Minibatch vs Local SGD for Heterogeneous Distributed Learning

Blake Woodworth
Toyota Technological Institute at Chicago
blake@ttic.edu

Kumar Kshitij Patel
Toyota Technological Institute at Chicago
kkpatel@ttic.edu

Nathan Srebro
Toyota Technological Institute at Chicago
nati@ttic.edu

Abstract

We analyze Local SGD (aka parallel or federated SGD) and Minibatch SGD in the heterogeneous distributed setting, where each machine has access to stochastic gradient estimates for a different, machine-specific, convex objective; the goal is to optimize w.r.t. the average objective; and machines can only communicate intermittently. We argue that, (i) Minibatch SGD (even without acceleration) dominates all existing analysis of Local SGD in this setting, (ii) accelerated Minibatch SGD is optimal when the heterogeneity is high, and (iii) present the first upper bound for Local SGD that improves over Minibatch SGD in a non-homogeneous regime.

1 Introduction

Given the massive scale of many modern machine learning models and datasets, it has become important to develop better methods for distributed training. A particularly important setting for distributed stochastic optimization/learning, and the one we will consider in this work is characterized by, (1) training data that is distributed across many parallel devices rather than centralized in a single node; (2) this data is distributed *heterogeneously*, meaning that each individual machine has data drawn from a *different* distribution; and (3) the frequency of communication between the devices is limited. The goal is to find a single consensus predictor that performs well on all the local distributions simultaneously [3, 4]. The heterogeneity of the data significantly increases the difficulty of distributed learning because the machines' local objectives may be completely different, so a perfect model for one distribution might be terrible for all the others. Limited communication between devices can make it even more challenging to find a good consensus.

One possible approach is using Minibatch Stochastic Gradient Descent (SGD). Between communications, each machine computes one large minibatch stochastic gradient using its local data; then the machines average their local minibatch gradients, yielding one extra-large minibatch gradient comprising data from all the machines' local distributions, which is used for a single SGD update. Minibatch SGD can also be accelerated to improve its convergence rate [8, 9]. This algorithm is simple, ubiquitous, and performs very successfully in a variety of settings.

However, there has recently been great interest in another algorithm, Local SGD (also known as Parallel SGD or Federated SGD) [14, 20, 27], which has been suggested as an improvement over the naïve approach of Minibatch SGD. For Local SGD, each machine independently runs SGD on its local objective and, each time they communicate, the machines average together their local iterates. Local SGD is a very appealing approach—unlike Minibatch SGD, each machine is constantly improving its local model's performance on the local objective, even when the machines are not communicating with each other, so the number of updates is decoupled from the number of communications. In addition to the intuitive benefits, Local SGD has also performed well in many applications [13, 25, 26].

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

But can we show that Local SGD is in fact better then Minibatch SGD for convex, heterogeneous objectives? Does it enjoy better guarantees, and in what settings? To answer these questions we need to analyze the performance of Local SGD and Minibach SGD.

A number of recent papers have analyzed the convergence properties of Local SGD in the heterogeneous data setting [e.g. 2, 11, 12, 22]. But, as we will discuss, none of these Local SGD guarantees can show improvement over Minibatch SGD, even without acceleration, and in many regimes they are much worse. Is this a weakness of the analysis? Can the bounds be improved to show that Local SGD is actually better than Minibatch SGD in certain regimes? Or even at least as good?

Until recently, the situation was similar also for homogeneous distributed optimization, where all the machines' data distributions are the same. There were many published analyses of Local SGD, none of which showed any improvement over Minibatch SGD, or even matched Minibatch SGD's performance. Only recently Woodworth et al. [23] settled the issue for the homogeneous case, showing that on the one hand, when communication is rare, Local SGD provably outperforms even accelerated Minibatch SGD, but on the other, Local SGD does not always match Minibatch SGD's performance guarantees and in some situations is provably *worse* than Minibatch SGD.

How does this situation play out in the more challenging, and perhaps more interesting, heterogeneous setting? Some have suggested that the more difficult heterogeneous setting is where we should expect Local SGD to really shine, and where its analysis becomes even more relevant. So, how does heterogeneity affect both Local SGD and Minibatch SGD, and the comparison between them? Do any existing analyses of Local SGD show improvement over Minibatch SGD in the heterogeneous setting? Is Local SGD still better than Minibatch SGD when communication is rare? Does the added complexity of heterogeneity perhaps necessitate the more sophisticated Local SGD approach, as some have suggested? What is the optimal method in this more challenging setting?

