Tighter Theory for Local SGD on Identical and Heterogeneous Data*

Ahmed Khaled
Cairo University, KAUST†

Konstantin Mishchenko
KAUST

Peter Richtárik
KAUST

Abstract

We provide a new analysis of local SGD, removing unnecessary assumptions and elaborating on the difference between two data regimes: identical and heterogeneous. In both cases, we improve the existing theory and provide values of the optimal stepsize and optimal number of local iterations. Our bounds are based on a new notion of variance that is specific to local SGD methods with different data. The tightness of our results is guaranteed by recovering known statements when we plug $H = 1$, where H is the number of local steps. The empirical evidence further validates the severe impact of data heterogeneity on the performance of local SGD.

1 Introduction

Modern hardware increasingly relies on the power of uniting many parallel units into one system. This approach requires optimization methods that target specific issues arising in distributed environments such as decentralized data storage. Not having data in one place implies that computing nodes have to communicate back and forth to keep moving toward the solution of the overall problem. A number of efficient first-, second-order and dual methods that are capable of reducing the communication overhead existed in the literature for a long time, some of which are in certain sense optimal.

Yet, when Federated Learning (FL) showed up, it turned out that the problem of balancing the communication and computation had not been solved. On the one hand, Minibatch Stochastic Gradient Descent (SGD), which averages the result of stochastic gradient steps computed in parallel on several machines, again demonstrated its computation efficiency. Seeking communication efficiency, Konečný et al. (2016); McMahan et al. (2017) proposed to use a natural variant of Minibatch SGD—*Local SGD* (Algorithm 1), which does a few SGD iterations locally on each involved node and only then computes the average. This approach saves a lot of time on communication, but, unfortunately, in terms of theory things were not as great as in terms of practice and there are still gaps in our understanding of Local SGD.

The idea of local SGD in fact is not recent, it traces back to the work of Mangasarian (1995) and has since been popular among practitioners from different communities. An asymptotic analysis can be found in Mangasarian (1995) and quite a few recent papers proved new convergence results, making the bounds tighter with every work. The theory has been developing in two important regimes: identical and heterogeneous data.

The identical data regime is more of interest if the data are actually stored in one place. In that case, we can access it on each computing device at no extra cost and get a fast, scalable method. Although not very general, this framework is already of interest to a wide audience due to its efficiency in training large-scale machine learning models (Lin et al., 2020). The first contribution of our work is to provide the fastest known rate of convergence for this regime under weaker assumptions than in prior work.

Federated learning, however, is done on a very large number of mobile devices, and is operating in a highly non-i.i.d. regime. To address this, we present the first analysis of Local SGD that applies to *arbitrarily heterogeneous data*, while all previous works assumed a certain type of similarity between the data or local gradients.

*This work extends two papers Khaled et al. (2019a;b) presented at the NeurIPS 2019 Federated Learning Workshop.

†This paper was prepared when the author was a research intern at KAUST.

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

Algorithm 1 Local SGD

Input: Stepsize $\gamma > 0$, initial vector $x_0 = x_0^m$ for all $m \in [M]$, synchronization timesteps $t_1, t_2, \ldots$.

for $t = 0, 1, \ldots$ **do**
  for $m = 1, \ldots, M$ in parallel **do**
    Sample $z_m \overset{\text{i.i.d.}}{\sim} \mathcal{D}_m$.
    if data is identical **then**
      Compute $g_t^m = g(f, x_t^m, z_m)$ such that $\mathbb{E}\left[g_t^m \mid x_t^m\right] = \nabla f(x_t^m)$.
    else
      Compute $g_t^m = g(f_m, x_t^m, z_m)$ such that $\mathbb{E}\left[g_t^m \mid x_t^m\right] = \nabla f_m(x_t^m)$.
    end if
    $x_{t+1}^m = \begin{cases} \frac{1}{m}\sum_{j=1}^M (x_t^j - \gamma g_t^j), & \text{if } t = t_p \text{ for some } p \in \mathbb{N} \\ x_t^m - \gamma g_t^m, & \text{otherwise.} \end{cases}$
  end for
end for

To explain the challenge of heterogeneity better, let us introduce the problem we are trying to solve. Given that there are M devices and corresponding local losses $f_m : \mathbb{R}^d \to \mathbb{R}$, we want to find

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) = \frac{1}{M} \sum_{m=1}^{M} f_m(x) \right\}. \tag{1}$$

In the case of identical data, we are able to obtain on each node an unbiased estimate of the gradient ∇f. In the case of heterogeneous data, m-th node can only obtain an unbiased estimate of the gradient ∇f_m. Data similarity can then be formulated in terms of the differences between functions $f_1, \ldots, f_M$. If the underlying data giving rise to the loss functions are i.i.d., the function share optima and one could even minimize them separately, averaging the results at the end. We will demonstrate this rigorously later in the paper.

If the data are dissimilar, however, we need to be much more careful since running SGD locally will yield solutions of local problems. Clearly, their average might not minimize the true objective (1), and this poses significant issues for the convergence of Local SGD.

To properly discuss the efficiency of local SGD, we also need a practical way of quantifying it. Normally, a method's efficiency is measured by the total number of times each function f_m is touched and the cost of the touches. On the other hand, in distributed learning we also care about how much information each computing node needs to communicate. In fact, when communication is as expensive as is the case in FL, we predominantly care about communication. The question we address in this paper, thus, can be posed as follows: how many times does each node need to communicate if we want to solve (1) up to accuracy ε? Equivalently, we can ask for the optimal *synchronization interval length* between communications, H, i.e. how many computation steps per one communication we can allow for. We next review related work and then present our contributions.

