{n}\sum_{i=1}^n \|\nabla f_i(y^k) - \nabla f_i(x^*)\|^2 + \frac{q}{n}\sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2\\ &\overset{(19)}{\leq} (1-q)\sigma_k^2 + 2Lq(f(x^k) - f(x^*)).\end{aligned}$$

Taking the full mathematical expectation on both sides of previous inequality and using the tower property (140) we get the result. □

Using Corollary E.3 we obtain the following theorem.

Theorem G.9. *Assume that $f_i(x)$ is μ-strongly convex and L-smooth for every $i \in [n]$. Then* `SS-Local-SGD` *satisfies Assumption E.1 with*

$$\widetilde{A} = 4L, \quad \widehat{A} = 0, \quad \widetilde{B} = 2, \quad \widehat{B} = 0, \quad \widetilde{F} = 4L^2, \quad \widehat{F} = 0, \quad \widetilde{D}_1 = \frac{2\sigma^2}{r}, \quad \widehat{D}_1 = \sigma^2, \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^n D_{1,i},$$

$$A' = 2L, \quad B' = 0, \quad F' = 2L^2, \quad D_1' = \frac{\sigma^2}{n},$$

$$\sigma_k^2 = \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^*)\right\|^2, \quad \rho = q, \quad C = Lq, \quad G = 0, \quad D_2 = 0,$$

$$H = \frac{128(1-p)(2+p)(2+q)\gamma^2}{3p^2q}, \quad D_3 = \frac{8(1-p)}{p^2}\left(\frac{2(p+2)\sigma^2}{r} + p\sigma^2\right)$$

under assumption that

$$\gamma \quad \leq \quad \min\left\{\frac{1}{4L}, \frac{p\sqrt{3}}{32L\sqrt{2(1-p)(2+p)\left(1+{}^1\!/_{(1-q)}\right)}}\right\}.$$

Moreover, for $\mu > 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \left(1 - \min\left\{\gamma\mu, \frac{q}{4}\right\}\right)^K \frac{\Phi^0}{\gamma} + 2\gamma\left(\frac{\sigma^2}{n} + \gamma\frac{16L(1-p)}{p^2}\left(\frac{2(p+2)\sigma^2}{r} + p\sigma^2\right)\right)$$

and when $\mu = 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \quad \leq \quad \frac{\Phi^0}{\gamma K} + 2\gamma\left(\frac{\sigma^2}{n} + \gamma\frac{16L(1-p)}{p^2}\left(\frac{2(p+2)\sigma^2}{r} + p\sigma^2\right)\right)$$

where $\Phi^0 = 2\|x^0 - x^*\|^2 + \frac{512L(1-p)(2+p)(2+q)\gamma^3\sigma_0^2}{3p^2q}$.

The theorem above together with Lemma I.2 implies the following result.

Corollary G.17. *Let assumptions of Theorem G.9 hold with $\mu > 0$. Then for*

$$\begin{aligned}
\gamma_0 \quad &= \quad \min\left\{\frac{1}{4L}, \frac{p\sqrt{3}}{32L\sqrt{2(1-p)(2+p)\left(1+{}^1\!/_{(1-q)}\right)}}\right\}, \\
\widetilde{\Phi}^0 \quad &= \quad 2\|x^0 - x^*\|^2 + \frac{512L(1-p)(2+p)(2+q)\gamma_0^3\sigma_0^2}{3p^2q}, \quad q = p, \\
\gamma \quad &= \quad \min\left\{\gamma_0, \frac{\ln\left(\max\left\{2, \min\left\{{}^{n\widetilde{\Phi}^0\mu^2K^2}\!/_{2\sigma^2}, {}^{p\widetilde{\Phi}^0\mu^3K^3}\!/_{32L(1-p)(3p+4)\sigma^2}\right\}\right\}\right)}{\mu K}\right\}, \quad r = \left[\frac{1}{p}\right],
\end{aligned}$$

for all K such that

$$\begin{aligned}
&\textit{either} \quad && \frac{\ln\left(\max\left\{2, \min\left\{{}^{n\widetilde{\Phi}^0\mu^2K^2}\!/_{2\sigma^2}, {}^{p\widetilde{\Phi}^0\mu^3K^3}\!/_{32L(1-p)(3p+4)\sigma^2}\right\}\right\}\right)}{K} \leq p \\
&\textit{or} \quad && \gamma_0 \leq \frac{\ln\left(\max\left\{2, \min\left\{{}^{n\widetilde{\Phi}^0\mu^2K^2}\!/_{2\sigma^2}, {}^{p\widetilde{\Phi}^0\mu^3K^3}\!/_{32L(1-p)(3p+4)\sigma^2}\right\}\right\}\right)}{\mu K}
\end{aligned}$$

we have that $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right]$ is of the order

$$\widetilde{\mathcal{O}}\left(\frac{\widetilde{\Phi}^0}{\gamma_0}\exp\left(-\min\left\{\frac{1}{p}, \gamma_0\mu\right\}K\right) + \frac{\sigma^2}{n\mu K} + \frac{L(1-p)\sigma^2}{p\mu^2K^2}\right).$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ *in this case* `SS-Local-SGD` *requires*

$$\widetilde{\mathcal{O}}\left(\frac{L}{p\mu} + \frac{\sigma^2}{n\mu\varepsilon} + \sqrt{\frac{L(1-p)\sigma^2}{p\mu^2\varepsilon}}\right)$$

iterations/oracle calls per node (in expectation) and $1/p$ *times less communication rounds.*

Combining Theorem G.9 and Lemma I.3 we derive the following result for the convergence of `SS-Local-SGD` in the case when $\mu = 0$.

Corollary G.18. *Let assumptions of Theorem G.9 hold with* $\mu = 0$. *Then for* $q = p$, $r = \lceil 1/p \rceil$ *and*

$$\begin{aligned}\gamma_0 &= \min\left\{\frac{1}{4L}, \frac{p\sqrt{3}}{32L\sqrt{2(1-p)(2+p)\left(1+1/(1-q)\right)}}\right\},\\ \gamma &= \min\left\{\gamma_0, \sqrt[3]{\frac{3p^3R_0^2}{256L(1-p)(2+p)^2\sigma_0^2}}, \sqrt{\frac{nR_0^2}{\sigma^2K}}, \sqrt[3]{\frac{pR_0^2}{16L(1-p)(3p+4)\sigma^2K}}\right\},\end{aligned}$$

where $R_0 = \|x^0 - x^*\|$, *we have that* $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right]$ *is of the order*

$$\mathcal{O}\left(\frac{LR_0^2 + \sqrt[3]{L(1-p)\sigma_0^2R_0^4}}{pK} + \sqrt{\frac{\sigma^2R_0^2}{nK}} + \frac{\sqrt[3]{LR_0^4(1-p)\sigma^2}}{p^{1/3}K^{2/3}}\right).$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ *in this case* `SS-Local-SGD` *requires*

$$\mathcal{O}\left(\frac{LR_0^2 + \sqrt[3]{L(1-p)\sigma_0^2R_0^4}}{p\varepsilon} + \frac{\sigma^2R_0^2}{n\varepsilon^2} + \frac{R_0^2\sqrt{L(1-p)\sigma^2}}{p^{1/2}\varepsilon^{3/2}}\right)$$

iterations/oracle calls per node (in expectation) and $1/p$ *times less communication rounds.*

Remark G.3. *To get the rate from Tbl. 4 it remains to apply the following inequality:*

$$\sigma_0^2 = \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^0) - \nabla f_i(x^*)\|^2 \overset{(6)}{\leq} L^2\|x^0 - x^*\|^2.$$

G.4.2 Expected Smoothness and Arbitrary Sampling

In this section we consider the same method `SS-Local-SGD`, but without assumption that the stochastic gradient has a uniformly bounded variance. Instead of this we consider the same setup as in Section G.1.2, i.e. we assume that each worker $i \in [n]$ at any point $x \in \mathbb{R}^d$ has an access to the unbiased estimator $\nabla f_{\xi_i}(x)$ of $\nabla f_i(x)$ satisfying Assumption G.1.

Lemma G.10. *Let* f_i *be convex and* L*-smooth for all* $i \in [n]$. *Let Assumption G.1 holds. Then for all* $k \geq 0$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}_k\left[g_i^k\right] = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k), \tag{113}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\bar{g}_i^k\|^2\right] \leq 8L\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2\mathbf{E}[\sigma_k^2] + 4L^2\mathbf{E}[V_k], \tag{114}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2\right] \leq 8\mathcal{L}\mathbf{E}\left[f(x^k) - f(x^*)\right] + 4\mathcal{L}L\mathbf{E}[V_k] + 2\sigma_*^2, \tag{115}$$

$$\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k\right\|^2\right] \leq 4\left(\frac{2\mathcal{L}}{n} + L\right)\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2L\left(\frac{2\mathcal{L}}{n} + L\right)\mathbf{E}[V_k] + \frac{2\sigma_*^2}{n}, \tag{116}$$

where $\sigma_k^2 \overset{def}{=} \frac{1}{n}\sum\limits_{i=1}^n \left\|\nabla f_{\xi_i^k}(y^k) - \nabla f_i(x^*)\right\|^2$ *and* $\sigma_*^2 \overset{def}{=} \frac{1}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i}\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2$.