In fact, Karimireddy et al. [11] recently argued that heterogeneity can be particularly problematic for Local-SGD, proving a lower bound for it that shows degradation as heterogeneity increases. In Section 4, we discuss how this lower bound implies that Local SGD is strictly *worse* than Minibatch SGD when the level of heterogeneity is very large. However, even with Karimireddy et al.'s lower bound, it is not clear whether or not Local SGD can improve over Minibatch SGD for merely moderately heterogeneous objectives.

In this paper, we expand on Karimireddy et al.'s observation about the ill-suitability of Local SGD to the heterogeneous setting. We prove that existing analysis of Local SGD for heterogeneous data cannot be substantially improved unless the setting is very near-homogeneous. This is disappointing for Local SGD, because it indicates that unless the level of heterogeneity is very low, the performance it can ensure is truly worse than Minibatch SGD, even without acceleration and regardless of the frequency of communication. At the same time, we provide a more refined analysis of Local SGD showing that Local SGD does, in fact, improve over Minibatch SGD when the level of heterogeneity is sufficiently small (i.e. the problem is not exactly homogeneous, but it is at least near-homogeneous). This is the first result to show that Local SGD improves over Minibatch SGD in *any* non-homogeneous setting.

We further show that with even moderately high heterogeneity (or when we do not restrict the dissimilarity between machines), Accelerated Minibatch SGD is in fact optimal for heterogeneous stochastic distributed optimization! This is because, as we show, Minibatch SGD and its accelerated variant are immune to the heterogeneity of the problem. Perhaps the most important conclusion of our study is that we identify a regime, in which the data is moderately heterogeneous, where Accelerated Minibatch SGD may *not* be optimal and where Local SGD is certainly worse than Minibatch SGD, and so new methods may be needed.

2 Setup

We consider heterogeneous distributed stochastic convex optimization/learning using M machines, each of which has access to its own data distribution $\mathcal{D}^m$. The goal is to find an approximate minimizer of the average of the local objectives:

$$\min_{x\in\mathbb{R}^d} F(x) := \frac{1}{M}\sum_{m=1}^{M} F_m(x) := \frac{1}{M}\sum_{m=1}^{M} \mathop{\mathbb{E}}_{z^m\sim\mathcal{D}^m} f(x;z^m) \tag{1}$$

This objective captures, for example, supervised learning where $z = (z_{\text{features}}, z_{\text{label}})$ and $f(x; z)$ is the loss of the predictor, parametrized by x, on the instance z. The per-machine distribution $\mathcal{D}^m$ can be thought of as the empirical distribution of data on machine m, or as source distribution which varies between servers, regions or devices.

We focus on a setting in which each machine performs local computations using samples from its own distribution, and is able to communicate with other machines periodically in order to build consensus. This situation arises, for example, when communication is expensive relative to local computation, so it is advantageous to limit the frequency of communication to improve performance.

Concretely, we consider distributed first-order algorithms where each machine computes a total of T stochastic gradients, and is limited to communicate with the others R times. We divide the optimization process into R rounds, where each round consists of each machine calculating and processing $K = T/R$ stochastic gradients, and then communicating with all other machines[1]. Each stochastic gradient for machine m is given by $\nabla f(x; z^m)$ for an independent $z^m \sim \mathcal{D}^m$.

A simple algorithm for this setting is **Minibatch SGD**. During each round, each machine computes K stochastic gradients $\{g^m_{r,k}\}_{k\in[K]}$ at the same point x_r, and communicates its average g^m_r. Then, the averages from all machines are averaged, to obtain an estimate g_r of the gradient of the overall objective. The estimate g_r is based on all KM stochastic gradients, and is used to obtain the iterate x_{r+1}. Overall, we perform R steps of SGD, each step based on a mini-batch of KM (non-i.i.d.) samples. To summarize, initializing at some x_0, Minibatch SGD operates as follows,

$$\begin{aligned} g^m_{r,k} &= \nabla f(x_r; z^m_{r,k}), \;\; z^m_{r,k} \sim \mathcal{D}^m, \;\; m = 1 \dots M, k = 0 \dots K-1, \\ g_r &= \frac{1}{M}\sum_{m=1}^{M} g^m_r \quad \text{where } g^m_r = \frac{1}{K}\sum_{k=1}^{K} g^m_{r,k}, \\ x_{r+1} &= x_r - \eta_r g_r, \end{aligned} \tag{2}$$