2 Related Work

While local SGD has been used among practitioners for a long time, see e.g. (Coppola, 2015; McDonald et al., 2010), its theoretical analysis has been limited until recently. Early theoretical work on the convergence of local methods exists as in (Mangasarian, 1995), but no convergence rate was given there. The previous work can mainly be divided into two groups: those assuming identical data (that all nodes have access to the same dataset) and those that allow each node to hold its own dataset. As might be expected, the analysis in the latter case is more challenging, more limited, and usually shows worse rates. We note that in recent work more sophisticated local stochastic gradient methods have been considered, for example with momentum (Yu et al., 2019a; Wang et al., 2019), with quantization (Reisizadeh et al., 2019; Basu et al., 2019), with adaptive stepsizes (Xie et al., 2019) and with various variance-reduction methods (Liang et al., 2019; Sharma et al., 2019; Karimireddy et al., 2019). Our work is complimentary to these approaches, and provides improved rates and analysis for the vanilla method.

2.1 Local SGD with Identical Data

The analysis of local SGD in this setting shows that a reduction in communication is possible without affecting the asymptotic convergence rate of Minibatch SGD with M nodes (albeit with usually worse dependence on constants). An overview of related work on local SGD for convex objectives is given in Table 1. We note that analysis for nonconvex objectives has been carried out in a few recent works (Zhou and Cong, 2018; Wang and Joshi, 2018; Jiang and Agrawal, 2018), but our fo-

Table 1: Existing theoretical bounds for local SGD for **identical data** with convex objectives.

Unbounded gradient	$H = T$ convergent	$C(T)$[a] f strongly convex	$C(T)$ f convex	Reference
✗	✗	$\Omega(\sqrt{MT})$	✗	Stich, 5/2018
✗	✗	$\Omega\left(\sqrt{MT}\right)$	✗	Basu et al., 6/2018
✓	✗	$\tilde{\Omega}\left(M\right)$	$\Omega(M^{3/2}T^{1/2})$	Stich and Karimireddy, 9/2019
✓	✗	$\tilde{\Omega}\left(M^{1/3}T^{1/3}\right)$[b]	-	Haddadpour et al., 10/2019
✓	✓	$\tilde{\Omega}(M)$	$\Omega(M^{3/2}T^{1/2})$	**THIS WORK, 9/2019-1/2020**

[a] $C(T)$ denotes the minimum number of communication steps required at each of T iterations to achieve a linear speedup in the number of nodes M.

[b] The PL inequality, a generalization of strong convexity, is assumed in (Haddadpour et al., 2019), but for comparison we specialize to strong convexity.

cus in this work is on convex objectives and hence they were not included in Table 1. The comparison shows that we attain superior rates in the strongly convex setting to previous work with the exception of the concurrent[1] work of Stich and Karimireddy (2019) and we attain these rates under less restrictive assumptions on the optimization process compared to them. We further provide a novel analysis in the convex case, which has not been previously explored in the literature, with the exception of (Stich and Karimireddy, 2019). Their analysis attains the same communication complexity but is much more pessimistic about possible values of H. In particular, it does not recover the convergence of one-shot averaging, i.e. substituting $H = T$ or even $H = T/M$ gives noninformative bounds, unlike our Theorem 1.

In addition to the works listed in the table, Dieuleveut and Patel (2019) also analyze local SGD for identical data under a Hessian smoothness assumption in addition to gradient smoothness, strong convexity, and uniformly bounded variance. However, we believe that there are issues in their proof that we explain in Section 11 in the supplementary material. As a result, the work is excluded from the table.

2.2 Local SGD with Heterogeneous Data

An overview of related work on local SGD in this setting is given in Table 2. In addition to the works in Table 2, Wang et al. (2018) analyze a local gradient descent method under convexity, bounded dissimilarity, and bounded gradients, but do not show convergence to arbitrary precisions. Li et al. (2020) analyze federated averaging (discussed below) in the strongly convex and nonconvex cases under bounded gradient norms. However, their result is not included in Table 2 because in the more general setting of federated averaging, their analysis and experiments suggest that retaining a linear speedup is not possible.

Local SGD is at the core of the *Federated Averaging* algorithm which is popular in federated learning applications (Konečný et al., 2016). Essentially, Federated Averaging is a variant of Local SGD with participating devices sampled randomly. This algorithm has been used in several machine learning applications such as mobile keyboard prediction (Hard et al., 2018), and strategies for improving its communication efficiency were explored in (Konečný et al., 2016). Despite its empirical success, little is known about convergence properties of this method and it has been observed to diverge when too many local steps are performed (McMahan et al., 2017). This is not so surprising as the majority of common assumptions are not satisfied; in particular, the data are typically very non-i.i.d. (McMahan et al., 2017), so the local gradients can point in different directions. This property of the data can be written for any vector x and indices i, j as

$$\|\nabla f_i(x) - \nabla f_j(x)\| \gg 1.$$

Unfortunately, it is very hard to analyze local methods without assuming a bound on the dissimilarity of $\nabla f_i(x)$ and $\nabla f_j(x)$. For this reason, almost all prior work assumed some regularity notion over the functions such as bounded dissimilarity (Yu et al., 2019a; Li et al., 2020; Yu et al., 2019b; Wang et al., 2018) or bounded gradient diversity (Haddadpour and Mahdavi, 2019) and addressed other less challenging aspects of federated learning such as decentralized communication, nonconvexity of the objective or unbalanced data partitioning. In fact, a common way to make the analysis simple is to assume Lipschitzness of

[1] Made available online one day after the first version of our work was.

Table 2: Existing theoretical bounds for local SGD with **heterogeneous data**.