Proof. First of all, (113) follows from (108). Next, using $\bar{g}_i^k = \nabla f_i(x_i^k) - \nabla f_{\bar{\xi}_i^k}(y^k) + \nabla f_{\bar{\xi}^k}(y^k)$ we get

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\bar{g}_i^k\|^2\right] &\overset{(136)}{\leq} \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_i(x_i^k) - \nabla f_i(x^*)\|^2\right] \\
&\quad + \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\left\|\nabla f_{\bar{\xi}_i^k}(y^k) - \nabla f_i(x^*) - (\nabla f_{\bar{\xi}^k}(y^k) - \nabla f(x^*))\right\|^2\right] \\
&\overset{(19),(139)}{\leq} \frac{4L}{n}\sum_{i=1}^n \mathbf{E}\left[D_{f_i}(x_i^k, x^*)\right] + \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\bar{\xi}_i^k}(y^k) - \nabla f_i(x^*)\|^2\right] \\
&\overset{(63)}{\leq} 8L\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2\mathbf{E}[\sigma_k^2] + 4L^2\mathbf{E}[V_k]
\end{aligned}$$

and

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2\right] &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x_i^k)\|^2\right] \\
&\overset{(139)}{\leq} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x^*)\|^2\right] \\
&\overset{(136)}{\leq} \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_{\xi_i^k}(x^*)\|^2\right] + \frac{2}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i^k}(x^*) - \nabla f_i(x^*)\|^2\right] \\
&\overset{(86)}{\leq} \frac{4\mathcal{L}}{n}\sum_{i=1}^n \mathbf{E}\left[D_{f_i}(x_i^k, x^*)\right] + 2\sigma_*^2 \\
&\overset{(63)}{\leq} 8\mathcal{L}\mathbf{E}\left[f(x^k) - f(x^*)\right] + 4\mathcal{L}L\mathbf{E}[V_k] + 2\sigma_*^2. \qquad (117)
\end{aligned}$$

Finally, we use independence of $\xi_1^k, \ldots, \xi_n^k$ and derive

$$\begin{aligned}
\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k\right\|^2\right] &= \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_{\xi_i^k}(x_i^k)\right\|^2\right] \\
&\overset{(140),(139)}{=} \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n (\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x_i^k))\right\|^2\right] + \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)\right\|^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{\xi_i^k}(x_i^k) - \nabla f_i(x_i^k)\|^2\right] + \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)\right\|^2\right] \\
&\overset{(117),(68)}{\leq} 4\left(\frac{2\mathcal{L}}{n} + L\right)\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2L\left(\frac{2\mathcal{L}}{n} + L\right)\mathbf{E}[V_k] + \frac{2\sigma_*^2}{n}
\end{aligned}$$

which finishes the proof. □

Lemma G.11. *Let f_i be convex and L-smooth for all $i \in [n]$ and Assumption G.1 holds. Then for all $k \geq 0$*

$$\mathbf{E}\left[\sigma_{k+1}^2\right] \leq (1-q)\mathbf{E}\left[\sigma_k^2\right] + 2q\left(\frac{2\mathcal{L}}{r} + L\right)\mathbf{E}\left[f(x^k) - f(x^*)\right] + \frac{2q\sigma_*^2}{r} \qquad (118)$$

where $\sigma_k^2 \overset{def}{=} \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_{\bar{\xi}_i^k}(y^k) - \nabla f_i(x^*)\right\|^2$ and $\sigma_*^2 \overset{def}{=} \frac{1}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i}\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2$.

Proof. By definition of y^{k+1} we have

$$\begin{aligned}\mathbf{E}\left[\sigma_{k+1}^2 \mid x_1^k,\ldots,x_n^k\right] &= \frac{1-q}{n}\sum_{i=1}^n \|\nabla f_{\bar{\xi}_i^k}(y^k)-\nabla f_i(x^*)\|^2 \\ &\quad +\frac{q}{n}\sum_{i=1}^n \mathbf{E}_{\bar{\xi}_i^{k+1}}\left[\|\nabla f_{\bar{\xi}_i^{k+1}}(x^k)-\nabla f_i(x^*)\|^2\right] \\ &\overset{(139)}{=} (1-q)\sigma_k^2+\frac{q}{n}\sum_{i=1}^n \|\nabla f_i(x^k)-\nabla f_i(x^*)\|^2 \\ &\quad +\frac{q}{n}\sum_{i=1}^n \mathbf{E}_{\bar{\xi}_i^{k+1}}\left[\|\nabla f_{\bar{\xi}_i^{k+1}}(x^k)-\nabla f_i(x^k)\|^2\right].\end{aligned}$$

Next, we use independence of $\bar{\xi}_{i,1}^{k+1},\bar{\xi}_{i,2}^{k+1},\ldots,\bar{\xi}_{i,r}^{k+1}$ for all $i\in[n]$ and derive

$$\begin{aligned}\mathbf{E}\left[\sigma_{k+1}^2 \mid x_1^k,\ldots,x_n^k\right] &= (1-q)\sigma_k^2+\frac{q}{n}\sum_{i=1}^n \|\nabla f_i(x^k)-\nabla f_i(x^*)\|^2 \\ &\quad +\frac{q}{nr^2}\sum_{i=1}^n\sum_{j=1}^r \mathbf{E}_{\bar{\xi}_{i,j}^{k+1}}\left[\|\nabla f_{\bar{\xi}_{i,j}^{k+1}}(x^k)-\nabla f_i(x^k)\|^2\right] \\ &\overset{(19),(139)}{\leq} (1-q)\sigma_k^2+2Lq\left(f(x^k)-f(x^*)\right) \\ &\quad +\frac{q}{nr^2}\sum_{i=1}^n\sum_{j=1}^r \mathbf{E}_{\bar{\xi}_{i,j}^{k+1}}\left[\|\nabla f_{\bar{\xi}_{i,j}^{k+1}}(x^k)-\nabla f_i(x^*)\|^2\right] \\ &\overset{(136)}{\leq} (1-q)\sigma_k^2+2Lq\left(f(x^k)-f(x^*)\right) \\ &\quad +\frac{2q}{nr^2}\sum_{i=1}^n\sum_{j=1}^r \mathbf{E}_{\bar{\xi}_{i,j}^{k+1}}\left[\|\nabla f_{\bar{\xi}_{i,j}^{k+1}}(x^k)-\nabla f_{\bar{\xi}_{i,j}^{k+1}}(x^*)\|^2\right] \\ &\quad +\frac{2q}{nr^2}\sum_{i=1}^n\sum_{j=1}^r \mathbf{E}_{\bar{\xi}_{i,j}^{k+1}}\left[\|\nabla f_{\bar{\xi}_{i,j}^{k+1}}(x^*)-\nabla f_i(x^*)\|^2\right] \\ &\overset{(86)}{\leq} (1-q)\sigma_k^2+2q\left(\frac{2\mathcal{L}}{r}+L\right)\left(f(x^k)-f(x^*)\right)+\frac{2q\sigma_*^2}{r}.\end{aligned}$$

Taking the full mathematical expectation on both sides of previous inequality and using the tower property (140) we get the result. □

Using Corollary E.3 we obtain the following theorem.

Theorem G.10. *Assume that $f_i(x)$ is μ-strongly convex and L-smooth for every $i\in[n]$. Let Assumption G.1 holds. Then* `SS-Local-SGD` *satisfies Assumption E.1 with*

$$\widetilde{A}=4L,\quad \widehat{A}=4\mathcal{L},\quad \widetilde{B}=2,\quad \widehat{B}=0,\quad \widetilde{F}=4L^2,\quad \widehat{F}=4\mathcal{L}L,\quad \widetilde{D}_1=0,$$

$$\widehat{D}_1=2\sigma_*^2,\quad \sigma_*^2=\frac{1}{n}\sum_{i=1}^n \mathbf{E}_{\xi_i}\|\nabla f_{\xi_i}(x^*)-\nabla f_i(x^*)\|^2,\quad A'=2\left(\frac{2\mathcal{L}}{n}+L\right),\quad B'=0,\quad F'=2L\left(\frac{2\mathcal{L}}{n}+L\right),$$

$$D_1'=\frac{2\sigma_*^2}{n},\quad \sigma_k^2=\frac{1}{n}\sum_{i=1}^n\left\|\nabla f_{\bar{\xi}_i^k}(y^k)-\nabla f_i(x^*)\right\|^2,\quad \rho=q,\quad C=q\left(\frac{2\mathcal{L}}{r}+L\right),\quad G=0,\quad D_2=\frac{2q\sigma_*^2}{r},$$

$$H=\frac{128(1-p)(2+p)(2+q)\gamma^2}{3p^2q},\quad D_3=\frac{8(1-p)}{p^2}\left(2p\sigma_*^2+\frac{32(2+p)\sigma_*^2}{3r}\right)$$

under assumption that

$$\gamma \quad \leq \quad \min\left\{\frac{1}{4\left(\frac{2\mathcal{L}}{n}+L\right)}, \frac{p\sqrt{3}}{32\sqrt{2L(1-p)\left((2+p)L+p\mathcal{L}+\frac{(2+p)(2\mathcal{L}/r+L)}{(1-q)}\right)}}\right\}.$$