Alternatively, we can perform the same stochastic gradient calculation and aggregation of g_r but replace the simple SGD updates with more sophisticated updates and carefully tuned momentum parameters. Throughout this paper, we use "**Accelerated Minibatch SGD**" to refer to two accelerated variants of SGD: AC-SA [8] for convex objectives, and multi-stage AC-SA [9] for strongly convex objectives. Algorithmic details including pseudo-code are provided in Appendix C.2.

Unlike Minibatch SGD, where each machine spends an entire round calculating stochastic gradients at the same point, **Local SGD** allows the machines to update local iterates throughout the round based on their own stochastic gradient estimates. Each round starts with a common iterate on all machines, then each machine executes K steps of SGD on its own local objective and communicates the final iterate. These iterates are averaged to form the starting point for the next round. Using $x^m_{r,k}$ to denote the iterate after r rounds and k local steps on machine m and initializing at $x^m_{0,0} = x_0$ for all $m \in [M]$, Local SGD operates as follows,

$$\begin{aligned} g^m_{r,k} &:= \nabla f(x^m_{r,k}; z^m_{r,k}), \;\; z^m_{r,k} \sim \mathcal{D}^m, \;\; m = 1 \dots M, k = 0 \dots K-1, \\ x^m_{r,k+1} &= x^m_{r,k} - \eta_{r,k} g^m_{r,k} \\ x^m_{r+1,0} &= x_{r+1} := \tfrac{1}{M}\sum\nolimits_{m=1}^{M} x^m_{r,K}. \end{aligned} \tag{3}$$

An alternative viewpoint: reducing communication. Another way of viewing the problem is as follows: consider as a baseline processing T stochastic gradient on each machine, but communicating at every step, i.e. T times, thus implementing T steps of SGD, using a mini-batch of M samples (one from each machine). Can we achieve the same performance as this baseline while communicating less frequently? Communicating only R times instead of T precisely brings us back to the model we are considering, and all the methods discussed above (R steps of MB-SGD with mini-batches of size $KM = TM/R$, or Local SGD with T steps per machine and R averaging steps) reduce the communication. Checking that $R < T$ rounds achieve the same accuracy as the dense communication baseline is a starting point, but the question is how small can we push R (while keeping $T = KR$ fixed) before accuracy degrades. Better error guarantees (in terms of K, M and R) mean we can use a smaller R with less degradation, and the smallest R with no asymptotic degradation can be directly calculated from the error guarantee [see, e.g., discussion in 6].

[1] Variable-length rounds of communication are also possible, but do not substantially change the picture.

In order to prove convergence guarantees, we rely on several assumptions about the problem. Central to our analysis will be a parameter ζ_*^2 which, in some sense, describes the level of heterogeneity in the problem. It is possible to analyze heterogeneous distributed optimization without any bound on the relatedness of the local objective—this is the typical setup in the consensus optimization literature [e.g. 4, 15, 16, 19] and indeed our analysis of Minibatch SGD does not rely on this parameter. Nevertheless, such an assumption *is* required by the existing analyses of Local SGD [2, 11, 12] which we would like to compare to, and, as we will show, is in fact necessary for the convergence of Local SGD. Following prior work [11, 12], we define

$$\zeta_*^2 = \tfrac{1}{M}\sum_{m=1}^{M}\left\|\nabla F_m(x^*)\right\|^2 \tag{4}$$

Since $\frac{1}{M}\sum_{m=1}^{M}\nabla F_m(x^*) = \nabla F(x^*) = 0$, this captures, in some sense, the variation in the local gradients *at the optimum*. When $\zeta_*^2 = 0$, all F_m share at least one minimizer, and when ζ_*^2 is large, there is great disagreement between the local objectives. While homogeneous objectives (i.e. $F_m = F$) have $\zeta_*^2 = 0$, the converse is not true! Even when $\zeta_*^2 = 0$, $\nabla F_m(x)$ might be different than $\nabla F_n(x)$ for $x \neq x^*$.