Unbounded gradient	Unbounded dissimilarity/diversity	$C(T)$ f strongly convex	$C(T)$ f convex	$C(T)$ f nonconvex	Reference
✗	✗	-	-	$\Omega\left(M^{3/4}T^{3/4}\right)$	Yu et al., 7/2018
✓	✗	-	-	$\Omega(T)$	Jiang and Agrawal, 12/2018
✗	✗	$\Omega\left(\sqrt{MT}\right)$	-	$\Omega\left(M^{3/4}T^{3/4}\right)$	Basu et al., 6/2019
✓	✗	$\Omega\left(M^{1/3}T^{1/3}\right)$	-	$\Omega\left(M^{3/2}T^{1/2}\right)$	Haddadpour and Mahdavi, 10/2019
✓	✓	-	$\Omega\left(M^{3/4}T^{3/4}\right)$	-	**THIS WORK, 1/2020**

local functions, $\|\nabla f_i(x)\| \leqslant G$ for any x and i. We argue that this assumption is pathological and should be avoided when seeking a meaningful convergence bound. First of all, in unconstrained strongly convex minimization this assumption can not be satisfied, making the analysis in works like (Stich, 2019) questionable. Second, there exists at least one method, whose convergence is guaranteed under bounded variance (Juditsky et al., 2011), but in practice the method diverges (Chavdarova et al., 2019; Mishchenko et al., 2019). Finally, under the bounded gradients assumption we have

$$\|\nabla f_i(x) - \nabla f_j(x)\| \leqslant \|\nabla f_i(x)\| + \|\nabla f_j(x)\| \leqslant 2G.$$

In other words, we lose control over the difference between the functions. Since G bounds not just dissimilarity, but also the gradients themselves, it makes the statements less insightful or even vacuous. For instance, it is not going to be tight if the data are actually i.i.d. since G in that case will remain a positive constant. In contrast, we will show that the rate should depend on a much more meaningful quantity,

$$\sigma_{\text{dif}}^2 \overset{\text{def}}{=} \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{z_m\sim\mathcal{D}_m}\left[\|\nabla f_m(x_*, z_m)\|^2\right],$$

where x_* is a fixed minimizer of f and $f_m(\cdot, z_m)$ for $z_m \sim \mathcal{D}$ are stochastic realizations of f_m (see the next section for the setting). Obviously, for all nondegenerate sampling distributions $\mathcal{D}_m$ the quantity σ_{dif} is finite and serves as a natural measure of variance in local methods. We note that an attempt to get more general convergence statement has been made by (Li et al., 2018), but unfortunately their guarantee is strictly worse than that of minibatch Stochastic Gradient Descent (SGD). In the overparameterized regime where $\sigma_{\text{dif}} = 0$, Zhang and Li (2019) prove the convergence of Local SGD with arbitrary H.

Our earlier workshop paper (Khaled et al., 2019a) explicitly analyzed Local Gradient Descent (Local GD) as opposed to Local SGD, where there is no stochasticity in the gradients. An analysis of Local GD for nonconvex objectives with the PL inequality and under bounded gradient diversity was subsequently carried out by Haddadpour and Mahdavi (2019).

3 Settings and Contributions

Assumption 1. Assume that the set of minimizers of (1) is nonempty. Each f_m is μ-strongly convex for $\mu \geqslant 0$ and L-smooth. That is, for all $x, y \in \mathbb{R}^d$

$$\frac{\mu}{2}\|x-y\|^2 \leqslant f_m(x) - f_m(y) - \langle \nabla f_m(y), x-y\rangle \leqslant \frac{L}{2}\|x-y\|^2.$$

When $\mu = 0$, we say that each f_m is just convex. When $\mu \neq 0$, we define $\kappa \overset{\text{def}}{=} \frac{L}{\mu}$, the condition number.

Assumption 1 formulates our requirements on the overall objective. Next, we have two different sets of assumptions on the stochastic gradients that model different scenarios, which also lead to different convergence rates.

Assumption 2. Given a function h, a point $x \in \mathbb{R}^d$, and a sample $z \sim \mathcal{D}$ drawn i.i.d. according to a distribution $\mathcal{D}$, the stochastic gradients $g = g(h, x, z)$ satisfy $\mathbb{E}_{z\sim\mathcal{D}}\left[g(h,x,z)\right] = \nabla h(x), \mathbb{E}_{z\sim\mathcal{D}}\left[\|g(h,x,z) - \nabla h(x)\|^2\right] \leqslant \sigma^2$.

Assumption 2 holds for example when $g(x,z) = \nabla h(x) + \xi_z$ for a random variable ξ_z of expected bounded squared norm: $\mathbb{E}_{z\sim\mathcal{D}}\left[\|\xi_z\|^2\right] \leqslant \sigma^2$. Assumption 2, however, typically does not hold for finite-sum problems where $g(x,z)$ is a gradient of the one functions in the finite-sum. To capture this setting, we consider the following assumption:

Assumption 3. Given an L-smooth and μ-strongly convex (possibly with $\mu = 0$) function $h : \mathbb{R}^d \to \mathbb{R}$ written as an expectation $h = \mathbb{E}_{z\sim\mathcal{D}}\left[h(x,z)\right]$, we assume that a stochastic gradient $g = g(h,x,z)$ is computed by $g(h,x,z) = \nabla h(x,z)$. We assume that $h(\cdot, z) : \mathbb{R}^d \to \mathbb{R}$ is almost-surely L-smooth and μ-strongly convex (with the same L and μ as h).