Moreover, for $\mu > 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K)-f(x^*)\right] \quad \leq \quad \left(1-\min\left\{\gamma\mu, \frac{q}{4}\right\}\right)^K \frac{\Phi^0}{\gamma} + 2\gamma\left(\frac{2\sigma_*^2}{n}+\gamma\frac{16L(1-p)}{p^2}\left(2p\sigma_*^2+\frac{32(2+p)\sigma_*^2}{3r}\right)\right)$$

and when $\mu = 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K)-f(x^*)\right] \quad \leq \quad \frac{\Phi^0}{\gamma K} + 2\gamma\left(\frac{2\sigma_*^2}{n}+\gamma\frac{16L(1-p)}{p^2}\left(2p\sigma_*^2+\frac{32(2+p)\sigma_*^2}{3r}\right)\right)$$

where $\Phi^0 = 2\|x^0-x^*\|^2 + \frac{512L(1-p)(2+p)(2+q)\gamma^3\mathbf{E}[\sigma_0^2]}{3p^2q}$.

The theorem above together with Lemma I.2 implies the following result.

Corollary G.19. *Let assumptions of Theorem G.10 hold with $\mu > 0$. Then for*

$$\begin{aligned}
\gamma_0 &= \min\left\{\frac{1}{4\left(\frac{2\mathcal{L}}{n}+L\right)}, \frac{p\sqrt{3}}{32\sqrt{2L(1-p)\left((2+p)L+p\mathcal{L}+\frac{(2+p)(2\mathcal{L}/r+L)}{(1-q)}\right)}}\right\},\\
\widetilde{\Phi}^0 &= 2\|x^0-x^*\|^2 + \frac{512L(1-p)(2+p)(2+q)\gamma_0^3\mathbf{E}[\sigma_0^2]}{p^2q}, \quad q=p,\\
\gamma &= \min\left\{\gamma_0, \frac{\ln\left(\max\left\{2, \min\left\{n\widetilde{\Phi}^0\mu^2K^2/4\sigma_*^2, p\widetilde{\Phi}^0\mu^3K^3/64L(1-p)(1+32(2+p)/3)\sigma_*^2\right\}\right\}\right)}{\mu K}\right\}, \quad r=\left\lceil\frac{1}{p}\right\rceil,
\end{aligned}$$

for all K such that

$$\begin{aligned}
&\textit{either} \quad \frac{\ln\left(\max\left\{2, \min\left\{n\widetilde{\Phi}^0\mu^2K^2/4\sigma_*^2, p\widetilde{\Phi}^0\mu^3K^3/64L(1-p)(1+32(2+p)/3)\sigma_*^2\right\}\right\}\right)}{K} \leq p\\
&\textit{or} \quad \gamma_0 \leq \frac{\ln\left(\max\left\{2, \min\left\{n\widetilde{\Phi}^0\mu^2K^2/4\sigma_*^2, p\widetilde{\Phi}^0\mu^3K^3/64L(1-p)(1+32(2+p)/3)\sigma_*^2\right\}\right\}\right)}{\mu K}
\end{aligned}$$

we have that $\mathbf{E}\left[f(\overline{x}^K)-f(x^*)\right]$ is of the order

$$\widetilde{\mathcal{O}}\left(\frac{\widetilde{\Phi}^0}{\gamma_0}\exp\left(-\min\left\{\frac{1}{p}, \gamma_0\mu\right\}K\right) + \frac{\sigma_*^2}{n\mu K} + \frac{L(1-p)\sigma_*^2}{p\mu^2K^2}\right).$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K)-f(x^*)\right] \leq \varepsilon$ in this case `SS-Local-SGD` *requires*

$$\widetilde{\mathcal{O}}\left(\frac{L}{p\mu}+\frac{\mathcal{L}}{n\mu}+\frac{\sqrt{\mathcal{L}L(1-p)}}{\sqrt{p}\mu}+\frac{\sigma_*^2}{n\mu\varepsilon}+\sqrt{\frac{L(1-p)\sigma_*^2}{p\mu^2\varepsilon}}\right)$$

iterations/oracle calls per node (in expectation) and $1/p$ times less communication rounds.

Combining Theorem G.10 and Lemma I.3 we derive the following result for the convergence of `SS-Local-SGD` in the case when $\mu = 0$.

Corollary G.20. *Let assumptions of Theorem G.10 hold with $\mu = 0$. Then for $q = p$, $r = \lceil 1/p \rceil$ and*

$$\begin{aligned}
\gamma_0 &= \min\left\{\frac{1}{4\left(\frac{2\mathcal{L}}{n}+L\right)}, \frac{p\sqrt{3}}{32\sqrt{2L(1-p)\left((2+p)L+p\mathcal{L}+\frac{(2+p)(2\mathcal{L}/r+L)}{(1-q)}\right)}}\right\},\\
\gamma &= \min\left\{\gamma_0, \sqrt[3]{\frac{p^3R_0^2}{256L(1-p)(2+p)^2\mathbf{E}[\sigma_0^2]}}, \sqrt{\frac{nR_0^2}{2\sigma_*^2K}}, \sqrt[3]{\frac{pR_0^2}{32L(1-p)\left(1+32(2+p)/3\right)\sigma_*^2K}}\right\},
\end{aligned}$$

where $R_0 = \|x^0 - x^*\|$*, we have that* $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right]$ *is of the order*

$$\mathcal{O}\left(\frac{\left(L + {}^{p\mathcal{L}}/_{n} + \sqrt{p(1-p)\mathcal{L}L}\right)R_0^2 + \sqrt[3]{L(1-p)\mathbf{E}[\sigma_0^2]R_0^4}}{pK} + \sqrt{\frac{\sigma_*^2 R_0^2}{nK}} + \frac{\sqrt[3]{LR_0^4(1-p)\sigma_*^2}}{p^{1/3}K^{2/3}}\right).$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ *in this case* `SS-Local-SGD` *requires*

$$\mathcal{O}\left(\frac{\left(L + {}^{p\mathcal{L}}/_{n} + \sqrt{p(1-p)\mathcal{L}L}\right)R_0^2 + \sqrt[3]{L(1-p)\mathbf{E}[\sigma_0^2]R_0^4}}{p\varepsilon} + \frac{\sigma_*^2 R_0^2}{n\varepsilon^2} + \frac{R_0^2\sqrt{L(1-p)\sigma_*^2}}{p^{1/2}\varepsilon^{3/2}}\right)$$

iterations/oracle calls per node (in expectation) and ${}^1/_p$ *times less communication rounds.*

Remark G.4. *To get the rate from Tbl. 4 it remains to apply the following inequality:*

$$\begin{aligned}
\mathbf{E}[\sigma_0^2] &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}_{\bar{\xi}_i^0}\left[\|\nabla f_{\bar{\xi}_i^0}(x^0) - \nabla f_i(x^*)\|^2\right] \\
&\overset{(139)}{=} \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^0) - \nabla f_i(x^*)\|^2 + \frac{1}{n}\sum_{i=1}^n \mathbf{E}_{\bar{\xi}_i^0}\left[\|\nabla f_{\bar{\xi}_i^0}(x^0) - \nabla f_i(x^0)\|^2\right] \\
&\overset{(19)}{\leq} 2L(f(x^0) - f(x^*)) + \frac{1}{nr^2}\sum_{i=1}^n\sum_{j=1}^r \mathbf{E}_{\bar{\xi}_{i,j}^0}\left[\|\nabla f_{\bar{\xi}_{i,j}^0}(x^0) - \nabla f_i(x^0)\|^2\right] \\
&\overset{(139)}{\leq} 2L(f(x^0) - f(x^*)) + \frac{1}{nr}\sum_{i=1}^n \mathbf{E}_{\xi_i}\left[\|\nabla f_{\xi_i}(x^0) - \nabla f_i(x^*)\|^2\right] \\
&\overset{(136)}{\leq} 2L(f(x^0) - f(x^*)) + \frac{2}{nr}\sum_{i=1}^n \mathbf{E}_{\xi_i}\left[\|\nabla f_{\xi_i}(x^0) - \nabla f_{\xi_i}(x^*)\|^2\right] \\
&\quad + \frac{2}{nr}\sum_{i=1}^n \mathbf{E}_{\xi_i}\left[\|\nabla f_{\xi_i}(x^*) - \nabla f_i(x^*)\|^2\right] \\
&\overset{r=\lceil 1/p\rceil,(86)}{\leq} 2\left(L + 2p\mathcal{L}\right)\left(f(x^0) - f(x^*)\right) + 2p\sigma_*^2.
\end{aligned}$$

G.5 `S*-Local-SGD*`

In this section we present doubly idealized algorithm for solving problem (1)+(3). Specifically, we choose b_i^k to the optimal shift $\nabla f_i(x^*)$ as per Case II, while a_i^k is selected as `SGD-star` gradient estimator (Gorbunov et al., 2020a), i.e.,

$$a_i^k = \nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(x^*) + \nabla f_i(x^*), \qquad b_i^k = \nabla f_i(x^*).$$

Note that now a_i^k serves as an ambitious target for the local variance reduced estimators, while b_i^k serves as an ambitious goal for the local shift. The resulting instance of (4) is presented as Algorithm 5 and called Star-Shifted `Local-SGD-star` (`S*-Local-SGD*`).