Throughout, we assume that $f(\cdot; z)$ is H**-smooth** for all z meaning

$$f(y;z) \leq f(x;z) + \langle \nabla f(x;z),\, y-x\rangle + \tfrac{H}{2}\|x-y\|^2 \qquad \forall_{x,y,z}, \tag{5}$$

We assume the **variance of the stochastic gradients** on each machine is bounded, either uniformly

$$\mathbb{E}_{z^m\sim\mathcal{D}^m}\left\|\nabla f(x;z^m) - \nabla F_m(x)\right\|^2 \leq \sigma_m^2 \quad \forall_{x,m} \quad \text{and} \quad \sigma^2 = \tfrac{1}{M}\sum_{m=1}^{M}\sigma_m^2 \tag{6}$$

or **only at the optimum** x^*, i.e.

$$\mathbb{E}_{z^m\sim\mathcal{D}^m}\left\|\nabla f(x^*;z^m) - \nabla F_m(x^*)\right\|^2 \leq \sigma_{*,m}^2. \quad \forall_m \quad \text{and} \quad \sigma_*^2 = \tfrac{1}{M}\sum_{m=1}^{M}\sigma_{*,m}^2 \tag{7}$$

We consider guarantees of two forms: For **strongly convex local objectives**, we consider guarantees that depend on the **parameter of strong convexity**, λ, for all the machines' objectives,

$$F_m(y) \geq F_m(x) + \langle \nabla F_m(x),\, y-x\rangle + \tfrac{\lambda}{2}\|x-y\|^2 \qquad \forall_{x,y,m} \tag{8}$$

as well as the **initial sub-optimality** $F(0) - F(x^*) \leq \Delta$, besides the smoothness, heterogeneity bound, and variance as discussed above.

We also consider guarantees for **weakly convex objectives** that just rely on each local objective F_m being convex (not necessarily strongly), as well as a bound on the norm of the optimum $\|x^*\| \leq B$, besides the smoothness, homogeneity bound, and variance as discussed above.

3 Minibatch SGD and Accelerated Minibatch SGD

To begin, we analyze the worst-case error of Minibatch and Accelerated Minibatch SGD in the heterogeneous setting. A simple observation is that, despite the heterogeneity of the objective, the minibatch gradients g_r (2) are unbiased estimates of ∇F, i.e. the overall objective's gradient:

$$\mathbb{E}g_r = \mathbb{E}\left[\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}\nabla f(x_r; z_{r,k}^m)\right] = \frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}\nabla F_m(x_r) = \nabla F(x_r) \tag{9}$$

We are therefore updating using unbiased estimates of ∇F, and can appeal to standard analysis for (accelerated) SGD. To do so, we calculate the variance of these estimates:

$$\mathbb{E}\|g_r - \nabla F(x_r)\|^2 = \mathbb{E}\left\|\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}\nabla f(x_r; z_{r,k}^m) - \nabla F(x_r)\right\|^2 \tag{10}$$

$$= \frac{1}{M^2K^2}\sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla f(x_r; z_{r,k}^m) - \nabla F_m(x_r)\right\|^2 \leq \frac{\sigma^2}{MK} \tag{11}$$

Interestingly, the variance is always reduced by MK and is not effected by the heterogeneity ζ_*. Plugging this calculation[2] into the analysis of SGD (see details in Appendix C) yields:

[2] A similar calculation establishes the variance at x^* is σ_*^2/MK.