When Assumption 3 is assumed **in the identical data setting**, we assume it is satisfied on each node

$m \in [M]$ with $h = f$ and distribution $\mathcal{D}_m$, and we define as a measure of variance at the optimum

$$\sigma_{\text{opt}}^2 \overset{\text{def}}{=} \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{z_m \sim \mathcal{D}_m}\left[\|\nabla f(x_*, z_m)\|^2\right].$$

Whereas **in the heterogeneous data** setting we assume that it is satisfied on each node $m \in [M]$ with $h = f_m$ and distribution $\mathcal{D}_m$, and we analogously define

$$\sigma_{\text{dif}}^2 \overset{\text{def}}{=} \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{z_m \sim \mathcal{D}_m}\left[\|\nabla f_m(x_*, z_m)\|^2\right].$$

Assumption 3 holds, for example, for finite-sum optimization problems with uniform sampling and permits direct extensions to more general settings such as expected smoothness Gower et al. (2019).

Our contributions are as follows:

1. In the identical data setting under Assumptions 1 and 2 with $\mu > 0$, we prove that the iteration complexity of Local SGD to achieve ε-accuracy is

   $$\tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu^2 M \varepsilon}\right)$$

   in squared distance from the optimum provided that $T = \Omega\left(\kappa\left(H-1\right)\right)$. This improves the communication complexity in prior work (see Table 1) with a tighter results compared to concurrent work (recovering convergence for $H = 1$ and $H = T$). When $\mu = 0$ we have that the iteration complexity of Minibatch SGD to attain an ε-accurate solution in functional suboptimality is

   $$\mathcal{O}\left(\frac{L^2\left\|x_0 - x_*\right\|^4}{M\varepsilon^2} + \frac{\sigma^4}{L^2 M \varepsilon^2}\right),$$

   provided that $T = \Omega\left(M^3 H^2\right)$. We further show that the same ε-dependence holds in both the $\mu > 0$ and $\mu = 0$ cases under Assumption 3. This has not been explored in the literature on Local SGD before, and hence we obtain the first results that apply to arbitrary convex and smooth finite-sum problems.

2. When the data on each node is different and Assumptions 1 and 3 hold with $\mu = 0$, the iteration complexity needed by Local SGD to achieve an ε-accurate solution in functional suboptimality is

   $$\mathcal{O}\left(\frac{L^2\left\|x_0 - x_*\right\|^4}{M\varepsilon^2} + \frac{\sigma_{\text{dif}}^4}{L^2 M \varepsilon^2}\right)$$

   provided that $T = \Omega(M^3 H^4)$. This improves upon previous work by not requiring any restrictive assumptions on the gradients and is the first analysis to capture true *data heterogeneity* between different nodes.

3. We verify our results by experimenting with logistic regression on multiple datasets, and investigate the effect of heterogeneity on the convergence speed.

4 Convergence Theory

The following quantity is crucial to the analysis of both variants of local SGD, and measures the deviation of the iterates from their average $\hat{x}_t$ over an epoch:

$$V_t \overset{\text{def}}{=} \frac{1}{M}\sum_{m=1}^{M}\left\|x_t^m - \hat{x}_t\right\|^2 \text{ where } \hat{x}_t \overset{\text{def}}{=} \frac{1}{M}\sum_{m=1}^{M} x_t^m.$$

To prove our results, we follow the line of work started by Stich (2019) and first show a recurrence similar to that of SGD up to an error term proportional to V_t, then we bound each V_t term individually or the sum of V_t's over an epoch. All proofs are relegated to the supplementary material.

4.1 Identical Data

Our first lemma presents a bound on the sequence of the V_t in terms of the synchronization interval H.

Lemma 1. Choose a stepsize $\gamma > 0$ such that $\gamma \leqslant \frac{1}{2L}$. Under Assumptions 1, and 2 we have that for Algorithm 1 with $\max_p |t_p - t_{p+1}| \leqslant H$ and with identical data, for all $t \geqslant 1$

$$\mathbb{E}\left[V_t\right] \leqslant \left(H-1\right)\gamma^2\sigma^2.$$

Combining Lemma 1 with perturbed iterate analysis as in (Stich, 2019) we can recover the convergence of local SGD for strongly-convex functions:

Theorem 1. Suppose that Assumptions 1, and 2 hold with $\mu > 0$. Then for Algorithm 1 run with identical data, a constant stepsize $\gamma > 0$ such that $\gamma \leqslant \frac{1}{4L}$, and $H \geqslant 1$ such that $\max_p |t_p - t_{p+1}| \leqslant H$,

$$\mathbb{E}\left[\left\|\hat{x}_T - x_*\right\|^2\right] \leqslant \left(1-\gamma\mu\right)^T\left\|x_0 - x_*\right\|^2 + \frac{\gamma\sigma^2}{\mu M} + \frac{2L\gamma^2\left(H-1\right)\sigma^2}{\mu}. \tag{2}$$

By (2) we see that the convergence of local SGD is the same as Minibatch SGD plus an additive error term which can be controlled by controlling the size of H, as the next corollary and the successive discussion show.

Corollary 1. Choosing $\gamma = \frac{1}{\mu a}$, with $a = 4\kappa + t$ for $t > 0$ and we take $T = 2a \log a$ steps. Then substituting in (2) and using that $1 - x \leqslant \exp(-x)$ and some algebraic manipulation we can conclude that,

$$\mathbb{E}\left[\|r_T\|^2\right] = \tilde{\mathcal{O}}\left(\frac{\|r_0\|^2}{T^2} + \frac{\sigma^2}{\mu^2 MT} + \frac{\kappa\sigma^2(H-1)}{\mu^2 T^2}\right).$$

where $r_t = \hat{x}_t - x_*$ and $\tilde{\mathcal{O}}(\cdot)$ ignores polylogarithmic and constant numerical factors.