Let us next provide the details on the convergence rate. In order to do so, let us identify the parameters of Assumption 4.1.

Lemma G.12. *Let* f_i *be convex and* L*-smooth and* $f_{i,j}$ *be convex and* $\max L_{ij}$*-smooth for all* $i \in [n]$*,* $j \in [m]$*.*

Algorithm 5 `S*-Local-SGD*`

Require: learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$, communication period $\tau \geq 1$

for $k = 0, 1, \ldots$ **do**
  for $i = 1, \ldots, n$ in parallel **do**
    Set $g_i^k = \nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(x_*)$ where $1 \leq j_i \leq m$ is sampled independently from all nodes
    if $k+1 \mod \tau = 0$ **then**
      $x_i^{k+1} = x^{k+1} = \frac{1}{n}\sum_{i=1}^n \left(x_i^k - \gamma g_i^k\right)$ ▷ averaging
    else
      $x_i^{k+1} = x_i^k - \gamma g_i^k$ ▷ local update
    end if
  end for
end for

Then for all $k \geq 0$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}_k\left[g_i^k\right] = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k), \tag{119}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\bar{g}_i^k\|^2\right] \leq 4L\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2L^2\mathbf{E}[V_k], \tag{120}$$

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2\right] \leq 4\max L_{ij}\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2L\max L_{ij}\mathbf{E}[V_k], \tag{121}$$

$$\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k\right\|^2\right] \leq 4\left(\frac{\max L_{ij}}{n} + L\right)\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2L\left(\frac{\max L_{ij}}{n} + L\right)\mathbf{E}[V_k]. \tag{122}$$

Proof. First of all,

$$\frac{1}{n}\sum_{i=1}^n \mathbf{E}_k\left[g_i^k\right] = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \left(\nabla f_{i,j}(x_i^k) - \nabla f_{i,j}(x^*)\right) = \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)$$

and, in particular, $\bar{g}_i^k = \mathbf{E}_k\left[g_i^k\right] = \nabla f_i(x_i^k) - \nabla f_i(x^*)$. Using this we derive

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\bar{g}_i^k\|^2\right] &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_i(x_i^k) - \nabla f_i(x^*)\|^2\right] \\ &\overset{(19)}{\leq} \frac{2L}{n}\sum_{i=1}^n \mathbf{E}\left[D_{f_i}(x_i^k, x^*)\right] \overset{(63)}{\leq} 4L\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2L^2\mathbf{E}[V_k]\end{aligned}$$

and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k - \bar{g}_i^k\|^2\right] &\overset{(139)}{\leq} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k\|^2\right] \\ &= \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{i,j}(x_i^k) - \nabla f_{i,j}(x^*)\|^2 \\ &\overset{(19)}{\leq} \frac{2\max L_{ij}}{n}\sum_{i=1}^n \mathbf{E}\left[D_{f_i}(x_i^k, x^*)\right] \\ &\overset{(63)}{\leq} 4\max L_{ij}\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2L\max L_{ij}\mathbf{E}[V_k].\end{aligned} \tag{123}$$

Finally, due to the independence of $j_1, j_2, \ldots, j_n$ we have

$$\begin{aligned}
\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k\right\|^2\right] &\overset{(139),(140)}{=} \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(x_*) - (\nabla f_i(x_i^k) - \nabla f_i(x^*))\right)\right\|^2\right] \\
&\quad + \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \left(\nabla f_i(x_i^k) - \nabla f_i(x^*)\right)\right\|^2\right] \\
&= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(x_*) - (\nabla f_i(x_i^k) - \nabla f_i(x^*))\|^2\right] \\
&\quad + \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)\right\|^2\right] \\
&\overset{(139)}{\leq} \frac{1}{n^2 m}\sum_{i=1}^n\sum_{j=1}^m \|\nabla f_{i,j}(x_i^k) - \nabla f_{i,j}(x^*)\|^2 + \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)\right\|^2\right] \\
&\overset{(123),(63)}{\leq} 4\left(\frac{\max L_{ij}}{n} + L\right)\mathbf{E}\left[f(x^k) - f(x^*)\right] + 2L\left(\frac{\max L_{ij}}{n} + L\right)\mathbf{E}[V_k].
\end{aligned}$$

□

Using Corollary E.1 we obtain the following theorem.

Theorem G.11. *Assume that $f_i(x)$ is μ-strongly convex and L-smooth and $f_{i,j}$ is convex and $\max L_{ij}$-smooth for every $i \in [n]$, $j \in [m]$. Then* `S*-Local-SGD*` *satisfies Assumption E.1 with*

$$\widetilde{A} = 2L, \quad \widehat{A} = 2\max L_{ij}, \quad \widetilde{B} = \widehat{B} = 0, \quad \widetilde{F} = 2L^2, \quad \widehat{F} = 2L\max L_{ij}, \quad \widetilde{D}_1 = \widehat{D}_1 = 0,$$
$$A' = 2\left(\frac{\max L_{ij}}{n} + L\right), \quad B' = 0, \quad F' = 2L\left(\frac{\max L_{ij}}{n} + L\right),$$
$$D_1' = 0, \quad \sigma_k^2 \equiv 0, \quad \rho = 1, \quad C = 0, \quad G = 0, \quad D_2 = 0, \quad H = 0, \quad D_3 = 0$$

under assumption that

$$\gamma \leq \min\left\{\frac{1}{4\left(\frac{\max L_{ij}}{n} + L\right)}, \frac{1}{8\sqrt{eL(\tau-1)\left(L(\tau-1) + \max L_{ij}\right)}}\right\}.$$

Moreover, for $\mu > 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq (1-\gamma\mu)^K \frac{2\|x^0 - x^*\|^2}{\gamma}$$

and when $\mu = 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{2\|x^0 - x^*\|^2}{\gamma K}.$$

The theorem above together with Lemma I.2 implies the following result.

Corollary G.21. *Let assumptions of Theorem G.11 hold with $\mu > 0$. Then for*

$$\gamma = \min\left\{\frac{1}{4\left(\frac{\max L_{ij}}{n} + L\right)}, \frac{1}{8\sqrt{eL(\tau-1)\left(L(\tau-1) + \max L_{ij}\right)}}\right\}$$

and for all $K \geq 1$ we have $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right]$ of order

$$\mathcal{O}\left(\left(L\tau + {\max L_{ij}}/{n} + \sqrt{(\tau-1)L\max L_{ij}}\right)\|x^0 - x^*\|^2 \exp\left(-\frac{\mu}{L\tau + {\max L_{ij}}/{n} + \sqrt{(\tau-1)L\max L_{ij}}}K\right)\right).$$

That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ *in this case* `S*-Local-SGD*` *requires*

$$\mathcal{O}\left(\left(\frac{L\tau}{\mu} + \frac{\max L_{ij}}{n\mu} + \frac{\sqrt{(\tau-1)L\max L_{ij}}}{\mu}\right)\log\frac{\left(L\tau + {}^{\max L_{ij}}/_{n} + \sqrt{(\tau-1)L\max L_{ij}}\right)\|x^0 - x^*\|^2}{\varepsilon}\right)$$

iterations/oracle calls per node and τ *times less communication rounds.*

Next, we derive the following result for the convergence of `S*-Local-SGD*` in the case when $\mu = 0$.