Theorem 1. *A weighted average of the Minibatch SGD iterates satisfies for a universal constant c*

$$\mathbb{E}F(\hat{x}) - F^* \leq c \cdot \left(\frac{HB^2}{R} + \frac{\sigma_* B}{\sqrt{MKR}} \right) \qquad \text{under the convex assumptions,}$$

$$\mathbb{E}F(\hat{x}) - F^* \leq c \cdot \left(\frac{H\Delta}{\lambda} \exp\left(-\frac{\lambda R}{2H} \right) + \frac{\sigma_*^2}{\lambda MKR} \right) \qquad \text{under the strongly convex assumptions.}$$

and Accelerated Minibatch SGD[3] *guarantees*

$$\mathbb{E}F(\hat{x}) - F^* \leq c \cdot \left(\frac{HB^2}{R^2} + \frac{\sigma B}{\sqrt{MKR}} \right) \qquad \text{under the convex assumptions,}$$

$$\mathbb{E}F(\hat{x}) - F^* \leq c \cdot \left(\Delta \exp\left(-\frac{\sqrt{\lambda} R}{c' \sqrt{H}} \right) + \frac{\sigma^2}{\lambda MKR} \right) \qquad \text{under the strongly convex assumptions.}$$

The most important feature of these guarantees is that they are completely independent of ζ_*^2. Since these upper bounds are known to be tight in the homogeneous case (where $\zeta_*^2 = 0$) [17, 18], this means that both algorithms are essentially immune to the heterogeneity of the problem, performing equally well for homogeneous objectives as they do for arbitrarily heterogeneous ones.

4 Local SGD for heterogeneous data

Recently, Bayoumi et al. [2] and Koloskova et al. [12] analyzed Local SGD for the heterogeneous data setting. Their guarantees in the convex case along with the analysis of (Accelerated) Minibatch SGD are summarized in Table 1, Table 2 in Appendix A shows the comparison in the strongly convex case. SCAFFOLD[4] [11], a related method for heterogeneous distributed optimization, is also included.

Upon inspection, among the previously published upper bounds for heterogeneous Local SGD (that we are aware of), Koloskova et al.'s is the tightest, and dominates the others. However, even this guarantee is the sum of the Minibatch SGD bound plus additional terms, and is thus worse in every regime, and cannot show improvement over Minibatch SGD. But does this reflect a weakness of their analysis, or a true weakness of Local SGD? Indeed, Woodworth et al. [23] showed a tighter upper bound for Local SGD *in the homogeneous case*, which improves over Koloskova et al. [12] (for $\zeta_* = 0$) and *does* show improvement over Minibatch SGD when communication is infrequent. But can we generalize Woodworth et al.'s bound also to the heterogeneous case? Optimistically, we might hope that the $(\zeta_*/R)^{2/3}$ dependence on heterogeneity in these bounds could be improved. After all, Minibatch SGD's rate is independent of ζ_*^2, so perhaps Local SGD's could be, too?

Unfortunately, it is already known that some dependence on ζ_* is necessary, as Karimireddy et al. [11] have shown a lower bound[5] of $\zeta_*^2/(\lambda R^2)$ in the strongly convex case, which suggests a lower bound of $\zeta_* B/R$ in the convex case. But perhaps the Koloskova et al. analysis can be improved to match this bound? If the $\zeta_* B/R$ term from Karimireddy et al.'s lower bound were possible, it would be lower order than HB^2/R for $\zeta_* < HB$, and we would see no slow-down until the level of heterogeneity is fairly large. On the other hand, if the $(\zeta_*/R)^{2/3}$ term from Koloskova et al. cannot be improved, then we see a slowdown as soon as $\zeta_* = \Omega(HB/R)$, i.e. even for very small ζ_*!

We now show that the poor dependence on ζ_* from the Koloskova et al. analysis cannot be improved. Consequently, for sufficiently heterogeneous data, Local SGD is strictly worse than Minibatch SGD, regardless of the frequency of communication, unless the level of heterogeneity is very small.

[3] This analysis can likely also be stated in terms of σ_*, but this does not easily follow from existing work.

[4] Karimireddy et al. [11] also analyzed a variant of heterogeneous Local SGD (refereed to as FEDAVG), that, as we discuss in Appendix B, ends up essentially equivalent to Minibatch SGD. The guarantee is thus not on the performance of Local SGD as in (3), but rather a loose upper bound for Minibatch SGD, and so we do not include it in the Tables. Karimireddy et al. consider the more general framework where only a random subset of the machines are used in each round—in the Tables here we present the analysis as it applies to our setting where all machines are used in each round. See Appendix B for more details.