Recovering fully synchronized Minibatch SGD. When $H = 1$ the error term vanishes and we obtain directly the ordinary rate of Minibatch SGD.

Linear speedup in the number of nodes M. We see that choosing $H = \mathcal{O}(T/M)$ leads to an asymptotic convergence rate of $\tilde{\mathcal{O}}\left(\frac{\sigma^2\kappa}{\mu^2 MT}\right)$ which shows the same linear speedup of Minibatch SGD but with worse dependence on κ. The number of communications in this case is then $C(T) = T/H = \tilde{\Omega}(M)$.

Local SGD vs Minibatch SGD. We assume that the statistical σ^2/T dependence dominates the dependence on the initial distance $\|x_0 - x_*\|^2/T^2$. From Corollary 1, we see that in order to achieve the same convergence guarantees as Minibatch SGD, we must have $H = \mathcal{O}\left(\frac{T}{\kappa M}\right)$, achieving a communication complexity of $\mathcal{O}(\kappa M)$. This is only possible when $T > \kappa M$. It follows that given a number of steps T the optimal H is $H = 1 + \lfloor T/(\kappa M)\rfloor$ achieving a communication complexity of $\tilde{\Omega}\left(\min(T, \kappa M)\right)$.

One-shot averaging. Putting $H = T + 1$ yields a convergence rate of $\tilde{\mathcal{O}}(\sigma^2\kappa/(\mu^2 T))$, showing no linear speedup but showing convergence, which improves upon all previous work. However, we admit that simply using Jensen's inequality to bound the distance of the average iterate $\mathbb{E}\left[\|\hat{x}_T - x_*\|^2\right]$ would yield a better asymptotic convergence rate of $\tilde{\mathcal{O}}(\sigma^2/(\mu^2 T))$. Under a Lipschitz Hessian assumption, Zhang et al. (2013) show that one-shot averaging can attain a linear speedup in the number of nodes, so one may do analysis of local SGD under this additional assumption to try to remove this gap, but this is beyond the scope of our work.

Similar results can be obtained for weakly convex functions, as the next Theorem shows.

Theorem 2. Suppose that Assumptions 1, 2 hold with $\mu = 0$ and that a constant stepsize γ such that $\gamma \geqslant 0$ and $\gamma \leqslant \frac{1}{4L}$ is chosen and that Algorithm 1 is run for identical data with $H \geqslant 1$ such that $\sup_p |t_p - t_{p+1}| \leqslant H$, then for $\bar{x}_T = \frac{1}{T}\sum_{t=1}^T \hat{x}_t$,

$$\mathbb{E}\left[f(\bar{x}_T) - f(x_*)\right] \leqslant \frac{2}{\gamma T}\|x_0 - x_*\|^2 + \frac{2\gamma\sigma^2}{M} + 4\gamma^2 L\sigma^2(H-1). \quad (3)$$

Theorem 2 essentially tells the same story as Theorem 1: convergence of local SGD is the same as Minibatch SGD up to an additive constant whose size can be controlled by controlling H.

Corollary 2. Assume that $T \geqslant M$. Choosing $\gamma = \frac{\sqrt{M}}{4L\sqrt{T}}$, then substituting in (3) we have,

$$\mathbb{E}\left[f(\bar{x}_T) - f(x_*)\right] \leqslant \frac{8\|x_0 - x_*\|^2}{\sqrt{MT}} + \frac{\sigma^2}{2L\sqrt{MT}} + \frac{\sigma^2 M(H-1)}{LT}.$$

Linear speedup and optimal H. From Corollary 2 we see that if we choose $H = \mathcal{O}(\sqrt{T}M^{-3/2})$ then we obtain a linear speedup, and the number of communication steps is then $C = T/H = \Omega\left(M^{3/2}T^{1/2}\right)$, and we get that the optimal H is then $H = 1 + \left\lfloor T^{1/2}M^{-3/2}\right\rfloor$.

The previous results were obtained under Assumption 2. Unfortunately, this assumption does not easily capture the finite-sum minimization scenario where $f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)$ and each stochastic gradient g_t is sampled uniformly at random from the sum.

Using smaller stepsizes and more involved proof techniques, we can show that our results still hold in the finite-sum setting. For strongly-convex functions, the next theorem shows that the same convergence guarantee as Theorem 1 can be attained.

Theorem 3. Suppose that Assumptions 1 and 3 hold with $\mu > 0$. Suppose that Algorithm 1 is run for identical data with $\max_p |t_p - t_{p+1}| \leqslant H$ for some $H \geqslant 1$ and with a stepsize $\gamma > 0$ chosen such that $\gamma \leqslant \min\left\{\frac{1}{4L\left(1+\frac{2}{M}\right)}, \frac{1}{\mu + 8L(H-1)}\right\}$. Then for any timestep t such that synchronization occurs,

$$\mathbb{E}\left[\|\hat{x}_t - x_*\|^2\right] \leqslant (1-\gamma\mu)^t\,\mathbb{E}\left[\|x_0 - x_*\|^2\right] + \frac{2\gamma\sigma_{\text{opt}}^2}{\mu M} + \frac{4\sigma_{\text{opt}}^2\gamma^2(H-1)L}{\mu}. \quad (4)$$

As a corollary, we can obtain an asymptotic convergence rate by choosing specific stepsizes γ and H.