Corollary G.22. *Let assumptions of Theorem G.11 hold with* $\mu = 0$*. Then for*

$$\gamma \quad = \quad \min\left\{\frac{1}{4\left(\frac{\max L_{ij}}{n} + L\right)}, \frac{1}{8\sqrt{eL(\tau-1)\left(L(\tau-1) + \max L_{ij}\right)}}\right\},$$

we have that $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right]$ *is of the order*

$$\mathcal{O}\left(\frac{\left(L\tau + {}^{\max L_{ij}}/_{n} + \sqrt{(\tau-1)L\max L_{ij}}\right)R_0^2}{K}\right),$$

where $R_0 = \|x^0 - x^*\|$*. That is, to achieve* $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ *in this case* `S*-Local-SGD*` *requires*

$$\mathcal{O}\left(\frac{\left(L\tau + {}^{\max L_{ij}}/_{n} + \sqrt{(\tau-1)L\max L_{ij}}\right)R_0^2}{\varepsilon}\right)$$

iterations/oracle calls per node and τ *times less communication rounds.*

G.6 `S-Local-SVRG`

Algorithm 6 Shifted Local `SVRG` (`S-Local-SVRG`) for minimizing local finite sums

Require: learning rate $\gamma > 0$, initial vector $x^0 \in \mathbb{R}^d$, probability of communication $p \in (0,1]$, probability of local full gradient computation $q \in (0,1]$, initialization $y^0 = x^0$

1: **for** $k = 0, 1, \ldots$ **do**
2: **for** $i = 1, \ldots, n$ in parallel **do**
3: Choose j_i uniformly at random from $[m]$
4: $g_i^k = \nabla f_{i,j_i}(x_i^k) - \nabla f_{i,j_i}(y^k) + \nabla f(y^k)$
5: $x_i^{k+1} = \begin{cases} x^{k+1}, & \text{w.p. } p, \\ x_i^k - \gamma g_i^k, & \text{w.p. } 1-p, \end{cases}$ where $x^{k+1} = \frac{1}{n}\sum_{i=1}^n (x_i^k - \gamma g_i^k)$
6: $y^{k+1} = \begin{cases} x^k, & \text{w.p. } q, \\ y^k, & \text{w.p. } 1-q \end{cases}$
7: **end for**
8: **end for**

In this section we are interested in problem (1)+(3). To solve this problem we propose a new method called Shifted `Local-SVRG` (`S-Local-SVRG`, see Algorithm 6).

We note that our analysis works even when updates in lines 5,6 are not independent. Moreover, in order for `S-Local-SVRG` to be efficient, we shall require $q \leq p$.

Remark G.5. *Unlike all other special cases, the rate of* `S-Local-SVRG` *can not be directly obtained from the theory of the local stochastic solver described in Section 4. Specifically, we construct the sequence* l_i^k *using* y^k *in contrast to* x_i^k *used in Section 4. While we could construct* l_i^k *from the local iterate sequences, setting it as the virtual iterates yields a tighter rate. We remark that such a choice is rather poor in general; we can implement it efficiently thanks to the specific structure of* `S-Local-SVRG`*.*

Lemma G.13. *Let f_i be convex and L-smooth and $f_{i,j}$ be convex and $\max L_{ij}$-smooth for all $i \in [n]$, $j \in [m]$. Then for all $k \geq 0$*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_k\left[g_i^k\right] = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^k), \tag{124}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\|\bar{g}_i^k\|^2\right] \leq 8L\mathbf{E}\left[f(x^k)-f(x^*)\right]+2\mathbf{E}[\sigma_k^2]+4L^2\mathbf{E}[V_k], \tag{125}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\|g_i^k-\bar{g}_i^k\|^2\right] \leq 8\max L_{ij}\mathbf{E}\left[f(x^k)-f(x^*)\right]+2\mathbf{E}[\sigma_k^2]+4L\max L_{ij}\mathbf{E}[V_k], \tag{126}$$

$$\begin{aligned}\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}g_i^k\right\|^2\right] \leq\ & 4\left(\frac{2\max L_{ij}}{n}+L\right)\mathbf{E}\left[f(x^k)-f(x^*)\right]+\frac{2}{n}\mathbf{E}[\sigma_k^2]\\ & +2L\left(\frac{2\max L_{ij}}{n}+L\right)\mathbf{E}[V_k],\end{aligned} \tag{127}$$

where $\sigma_k^2 \overset{def}{=} \frac{1}{nm}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m}\left\|\nabla f_{i,j}(y^k)-\nabla f_{i,j}(x^*)\right\|^2+\frac{1}{n}\sum\limits_{i=1}^{n}\left\|\nabla f_i(y^k)-\nabla f_i(x^*)\right\|^2$.

Proof. First of all, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_k\left[g_i^k\right] &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}_k\left[\nabla f_{i,j^k}(x_i^k)-\nabla f_{i,j_i}(y^k)+\nabla f(y^k)\right]\\ &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\nabla f_{i,j}(x_i^k)-\nabla f_{i,j}(y^k)+\nabla f(y^k)\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(x_i^k)\end{aligned}$$

and, in particular, $\bar{g}_i^k = \mathbf{E}_k[g_i^k] = \nabla f_i(x_i^k)-\nabla f_i(y^k)+\nabla f(y^k)$. Using this we get

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[\|\bar{g}_i^k\|^2\right] &\overset{(136)}{\leq} \frac{2}{n}\sum_{i=1}^{n}\mathbf{E}\left[\|\nabla f_i(x_i^k)-\nabla f_i(x^*)\|^2\right]\\ &\quad+\frac{2}{n}\sum_{i=1}^{n}\mathbf{E}\left[\|\nabla f_i(y^k)-\nabla f_i(x^*)-(\nabla f(y^k)-\nabla f(x^*))\|^2\right]\\ &\overset{(19),(139)}{\leq} \frac{4L}{n}\sum_{i=1}^{n}\mathbf{E}\left[D_{f_i}(x_i^k,x^*)\right]+\frac{2}{n}\sum_{i=1}^{n}\mathbf{E}\left[\|\nabla f_i(y^k)-\nabla f_i(x^*)\|^2\right]\\ &\overset{(63)}{\leq} 8L\mathbf{E}\left[f(x^k)-f(x^*)\right]+2\mathbf{E}[\sigma_k^2]+4L^2\mathbf{E}[V_k]\end{aligned}$$

and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|g_i^k-\bar{g}_i^k\|^2\right] &= \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{i,j_i}(x_i^k)-\nabla f_{i,j_i}(y^k)-(\nabla f_i(x_i^k)-\nabla f_i(y^k))\|^2\right]\\ &\overset{(139)}{\leq} \frac{1}{n}\sum_{i=1}^n \mathbf{E}\left[\|\nabla f_{i,j_i}(x_i^k)-\nabla f_{i,j_i}(y^k)\|^2\right]\\ &\overset{(136)}{\leq} \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \mathbf{E}\left[\|\nabla f_{i,j}(x_i^k)-\nabla f_{i,j}(x^*)\|^2\right]\\ &\quad +\frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m \mathbf{E}\left[\|\nabla f_{i,j}(y^k)-\nabla f_{i,j}(x^*)\|^2\right]\\ &\overset{(19)}{\leq} \frac{4\max L_{ij}}{n}\sum_{i=1}^n \mathbf{E}\left[D_{f_i}(x_i^k,x^*)\right]+2\mathbf{E}[\sigma_k^2]\\ &\overset{(63)}{\leq} 8\max L_{ij}\mathbf{E}\left[f(x^k)-f(x^*)\right]+2\mathbf{E}[\sigma_k^2]+4L\max L_{ij}\mathbf{E}[V_k]. \end{aligned} \tag{128}$$

Finally, using independence of $j_1, j_2, \ldots, j_n$ we derive

$$\begin{aligned}\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i^k\right\|^2\right] &\overset{(139),(124)}{=} \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)\right\|^2\right]\\ &\quad +\mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n (\nabla f_{i,j_i}(x_i^k)-\nabla f_{i,j_i}(y^k)-(\nabla f_i(x_i^k)-\nabla f_i(y^k)))\right\|^2\right]\\ &= \mathbf{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i^k)\right\|^2\right]\\ &\quad +\frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\left[\|(\nabla f_{i,j_i}(x_i^k)-\nabla f_{i,j_i}(y^k)-(\nabla f_i(x_i^k)-\nabla f_i(y^k)))\|^2\right]\\ &\overset{(68),(128)}{\leq} 4\left(\frac{2\max L_{ij}}{n}+L\right)\mathbf{E}\left[f(x^k)-f(x^*)\right]+\frac{2}{n}\mathbf{E}[\sigma_k^2]+2L\left(\frac{2\max L_{ij}}{n}+L\right)\mathbf{E}[V_k]. \end{aligned}$$

□

Lemma G.14. *Let f_i be convex and L-smooth and $f_{i,j}$ be convex and $\max L_{ij}$-smooth for all $i\in[n]$, $j\in[m]$. Then for all $k\geq 0$*

$$\mathbf{E}\left[\sigma_{k+1}^2\right] \leq (1-q)\,\mathbf{E}\left[\sigma_k^2\right]+2(L+\max L_{ij})q\mathbf{E}\left[f(x^k)-f(x^*)\right] \tag{129}$$

where $\sigma_k^2 \overset{def}{=} \frac{1}{nm}\sum\limits_{i=1}^n\sum\limits_{j=1}^m \left\|\nabla f_{i,j}(y^k)-\nabla f_{i,j}(x^*)\right\|^2+\frac{1}{n}\sum\limits_{i=1}^n \left\|\nabla f_i(y^k)-\nabla f_i(x^*)\right\|^2$.