[5] As stated by Karimireddy et al. [11, Theorem II], the lower bound is for their FEDAVG method, which is the same as (13) in Section 6. However, their lower bound should be qualified, since with an optimal choice of stepsize parameters the ζ_*-dependence can be avoided and the lower bound does not hold (see Section 6 and Appendix B). The more accurate statement is that their lower bound is for "traditional" Local SGD, i.e. when $\eta_{\text{inner}} = \eta_{\text{outer}}$ in the notation of Section 6, or $\eta_g = 1$ in Karimireddy et al.'s notation.

Method/Analysis	Worst-Case Error (i.e. $\mathbb{E}F(\hat{x}) - F^* \lesssim$)
Minibatch SGD Theorem 1	$\frac{HB^2}{R} + \frac{\sigma_* B}{\sqrt{MKR}}$
Accelerated Minibatch SGD Theorem 1	$\frac{HB^2}{R^2} + \frac{\sigma B}{\sqrt{MKR}}$
Local SGD Koloskova et al. [12]	$\frac{HB^2}{R} + \frac{\sigma_* B}{\sqrt{MKR}} + \frac{\left(H\zeta_*^2 B^4\right)^{1/3}}{R^{2/3}} + \frac{\left(H\sigma_*^2 B^4\right)^{1/3}}{K^{1/3}R^{2/3}}$
Local SGD Bayoumi et al. [2]	$\frac{HB^2}{R} + \frac{B\sqrt{\sigma_*^2+\zeta_*^2}}{\sqrt{MKR}} + \frac{\left(H(\sigma_*^2+\zeta_*^2) B^4\right)^{1/3}}{R^{2/3}}$
SCAFFOLD Karimireddy et al. [11]	$\frac{HB^2}{R} + \frac{\sigma B}{\sqrt{MKR}} + \frac{\zeta_*^2}{HR} + \frac{\sigma\zeta_*}{H\sqrt{MKR}}$
Local SGD Theorem 3	$\frac{HB^2}{\mathbf{K}R} + \frac{\sigma_* B}{\sqrt{MKR}} + \frac{\left(H\bar{\zeta}^2 B^4\right)^{1/3}}{R^{2/3}} + \frac{\left(H\sigma^2 B^4\right)^{1/3}}{K^{1/3}R^{2/3}}$
Local SGD Lower Bound Theorem 2	$\min\left\{\frac{HB^2}{R}, \frac{\left(H\zeta_*^2 B^4\right)^{1/3}}{R^{2/3}}\right\} + \frac{\sigma B}{\sqrt{MKR}} + \frac{\left(H\sigma^2 B^4\right)^{1/3}}{K^{2/3}R^{2/3}}$
Algorithm-Independent Lower Bound Theorem 4	$\min\left\{\frac{HB^2}{R^2}, \frac{\zeta_*^2}{HR^2}\right\} + \frac{\sigma B}{\sqrt{MKR}}$

Table 1: Guarantees under the convex assumptions. See (12) for a definition and discussion of $\bar{\zeta}$.

Theorem 2. *For any M, K, and R there exist objectives in four dimensions such that Local SGD initialized at zero and using any fixed stepsize η will have suboptimality at least*

$$\mathbb{E}F(\hat{x}) - F^* \geq c \cdot \left(\min\left\{\frac{HB^2}{R}, \frac{\left(H\zeta_*^2 B^4\right)^{1/3}}{R^{2/3}}\right\} + \frac{\left(H\sigma^2 B^4\right)^{1/3}}{K^{2/3}R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}}\right)$$

$$\mathbb{E}F(\hat{x}) - F^* \geq c \cdot \left(\min\left\{\Delta \exp\left(-\frac{6\lambda R}{H}\right), \frac{H\zeta_*^2}{\lambda^2 R^2}\right\} + \min\left\{\Delta, \frac{H\sigma^2}{\lambda^2 K^2 R^2}\right\} + \frac{\sigma^2}{\lambda MKR}\right)$$

under the convex and strongly convex assumptions (for $H \geq 16\lambda$), respectively.

This is proven in Appendix D using a similar approach to a lower bound for the homogeneous case from Woodworth et al. [23], and it is conceptually similar to the lower bounds for heterogeneous objectives of Karimireddy et al. [11]. Koloskova et al. [12] also prove a lower bound, but specifically for 1-strongly convex objectives, which obscures the important role of the strong convexity parameter.