Corollary 3. Let $a = 18\kappa t$ for some $t > 0$, let $H \leqslant t$ and choose $\gamma = \frac{1}{\mu a} \leqslant \frac{1}{9LH}$. We substitute in (4) and take $T = 18a\log a$ steps, then for $r_t \overset{\text{def}}{=} \hat{x}_t - x_*$,

$$\mathbb{E}\left[\|r_t\|^2\right] = \tilde{\mathcal{O}}\left(\|r_0\|^2 + \frac{\sigma_{\text{opt}}^2}{\mu^2 MT} + \frac{\sigma_{\text{opt}}^2\kappa(H-1)}{\mu^2 T^2}\right).$$

Substituting $H = 1 + \lfloor t/M\rfloor = 1 + \lfloor T/(18\kappa M)\rfloor$ in Corollary 3 we get an asymptotic convergence rate of $\tilde{\mathcal{O}}\left(\frac{\sigma_{\text{opt}}^2}{TM}\right)$. This preserves the rate of minibatch SGD

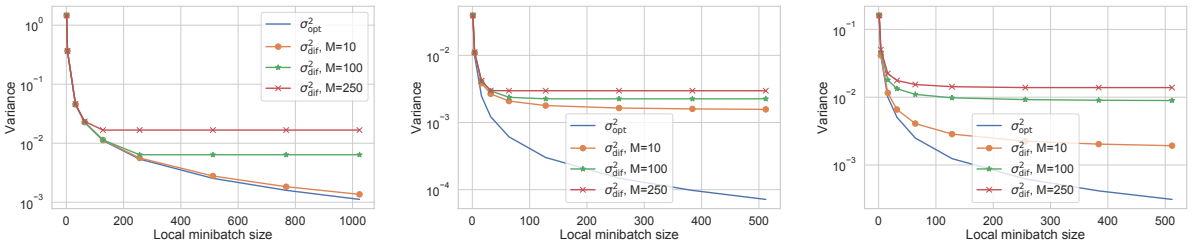

Figure 1: The effect of the dataset and number of workers M on the variance parameters. Left: 'a8a', middle: 'mushrooms', right: 'w8a' dataset. We use uniform sampling of data points, so σ^2_{opt} is the same as σ^2_{dif} with $M = 1$, while for higher values of M the value of σ^2_{dif} might be drastically larger than σ^2_{opt}.

up to problem-independent constants and polylogarithmic factors, but with possibly fewer communication steps.

Theorem 4. Suppose that Assumptions 1 and 3 hold with $\mu = 0$, that a stepsize $\gamma \leqslant \frac{1}{10LH}$ is chosen and that Algorithm 1 is run on $M \geqslant 2$ nodes with identical data and with $\sup_p |t_p - t_{p+1}| \leqslant H$, then for any timestep T such that synchronization occurs we have for $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} \hat{x}_t$ that

$$\mathbb{E}\left[f(\bar{x}_T) - f(x_*)\right] \leqslant \frac{10\|x_0 - x_*\|^2}{\gamma T} + \frac{20\gamma\sigma^2_{\text{opt}}}{M} + 40\gamma^2 L\sigma^2_{\text{opt}}\left(H-1\right). \tag{5}$$

Corollary 4. Let $H \leqslant \frac{\sqrt{T}}{\sqrt{M}}$, then for $\gamma = \frac{\sqrt{M}}{10L\sqrt{T}}$ we see that $\gamma \leqslant \frac{1}{10LH}$, and plugging it into (5) yields

$$\mathbb{E}\left[f(\bar{x}_T) - f(x_*)\right] \leqslant \frac{100L\|x_0 - x_*\|^2}{\sqrt{TM}} + \frac{2\sigma^2_{\text{opt}}}{L\sqrt{TM}} + \frac{2\sigma^2_{\text{opt}} M(H-1)}{5LT}.$$

This is the same result as Corollary 2, and hence we see that choosing $H = \mathcal{O}\left(T^{1/2}M^{-3/2}\right)$ (when $T > M^3$) yields a linear speedup in the number of nodes M.

4.2 Heterogeneous Data

We next show that for arbitrarily heterogeneous convex objectives, the convergence of Local SGD is the same as Minibatch SGD plus an error that depends on H.

Theorem 5. Suppose that Assumptions 1 and 3 hold with $\mu = 0$ and for heterogeneous data. Then for Algorithm 1 run for different data with $M \geqslant 2$, $\max_p |t_p - t_{p+1}| \leqslant H$, and a stepsize $\gamma > 0$ such that $\gamma \leqslant \min\left\{\frac{1}{4L}, \frac{1}{8L(H-1)}\right\}$, then we have

$$\mathbb{E}\left[f(\bar{x}_T) - f(x_*)\right] \leqslant \frac{4\|r_0\|^2}{\gamma T} + \frac{20\gamma\sigma^2_{\text{dif}}}{M} + 16\gamma^2 L(H-1)^2\sigma^2_{\text{dif}}.$$

where $\bar{x}_T \overset{\text{def}}{=} \frac{1}{T}\sum_{i=0}^{T-1}\hat{x}_i$ and $r_0 = x_0 - x_*$.

Dependence on σ_{dif}. We see that the convergence guarantee given by Theorem 5 shows a dependence on σ_{dif}, which measures the heterogeneity of the data distribution. In typical (non-federated) distributed learning settings where data is distributed before starting training, this term can very quite significantly depending on how the data is distributed.

Dependence on H. We further note that the dependence on H in Theorem 5 is quadratic rather than linear. This translates to a worse upper bound on the synchronization interval H that still allows convergence, as the next corollary shows.