Proof. First of all, we introduce new notations:

$$\sigma_{k,1}^2 \overset{\text{def}}{=} \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \left\|\nabla f_{i,j}(y^k)-\nabla f_{i,j}(x^*)\right\|^2, \quad \sigma_{k,2}^2=\frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(y^k)-\nabla f_i(x^*)\right\|^2.$$

Secondly, by definition of y^{k+1} we have

$$\begin{aligned}\mathbf{E}\left[\sigma_{k+1,1}^2 \mid x_1^k,\ldots,x_n^k\right] &= \frac{1-q}{nm}\sum_{i=1}^n\sum_{j=1}^m \left\|\nabla f_{i,j}(y^k)-\nabla f_{i,j}(x^*)\right\|^2+\frac{q}{nm}\sum_{i=1}^n\sum_{j=1}^m \left\|\nabla f_{i,j}(x^k)-\nabla f_{i,j}(x^*)\right\|^2\\ &\overset{(19)}{\leq} (1-q)\sigma_{k,1}^2+2q\max L_{ij}(f(x^k)-f(x^*)), \end{aligned}$$

hence

$$\mathbf{E}\left[\sigma_{k+1,1}^2\right] \leq (1-q)\mathbf{E}\left[\sigma_{k,1}^2\right] + 2q\max L_{ij}\mathbf{E}\left[f(x^k) - f(x^*)\right]. \tag{130}$$

Next, the definition of y^{k+1} implies

$$\begin{aligned}\mathbf{E}\left[\sigma_{k+1,2}^2 \mid x_1^k, \ldots, x_n^k\right] &= \frac{1-q}{n}\sum_{i=1}^n \|\nabla f_i(y^k) - \nabla f_i(x^*)\|^2 + \frac{q}{n}\sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 \\ &\overset{(19)}{\leq} (1-q)\sigma_k^2 + 2Lq(f(x^k) - f(x^*)),\end{aligned}$$

hence

$$\mathbf{E}\left[\sigma_{k+1,2}^2\right] \leq (1-q)\mathbf{E}\left[\sigma_{k,2}^2\right] + 2Lq\mathbf{E}\left[f(x^k) - f(x^*)\right]. \tag{131}$$

Finally, we combine obtained inequalities and get

$$\begin{aligned}\mathbf{E}\left[\sigma_{k+1}\right] &= \mathbf{E}\left[\sigma_{k+1,1}^2\right] + \mathbf{E}\left[\sigma_{k+1,2}^2\right] \\ &\overset{(130),(131)}{\leq} (1-q)\left(\mathbf{E}\left[\sigma_{k,1}^2\right] + \mathbf{E}\left[\sigma_{k,2}^2\right]\right) + 2(L + \max L_{ij})q\mathbf{E}\left[f(x^k) - f(x^*)\right] \\ &= (1-q)\,\mathbf{E}\left[\sigma_k^2\right] + 2(L + \max L_{ij})q\mathbf{E}\left[f(x^k) - f(x^*)\right],\end{aligned}$$

which concludes the proof. □

Using Corollary E.3 we obtain the following theorem.

Theorem G.12. *Assume that f_i is μ-strongly convex and L-smooth and $f_{i,j}$ is convex and $\max L_{ij}$-smooth for all $i \in [n]$, $j \in [m]$. Then* `S-Local-SVRG` *satisfies Assumption E.1 with*

$$\widetilde{A} = 4L, \quad \widehat{A} = 4\max L_{ij}, \quad \widetilde{B} = \widehat{B} = 2, \quad \widetilde{F} = 4L^2, \quad \widehat{F} = 4L\max L_{ij} \quad \widetilde{D}_1 = \widehat{D}_1 = 0,$$

$$A' = \frac{4\max L_{ij}}{n} + 2L, \quad B' = \frac{2}{n}, \quad F' = 2L\left(\frac{2\max L_{ij}}{n} + L\right), \quad D_1' = 0,$$

$$\sigma_k^2 = \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \left\|\nabla f_{i,j}(y^k) - \nabla f_{i,j}(x^*)\right\|^2 + \frac{1}{n}\sum_{i=1}^n \left\|\nabla f_i(y^k) - \nabla f_i(x^*)\right\|^2,$$

$$\rho = q, \quad C = (L + \max L_{ij})q, \quad G = 0, \quad D_2 = 0, \quad H = \frac{256(1-p^2)(2+q)\gamma^2}{3p^2q}, \quad D_3 = 0$$

under assumption that

$$\gamma \leq \min\left\{\frac{1}{56\max L_{ij}/3n + 4L + 32L/3n}, \frac{p\sqrt{3}}{32\sqrt{2L(1-p)\left(L(2+p) + p\max L_{ij} + 4(L+\max L_{ij})(1+p)/(1-q)\right)}}\right\}.$$

Moreover, for $\mu > 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \left(1 - \min\left\{\gamma\mu, \frac{q}{4}\right\}\right)^K \frac{2\|x^0 - x^*\|^2 + \frac{16\gamma^2\sigma_0^2}{nq} + \frac{1024L(1-p^2)(2+q)\gamma^3\sigma_0^2}{3p^2q}}{\gamma}$$

and when $\mu = 0$ we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \frac{2\|x^0 - x^*\|^2 + \frac{16\gamma^2\sigma_0^2}{nq} + \frac{1024L(1-p^2)(2+q)\gamma^3\sigma_0^2}{3p^2q}}{\gamma K}.$$

The theorem above together with Lemma I.2 implies the following result.

Corollary G.23. *Let assumptions of Theorem G.12 hold with $\mu > 0$. Then for $q = 1/m$, $m \geq 1/p$,*

$$\gamma = \min\left\{\frac{1}{56\max L_{ij}/3n + 4L + 32L/3n}, \frac{p\sqrt{3}}{32\sqrt{2L(1-p)\left(L(2+p) + p\max L_{ij} + 4(L+\max L_{ij})(1+p)/(1-q)\right)}}\right\}$$

and for all $K \geq 1$ we have $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right]$ of order

$$\mathcal{O}\left(\left(\frac{L}{p} + \frac{\max L_{ij}}{n} + \frac{\sqrt{(1-p)L\max L_{ij}}}{p}\right)\Phi^0 \exp\left(-\min\left\{\frac{\mu}{\frac{L}{p} + \frac{\max L_{ij}}{n} + \frac{\sqrt{(1-p)L\max L_{ij}}}{p}}, \frac{1}{m}\right\}K\right)\right),$$

where $\Phi^0 = 2\|x^0 - x^*\|^2 + \frac{16\gamma^2\sigma_0^2}{nq} + \frac{1024L(1-p^2)(2+q)\gamma^3\sigma_0^2}{3p^2q}$. That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ in this case `S-Local-SVRG` *requires*

$$K = \mathcal{O}\left(\left(m + \frac{L}{p\mu} + \frac{\max L_{ij}}{n\mu} + \frac{\sqrt{(1-p)L\max L_{ij}}}{p\mu}\right)\log\frac{\left(\frac{L}{p} + \frac{\max L_{ij}}{n} + \frac{\sqrt{(1-p)L\max L_{ij}}}{p}\right)\Phi^0}{\varepsilon}\right)$$

iterations/oracle calls per node (in expectation) and $1/p$ times less communication rounds.

That is, `S-Local-SVRG` is the first implementable linearly converging stochastic method with local updates with a convergence guarantee in terms of the number of communications that is not worse than that of `GD` even in the arbitrary heterogeneous data regime.

Next, we derive the following result for the convergence of `S-Local-SVRG` in the case when $\mu = 0$.

Corollary G.24. *Let assumptions of Theorem G.12 hold with $\mu = 0$. Then for $q = 1/m$, $m \geq 1/p$ and*

$$\begin{aligned}\gamma_0 &= \min\left\{\frac{1}{56\max L_{ij}/3n + 4L + 32L/3n}, \frac{p\sqrt{3}}{32\sqrt{2L(1-p)\left(L(2+p) + p\max L_{ij} + 4(L+\max L_{ij})(1+p)/(1-q)\right)}}\right\},\\ \gamma &= \min\left\{\gamma_0, \sqrt{\frac{nR_0^2}{8m\sigma_0^2}}, \sqrt[3]{\frac{3p^2R_0^2}{512L(1-p^2)(2m+1)\sigma_0^2}}\right\}\end{aligned}$$

we have

$$\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] = \mathcal{O}\left(\frac{\left(L + p\max L_{ij}/n + \sqrt{(1-p)L\max L_{ij}}\right)R_0^2}{pK} + \frac{\sqrt{m\sigma_0^2R_0^2}}{\sqrt{n}K} + \frac{\sqrt[3]{Lm\sigma_0^2R_0^4}}{p^{2/3}K}\right),$$

where $R_0 = \|x^0 - x^*\|$. That is, to achieve $\mathbf{E}\left[f(\overline{x}^K) - f(x^*)\right] \leq \varepsilon$ in this case `S-Local-SVRG` *requires*

$$K = \mathcal{O}\left(\frac{\left(L + p\max L_{ij}/n + \sqrt{(1-p)L\max L_{ij}}\right)R_0^2}{p\varepsilon} + \frac{\sqrt{m\sigma_0^2R_0^2}}{\sqrt{n}\varepsilon} + \frac{\sqrt[3]{Lm\sigma_0^2R_0^4}}{p^{2/3}\varepsilon}\right)$$

iterations/oracle calls per node (in expectation) and $1/p$ times less communication rounds.