In the convex case, this lower bound closely resembles the upper bound of Koloskova et al. [12]. Focusing on the case $H = B = \sigma^2 = 1$ to emphasize the role of ζ_*^2, the only gaps are between **(i)** a term $1/(K^{1/3}R^{2/3})$ vs $1/(K^{2/3}R^{2/3})$—a gap which also exists in the homogeneous case [23]—and **(ii)** another term $\max\left\{1/R, (\zeta_*/R)^{2/3}\right\}$ vs $\min\left\{1/R,\ (\zeta_*/R)^{2/3}\right\}$. Consequently, for $\zeta_*^2 \geq 1/R \Rightarrow (\zeta_*/R)^{2/3} \geq 1/R$, the lower bound shows that Local SGD has error at least $1/R + 1/\sqrt{MKR}$ and thus performs strictly worse than Minibatch, regardless of K. This is quite surprising—Local SGD is often suggested as an improvement over Minibatch SGD for the heterogeneous setting, yet we see that even a small degree of heterogeneity can make it much worse. Furthermore, increasing the duration of each round, K, is often thought of as more beneficial for Local SGD than Minibatch SGD, but the lower bound indicates it does little to help Local SGD in the heterogeneous setting. We make a qualitatively similar comparison for the strongly convex case in Appendix A.

The lower bounds indicate that it is not possible to radically improve over the Koloskova et al. analysis, and thus over Minibatch SGD for even moderate heterogeneity, without stronger assumptions. In order to obtain an improvement over Minibatch SGD in a heterogeneous setting, at least with very low heterogeneity, we introduce a modification to the heterogeneity measure ζ_* which bounds the difference between the local objectives' gradients everywhere (not just at x^*):

$$\sup_x \tfrac{1}{M} \sum\nolimits_{m=1}^{M} \|\nabla F_m(x) - \nabla F(x)\|^2 \leq \bar{\zeta}^2 \tag{12}$$

This quantity precisely captures homogeneity since $\bar{\zeta}^2 = 0$ if and only if $F_m = F$ (up to an irrelevant additive constant). In terms of $\bar{\zeta}^2$, we are able to analyze Local SGD and see a smooth transition from the heterogeneous ($\bar{\zeta}^2$ large) to homogeneous ($\bar{\zeta}^2 = 0$) setting.