Corollary 5. Choose $H \leqslant \frac{\sqrt{T}}{\sqrt{M}}$, then $\gamma = \frac{\sqrt{M}}{8L\sqrt{T}} \leqslant \frac{1}{8HL}$, and hence applying the result of Theorem 5,

$$\mathbb{E}\left[f(\bar{x}_T) - f(x_*)\right] \leqslant \frac{32L\|x_0 - x_*\|^2}{\sqrt{MT}} + \frac{5\sigma^2_{\text{dif}}}{2L\sqrt{MT}} + \frac{\sigma^2_{\text{dif}} M(H-1)^2}{4LT}.$$

Optimal H. By Corollary 5 we see that the optimal value of H is $H = 1 + \left\lfloor T^{1/4}M^{-3/4}\right\rfloor$, which gives $\mathcal{O}\left(\frac{1}{\sqrt{MT}}\right)$ convergence rate. Thus, the same convergence rate is attained provided that communication is more frequent compared to the identical data regime.

5 Experiments

All experiments described below were run on logistic regression problem with ℓ_2 regularization of order $\frac{1}{n}$. The datasets were taken from the LIBSVM library (Chang and Lin, 2011). The code was written in Python using MPI (Dalcin et al., 2011) and run on Intel(R) Xeon(R) Gold 6146 CPU @3.20GHz cores in parallel.

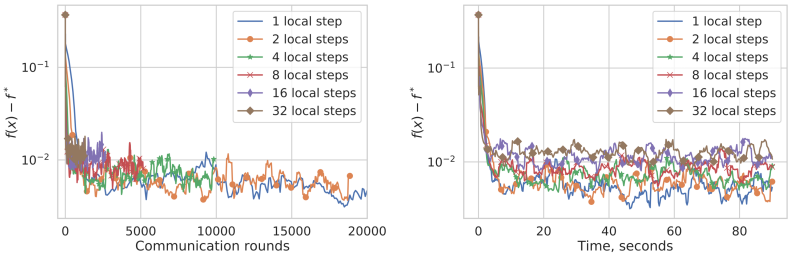

Figure 2: Results on 'a9a' dataset, with stepsize $\frac{1}{L}$. For any value of local iterations H the method converged to a neighborhood within a small number of communication rounds due to large stepsizes.

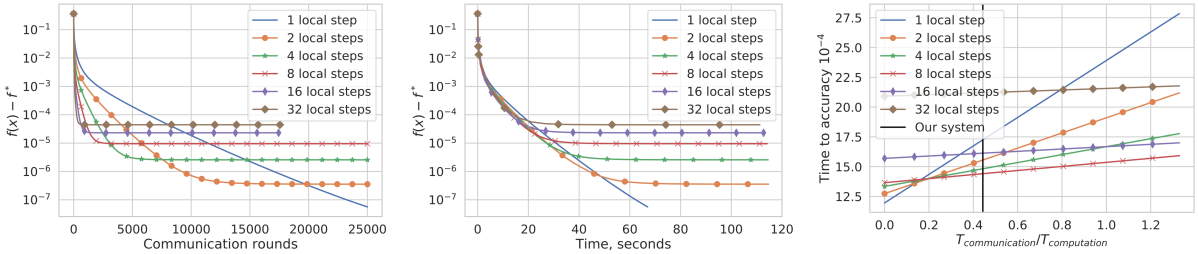

Figure 3: Convergence on heterogeneous data with different number of local steps on the 'a5a' dataset. 1 local step corresponds to fully synchronized gradient descent. Left: convergence in terms of communication rounds, which shows a clear advantage of local GD when only limited accuracy is required. Mid plot: wall-clock time might improve only slightly if communication is cheap. Right: what changes with different communication cost.

5.1 Variance measures

We provide values of σ_{dif}^2 and σ_{opt}^2 in Figure 1 for different datasets, minibatch sizes and M. The datasets were split evenly without any data reshuffling and no overlaps. For any $M > 1$, the value of σ_{dif} is lower bounded by $\frac{1}{M}\sum_{m=1}^{M} \| \nabla f_m(x_*) \|^2$ which explains the difference between identical and heterogeneous data.

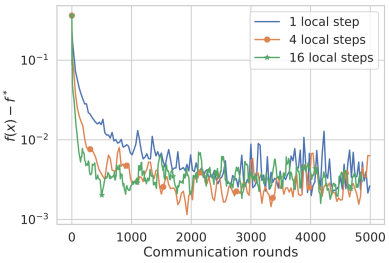

Figure 4: Convergence of local SGD on heterogeneous data with different number of local steps on the 'a5a' dataset.

5.2 Identical Data

For identical data we used $M = 20$ nodes and 'a9a' dataset. We estimated L numerically and ran two experiments, with stepsizes $\frac{1}{L}$ and $\frac{0.05}{L}$ and minibatch size equal 1. In both cases we observe convergence to a neighborhood, although of a different radius. Since we run the experiments on a single machine, the communication is very cheap and there is little gain in time required for convergence. However, the advantage in terms of required communication rounds is self-evident and can lead to significant time improvement under slow communication networks. The results are provided here in Figure 2 and in the supplementary material in Figure 5.

5.3 Heterogeneous Data

Since our architecture leads to a very specific trade-off between computation and communication, we provide plots for the case the communication time relative to gradient computation time is higher or lower. To see the impact of σ_{dif}, in all experiments we use full gradients ∇f_m and constant stepsize $\frac{1}{L}$. The data partitioning is not i.i.d. and is done based on the index in the original dataset. The results are provided in Figure 3 and in the supplementary material in Figure 6. In cases where communication is significantly more expensive than gradient computation, local methods are much faster for imprecise convergence.

References

Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-SGD: Distributed SGD with Quantization, Sparsification and Local Computations. In *Advances in Neural Information Processing Systems 32*, pages 14668–14679. 2019.

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

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing Noise in GAN Training with Variance Reduced Extragradient. In *Advances in Neural Information Processing Systems 32*, pages 391–401. Curran Associates, Inc., 2019.