Remark G.6. *To get the rate from Tbl. 4 it remains to apply the following inequality:*

$$\begin{aligned}\sigma_0^2 &= \frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m\left\|\nabla f_{i,j}(x^0) - \nabla f_{i,j}(x^*)\right\|^2 + \frac{1}{n}\sum_{i=1}^n\left\|\nabla f_i(x^0) - \nabla f_i(x^*)\right\|^2\\ &\overset{(6)}{\leq} 2\left(\max L_{ij}^2 + L^2\right)\|x^0 - x^*\|^2.\end{aligned}$$

H Basic Facts

For all $a, b, x_1, \ldots, x_n \in \mathbb{R}^d$, $\beta > 0$ and $p \in (0, 1]$ the following inequalities hold

$$\langle a, b\rangle \leq \frac{\|a\|^2}{2\beta} + \frac{\beta\|b\|^2}{2}, \tag{132}$$

$$\langle a - b, a + b\rangle = \|a\|^2 - \|b\|^2, \tag{133}$$

$$\frac{1}{2}\|a\|^2 - \|b\|^2 \leq \|a + b\|^2, \tag{134}$$

$$\|a + b\|^2 \leq (1 + \beta)\|a\|^2 + (1 + 1/\beta)\|b\|^2, \tag{135}$$

$$\left\|\sum_{i=1}^n x_n\right\|^2 \leq n\sum_{i=1}^n \|x_i\|^2, \tag{136}$$

$$\left(1 - \frac{p}{2}\right)^{-1} \leq 1 + p, \tag{137}$$

$$\left(1 + \frac{p}{2}\right)(1 - p) \leq 1 - \frac{p}{2}. \tag{138}$$

Variance decomposition. For a random vector $\xi \in \mathbb{R}^d$ and any deterministic vector $x \in \mathbb{R}^d$, the variance of ξ can be decomposed as

$$\mathbf{E}\left[\|\xi - \mathbf{E}[\xi]\|^2\right] = \mathbf{E}\left[\|\xi - x\|^2\right] - \|\mathbf{E}[\xi] - x\|^2 \tag{139}$$

Tower property of mathematical expectation. For random variables $\xi, \eta \in \mathbb{R}^d$ we have

$$\mathbf{E}[\xi] = \mathbf{E}\left[\mathbf{E}[\xi \mid \eta]\right] \tag{140}$$

under assumption that all expectations in the expression above are well-defined.

I Technical Lemmas

We now present a key technical lemma enabling our analysis. This is a refined version of Lemma 14 from (Stich and Karimireddy, 2019).

Lemma I.1 (see also Lemma 14 from (Stich and Karimireddy, 2019)). *For any τ random vectors $\xi_1, \ldots, \xi_\tau \in \mathbb{R}^d$ such that for all $t = 2, \ldots, \tau$ random vector ξ_t depends on $\xi_1, \ldots, \xi_{t-1}$ and does not depend on $\xi_{t+1}, \ldots, \xi_\tau$ the following inequality holds*

$$\mathbf{E}\left[\left\|\sum_{t=1}^{\tau}\xi_t\right\|^2\right] \leq e\tau\sum_{t=1}^{\tau}\mathbf{E}\left[\left\|\mathbf{E}_t[\xi_t]\right\|^2\right] + e\sum_{t=1}^{\tau}\mathbf{E}\left[\left\|\xi_t - \mathbf{E}_t[\xi_t]\right\|^2\right], \tag{141}$$

where $\mathbf{E}_t[\cdot]$ denotes the conditional expectation $\mathbf{E}[\cdot \mid \xi_{t-1}, \ldots, \xi_1]$.

Proof. First of all, if $\tau = 1$ then (141) immediately follows from variance decompostion (139). Otherwise ($\tau > 1$) for all $l = 1, \ldots, \tau$ we have

$$\begin{aligned}\mathbf{E}_l\left[\left\|\sum_{t=1}^{l}\xi_t\right\|^2\right] &\overset{(139)}{=} \left\|\mathbf{E}_l[\xi_l] + \sum_{t=1}^{l-1}\xi_t\right\|^2 + \mathbf{E}_l\left[\left\|\xi_l - \mathbf{E}_l[\xi_l]\right\|^2\right] \\ &\overset{(135)}{\leq} \left(1 + \frac{1}{\tau-1}\right)\left\|\sum_{t=1}^{l-1}\xi_t\right\|^2 + \tau\left\|\mathbf{E}_l[\xi_l]\right\|^2 + \mathbf{E}_l\left[\left\|\xi_l - \mathbf{E}_l[\xi_l]\right\|^2\right].\end{aligned}$$

Taking full mathematical expectation and using tower property (140) we derive

$$\mathbf{E}\left[\left\|\sum_{t=1}^{l}\xi_t\right\|^2\right] \leq \left(1 + \frac{1}{\tau-1}\right)\mathbf{E}\left[\left\|\sum_{t=1}^{l-1}\xi_t\right\|^2\right] + \tau\mathbf{E}\left[\left\|\mathbf{E}_l[\xi_l]\right\|^2\right] + \mathbf{E}\left[\left\|\xi_l - \mathbf{E}_l[\xi_l]\right\|^2\right]$$

for all $l = 1, \ldots, \tau$. Unrolling the recurrence for $\mathbf{E}\left[\left\|\sum_{t=1}^{l}\xi_t\right\|^2\right]$ we obtain

$$\mathbf{E}\left[\left\|\sum_{t=1}^{\tau}\xi_t\right\|^2\right] \leq \tau\sum_{t=1}^{\tau}\left(1 + \frac{1}{\tau-1}\right)^{\tau-t}\mathbf{E}\left[\left\|\mathbf{E}_t[\xi_t]\right\|^2\right] + \sum_{t=1}^{\tau}\left(1 + \frac{1}{\tau-1}\right)^{\tau-t}\mathbf{E}\left[\left\|\xi_t - \mathbf{E}_t[\xi_t]\right\|^2\right].$$

Since $\left(1 + \frac{1}{\tau-1}\right)^{\tau-t} \leq \left(1 + \frac{1}{\tau-1}\right)^{\tau-1} \leq e$ for all $t = 1, \ldots, \tau$ we get (141). □

Lemma I.2 (see also Lemma 2 from (Stich, 2019)). *Let $\{r_k\}_{k\geq 0}$ satisfy*

$$r_K \leq \frac{a}{\gamma W_K} + c_1\gamma + c_2\gamma^2 \tag{142}$$

for all $K \geq 0$ with some constants $a, c_2 \geq 0$, $c_1 \geq 0$ where $\{w_k\}_{k\geq 0}$ and $\{W_K\}_{K\geq 0}$ are defined in (12), $\gamma \leq \frac{1}{h}$. Then for all K such that

$$\begin{aligned}&\textit{either} \quad \frac{\ln\left(\max\{2, \min\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\}\}\right)}{K} \leq \rho \\ &\textit{or} \quad \frac{1}{h} \leq \frac{\ln\left(\max\{2, \min\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\}\}\right)}{\mu K}\end{aligned}$$

and

$$\gamma = \min\left\{\frac{1}{h}, \frac{\ln\left(\max\{2, \min\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\}\}\right)}{\mu K}\right\} \tag{143}$$

we have that

$$r_K = \widetilde{\mathcal{O}}\left(ha\exp\left(-\min\left\{\frac{\mu}{h}, \rho\right\}K\right) + \frac{c_1}{\mu K} + \frac{c_2}{\mu^2K^2}\right). \tag{144}$$

Proof. Since $W_K \geq w_K = (1-\eta)^{-(K+1)}$ we have

$$r_K \quad \leq \quad (1-\eta)^{K+1}\frac{a}{\gamma} + c_1\gamma + c_2\gamma^2 \leq \frac{a}{\gamma}\exp\left(-\eta(K+1)\right) + c_1\gamma + c_2\gamma^2. \tag{145}$$

Next we consider two possible situations.