Gregory F. Coppola. *Iterative parameter mixing for distributed large-margin training of structured predictors for natural language processing*. PhD thesis, University of Edinburgh, UK, 2015.

Lisandro D. Dalcin, Rodrigo R. Paz, Pablo A. Kler, and Alejandro Cosimo. Parallel distributed computing using Python. *Advances in Water Resources*, 34 (9):1124–1139, 2011.

Aymeric Dieuleveut and Kumar Kshitij Patel. Communication trade-offs for Local-SGD with large step size. In *Advances in Neural Information Processing Systems 32*, pages 13579–13590. 2019.

Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General Analysis and Improved Rates. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5200–5209, 2019.

Farzin Haddadpour and Mehrdad Mahdavi. On the Convergence of Local Descent Methods in Federated Learning. *arXiv preprint arXiv:1910.14425*, 2019.

Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Local SGD with Periodic Averaging: Tighter Analysis and Adaptive Synchronization. In *Advances in Neural Information Processing Systems 32*, pages 11080–11092, 2019.

Andrew Hard, Kanishka Rao, Rajiv Mathews, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated Learning for Mobile Keyboard Prediction. *arXiv preprint arXiv:1811.03604*, 2018.

Peng Jiang and Gagan Agrawal. A Linear Speedup Analysis of Distributed Deep Learning with Sparse and Quantized Communication. In *Advances in Neural Information Processing Systems 31*, pages 2525–2536. 2018.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational Inequalities with Stochastic Mirror-Prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic Controlled Averaging for On-Device Federated Learning. *arXiv preprint arXiv:1910.06378*, 2019.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. First Analysis of Local GD on Heterogeneous Data. *arXiv preprint arXiv:1909.04715*, 2019a.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Better Communication Complexity for Local SGD. *arXiv preprint arXiv:1909.04746v1*, 2019b.

Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated Learning: Strategies for Improving Communication Efficiency. In *NIPS Private Multi-Party Machine Learning Workshop*, 2016.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks. *arXiv preprint arXiv:1812.06127*, 2018.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the Convergence of FedAvg on Non-IID Data. In *International Conference on Learning Representations*, 2020.

Xianfeng Liang, Shuheng Shen, Jingchang Liu, Zhen Pan, Enhong Chen, and Yifei Cheng. Variance Reduced Local SGD with Lower Communication Complexity. *arXiv preprint arXiv:1912.12844*, 2019.

Tao Lin, Sebastian U. Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't Use Large Mini-batches, Use Local SGD. In *International Conference on Learning Representations*, 2020.

Olvi L. Mangasarian. Parallel Gradient Distribution in Unconstrained Optimization. *SIAM Journal on Control and Optimization*, 33(6):1916–1925, 1995.

Ryan McDonald, Keith Hall, and Gideon Mann. Distributed Training Strategies for the Structured Perceptron. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 456–464, 2010. ISBN 1-932432-65-5.

H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Ar-

cas. Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting Stochastic Extragradient. *to appear in the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. FedPAQ: A Communication-Efficient Federated Learning Method with Periodic Averaging and Quantization. *arXiv preprint arXiv:1909.13014*, 2019.

Pranay Sharma, Prashant Khanduri, Saikiran Bulusu, Ketan Rajawat, and Pramod K. Varshney. Parallel Restarted SPIDER – Communication Efficient Distributed Nonconvex Optimization with Optimal Computation Complexity. *arXiv preprint arXiv:1912.06036*, 2019.

Sebastian U. Stich. Local SGD Converges Fast and Communicates Little. In *International Conference on Learning Representations*, 2019.

Sebastian U. Stich and Sai Praneeth Karimireddy. The Error-Feedback Framework: Better Rates for SGD with Delayed Gradients and Compressed Communication. *arXiv preprint arXiv:1909.05350*, 2019.

Jianyu Wang and Gauri Joshi. Cooperative SGD: A Unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

Jianyu Wang, Vinayak Tantia, Nicolas Ballas, and Michael Rabbat. SlowMo: Improving Communication-Efficient Distributed SGD with Slow Momentum. *arXiv preprint arXiv:1910.00643*, 2019.

Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K. Leung, Christian Makaya, Ting He, and Kevin Chan. When Edge Meets Learning: Adaptive Control for Resource-Constrained Distributed Machine Learning. *arXiv:1804.05271*, 2018.

Cong Xie, Oluwasanmi Koyejo, Indranil Gupta, and Haibin Lin. Local AdaAlter: Communication-Efficient Stochastic Gradient Descent with Adaptive Learning Rates. *arXiv preprint arXiv:1911.09030*, 2019.

Hao Yu, Rong Jin, and Sen Yang. On the Linear Speedup Analysis of Communication Efficient Momentum SGD for Distributed Non-Convex Optimization. In *Proceedings of the 36th International Conference on Machine Learning*, 2019a.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel Restarted SGD with Faster Convergence and Less Communication: Demystifying Why Model Averaging Works for Deep Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33: 5693–5700, 2019b.

Chi Zhang and Qianxiao Li. Distributed Optimization for Over-Parameterized Learning. *arXiv preprint arXiv:1906.06205*, 2019.

Yuchen Zhang, John Duchi, Michael I. Jordan, and Martin J. Wainwright. Information-theoretic lower bounds for distributed statistical estimation with communication constraints. In *Advances in Neural Information Processing Systems 26*, pages 2328–2336. 2013.

Fan Zhou and Guojing Cong. On the Convergence Properties of a K-step Averaging Stochastic Gradient Descent Algorithm for Nonconvex Optimization. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.