1. If $\frac{1}{h} \geq \frac{\ln\left(\max\left\{2,\min\left\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\right\}\right\}\right)}{\mu K}$ then we choose $\gamma = \frac{\ln\left(\max\left\{2,\min\left\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\right\}\right\}\right)}{\mu K}$ and get that

$$\begin{aligned} r_K &\overset{(145)}{\leq} \frac{a}{\gamma}\exp\left(-\eta(K+1)\right) + c_1\gamma + c_2\gamma^2 \\ &= \widetilde{\mathcal{O}}\left(a\mu K\exp\left(-\min\left\{\rho, \frac{\ln\left(\max\left\{2,\min\left\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\right\}\right\}\right)}{K}\right\}K\right)\right) \\ &\quad + \widetilde{\mathcal{O}}\left(\frac{c_1}{\mu K} + \frac{c_2}{\mu^2K^2}\right). \end{aligned}$$

   Since $\frac{\ln\left(\max\left\{2,\min\left\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\right\}\right\}\right)}{K} \leq \rho$ we have

$$\begin{aligned} r_K &= \widetilde{\mathcal{O}}\left(a\mu K\exp\left(-\ln\left(\max\left\{2,\min\left\{\frac{a\mu^2K^2}{c_1}, \frac{a\mu^3K^3}{c_2}\right\}\right\}\right)\right)\right) \\ &\quad + \widetilde{\mathcal{O}}\left(\frac{c_1}{\mu K} + \frac{c_2}{\mu^2K^2}\right) \\ &= \widetilde{\mathcal{O}}\left(\frac{c_1}{\mu K} + \frac{c_2}{\mu^2K^2}\right). \end{aligned}$$

2. If $\frac{1}{h} \leq \frac{\ln\left(\max\left\{2,\min\left\{a\mu^2K^2/c_1, a\mu^3K^3/c_2\right\}\right\}\right)}{\mu K}$ then we choose $\gamma = \frac{1}{h}$ which implies that

$$\begin{aligned} r_K &\overset{(145)}{\leq} ha\exp\left(-\min\left\{\frac{\mu}{h}, \frac{\rho}{4}\right\}(K+1)\right) + \frac{c_1}{h} + \frac{c_2}{h^2} \\ &= \widetilde{\mathcal{O}}\left(ha\exp\left(-\min\left\{\frac{\mu}{h}, \rho\right\}K\right) + \frac{c_1}{\mu K} + \frac{c_2}{\mu^2K^2}\right). \end{aligned}$$

Combining the obtained bounds we get the result. □

Lemma I.3. *Let $\{r_k\}_{k\geq 0}$ satisfy*

$$r_K \leq \frac{a}{\gamma K} + \frac{b_1\gamma}{K} + \frac{b_2\gamma^2}{K} + c_1\gamma + c_2\gamma^2 \tag{146}$$

for all $K \geq 0$ with some constants $a > 0$, $b_1, b_2, c_1, c_2 \geq 0$ where $\gamma \leq \gamma_0$. Then for all K and

$$\gamma = \min\left\{\gamma_0, \sqrt{\frac{a}{b_1}}, \sqrt[3]{\frac{a}{b_2}}, \sqrt{\frac{a}{c_1K}}, \sqrt[3]{\frac{a}{c_2K}}\right\}$$

we have that

$$r_K = \mathcal{O}\left(\frac{a}{\gamma_0 K} + \frac{\sqrt{ab_1}}{K} + \frac{\sqrt[3]{a^2b_2}}{K} + \sqrt{\frac{ac_1}{K}} + \frac{\sqrt[3]{a^2c_2}}{K^{2/3}}\right). \tag{147}$$

Proof. We have

$$\begin{aligned} r_K &\leq \frac{a}{\gamma K} + \frac{b_1\gamma}{K} + \frac{b_2\gamma^2}{K} + c_1\gamma + c_2\gamma^2 \\ &\leq \frac{a}{\min\left\{\gamma_0, \sqrt{\frac{a}{b_1}}, \sqrt[3]{\frac{a}{b_2}}, \sqrt{\frac{a}{c_1K}}, \sqrt[3]{\frac{a}{c_2K}}\right\}K} + \frac{b_1}{K}\cdot\sqrt{\frac{a}{b_1}} + \frac{b_2}{K}\cdot\sqrt[3]{\frac{a^2}{b_2^2}} + c_1\cdot\sqrt{\frac{a}{c_1K}} + c_2\left(\sqrt[3]{\frac{a}{c_2K}}\right)^2 \\ &= \mathcal{O}\left(\frac{a}{\gamma_0 K} + \frac{\sqrt{ab_1}}{K} + \frac{\sqrt[3]{a^2b_2}}{K} + \sqrt{\frac{ac_1}{K}} + \frac{\sqrt[3]{a^2c_2}}{K^{2/3}}\right). \end{aligned}$$

□

Table 7: The parameters for which the methods from Table 2 satisfy Assumption 2.3/E.1. Absolute constants were omitted. The meaning of the expressions appearing in the table, as well as their justification, is detailed in Section 5. 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" inequality (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.

Method, Setting	A, $\widetilde{A}$, $\widehat{A}$, A'	B, $\widetilde{B}$, $\widehat{B}$, B'	ρ	C	F, $\widetilde{F}$, $\widehat{F}$, F'	G	D_1', D_1, $\widetilde{D}_1$, $\widehat{D}_1$, D_2, D_3
`Local-SGD` UBV, ζ-Het.	L, $-$, $-$, L	0, $-$, $-$, 0	1	0	L^2, $-$, $-$, L^2	0	$\frac{\sigma^2}{n}$, $\sigma^2+\zeta_*^2$, $-$, $-$, 0, $\tau\sigma^2+\tau^2\zeta^2$
`Local-SGD` UBV, Het.	$-$, L, 0, L	$-$, 0, 0, 0	1	0	$-$, L^2, 0, L^2	0	$\frac{\sigma^2}{n}$, $-$, ζ_*^2, σ^2, 0, $(\tau-1)\sigma^2+(\tau-1)^2\zeta_*^2$
`Local-SGD` ES, ζ-Het.	$\mathcal{L}$, $-$, $-$, $\frac{\mathcal{L}}{n}+L$	0, $-$, $-$, 0	1	0	$\mathcal{L}L$, $-$, $-$, $\frac{\mathcal{L}L}{n}+L^2$	0	$\frac{\sigma_*^2}{n}$, $\sigma_*^2+\zeta_*^2$, $-$, $-$, 0, $(\tau-1)\left(\sigma_*^2+\zeta_*^2+\frac{\zeta_*^2}{\gamma\mu}\right)$
`Local-SGD` ES, Het.	$-$, L, $\mathcal{L}$, $\frac{\mathcal{L}}{n}+L$	$-$, 0, 0, 0	1	0	$-$, L^2, $\mathcal{L}L$, $\frac{\mathcal{L}L}{n}+L^2$	0	$\frac{\sigma_*^2}{n}$, $-$, ζ_*^2, σ_*^2, 0, $(\tau-1)\sigma_*^2+(\tau-1)^2\zeta_*^2$
`Local-SVRG` simple, ζ-Het.	$\max L_{ij}$, $-$, $-$, $\frac{\max L_{ij}}{n}+L$	1, $-$, $-$, $\frac{1}{n}$	q	$\max L_{ij} q$	$\max L_{ij}L$, $-$, $-$, $\frac{\max L_{ij}L}{n}+L^2$	$\max L_{ij} Lq$	0, ζ_*^2, $-$, $-$, 0, $(\tau-1)\left(\zeta_*^2+\frac{\zeta_*^2}{\gamma\mu}\right)$
`Local-SVRG` simple, Het.	$-$, L, $\max L_{ij}$, $\frac{\max L_{ij}}{n}+L$	$-$, 0, 1, $\frac{1}{n}$	q	$\max L_{ij} q$	$-$, L^2, $\max L_{ij}L$, $\frac{\max L_{ij}L}{n}+L^2$	$\max L_{ij} Lq$	0, $-$, ζ_*^2, 0, 0, $(\tau-1)^2\zeta_*^2$
`S*-Local-SGD` UBV, Het.	$-$, L, 0, L	$-$, 0, 0, 0	1	0	$-$, L^2, 0, l^2	0	$\frac{\sigma^2}{n}$, $-$, 0, σ^2, $(\tau-1)\sigma^2$
`SS-Local-SGD` UBV, Het., $p=q$, $r=\lceil 1/p \rceil$	$-$, L, 0, L	$-$, 1, 0, 0	p	Lp	$-$, L^2, 0, L^2	0	$\frac{\sigma^2}{n}$, $-$, $p\sigma^2$, σ^2, 0, $\frac{(1-p)\sigma^2}{p}$
`SS-Local-SGD` ES, Het., $p=q$, $r=\lceil 1/p \rceil$	$-$, L, $\mathcal{L}$, $\frac{\mathcal{L}}{n}+L$	$-$, 1, 0, 0	p	$Lp+\mathcal{L}p^2$	$-$, L^2, $\mathcal{L}L$, $\frac{\mathcal{L}L}{n}+L^2$	0	$\frac{\sigma_*^2}{n}$, $-$, 0, σ_*^2, $p^2\sigma_*^2$, $\frac{(1-p)\sigma_*^2}{p}$
`S*-Local-SGD*` simple, Het.	$-$, L, $\max L_{ij}$, $\frac{\max L_{ij}}{n}+L$	$-$, 0, 0, 0	p	0	$-$, L^2, $\max L_{ij}L$, $\frac{L\max L_{ij}}{n}+L^2$	0	0, $-$, 0, 0, 0, 0
`S-Local-SVRG` simple, Het., $q=\frac{1}{m}$, $m\geq\frac{1}{p}$	$-$, L, $\max L_{ij}$, $\frac{\max L_{ij}}{n}+L$	$-$, 1, 1, $\frac{1}{n}$	$\frac{1}{m}$	$\frac{L+\max L_{ij}}{m}$	$-$, L^2, $\max L_{ij}L$, $\frac{L\max L_{ij}}{n}+L^2$	0	0, $-$, 0, 0, 0, 0