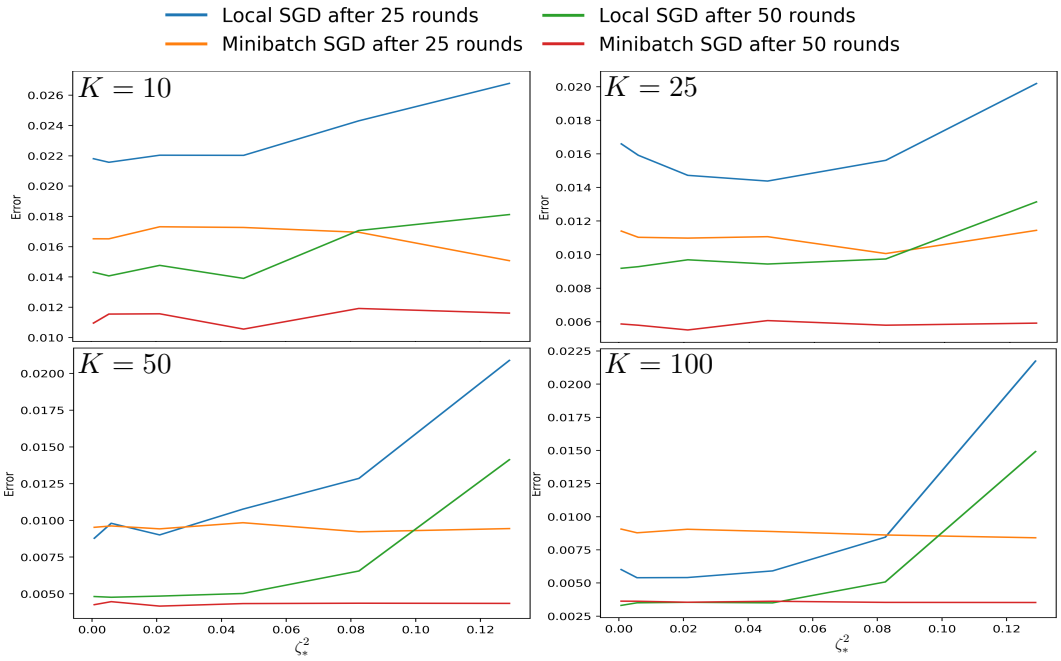

Figure 1: Binary logistic regression between even vs odd digits of MNIST. Twenty-five "tasks" were constructed, one for each combination of i vs j for even i and odd j. For $p \in \{0, 20, 40, 60, 80, 100\}$, we assigned to each of $M = 25$ machines $p\%$ data from task m, and $(100 - p)\%$ data from a mixture of all tasks. For several choices of R and K, we plot the error (averaged over four runs) versus the value of ζ_*^2 resulting from each choice of p. For both algorithms, we used the best fixed stepsize for each choice of K, R, and ζ_* individually. Additional details are provided in Appendix F.

Theorem 3. *When* $\sup_x \frac{1}{M}\sum_{m=1}^M \|\nabla F_m(x) - \nabla F(x)\|^2 \leq \bar{\zeta}^2$, *an average of the Local SGD iterates guarantees under the convex and strongly convex assumptions, respectively*

$$\mathbb{E}F(\hat{x}) - F^* \leq c \cdot \left(\frac{HB^2}{KR} + \frac{\left(H\bar{\zeta}^2 B^4\right)^{1/3}}{R^{2/3}} + \frac{\left(H\sigma^2 B^4\right)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\sigma_* B}{\sqrt{MKR}} \right),$$

$$\mathbb{E}F(\hat{x}) - F^* \leq c \cdot \left(\frac{H^2 B^2}{HKR + \lambda K^2 R^2} + \left(\frac{H\bar{\zeta}^2}{\lambda^2 R^2} + \frac{H\sigma^2}{\lambda^2 KR^2} \right) \log\left(\frac{H}{\lambda} + KR \right) + \frac{\sigma_*^2}{\lambda MKR} \right).$$

We prove the Theorem in Appendix E. This is the first analysis of Local SGD, or any other method for heterogeneous distributed optimization, showing any improvement over Minibatch SGD in any heterogeneous regime[6]. When $\bar{\zeta} = 0$, Theorem 3 reduces to the homogeneous analysis of Local SGD of Woodworth et al. [23, Theorem 2], which already showed that in that case, we see improvement when $K \gtrsim R$. Theorem 3 degrades smoothly when $\bar{\zeta} > 0$, and shows improvement for Local SGD over Minibatch SGD also when $\bar{\zeta}^2 \lesssim 1/R$ in the convex case, i.e. with low, yet positive, heterogeneity. It is yet unclear whether this rate of convergence can be ensured in terms of ζ_*^2 rather than $\bar{\zeta}^2$.

Experimental evidence Finally, while Theorem 2 proves that Local SGD is worse than Minibatch SGD unless ζ_*^2 is very small *in the worst case*, one might hope that for "normal" heterogeneous problems, Local SGD might perform better than its worst case error suggests. However, a simple binary logistic regression experiment on MNIST indicates that this behavior likely extends significantly beyond the worst case. The results, depicted in Figure 1, show that Local SGD performs worse than Minibatch SGD unless both ζ_* is very small and K is large. Finally, we also observe that Minibatch SGD's performance is essentially unaffected by ζ_* empirically as predicted by theory.

[6]Karimireddy et al. [11] establish a guarantee for SCAFFOLD that improves over Minibatch SGD in a setting where only a random subset of the machines are available in each iteration and ζ_* is sufficiently small. Here we refer to the distributed optimization setting of Section 2, where all machines are used in each iteration.

5 Accelerated Minibatch SGD is optimal for highly heterogeneous data

In the previous section, we showed that Local SGD is strictly worse than Minibatch SGD whenever $\zeta_*^2 \geq HB/R$ (in the convex case). But perhaps a different method can improve over Accelerated Minibatch SGD? After all, Accelerated Minibatch SGD's convergence rate can only partially be improved by increasing K, the amount of local computation per round of communication. Even when $K \to \infty$, Accelerated Minibatch SGD is only guaranteed suboptimality $1/R^2$ or $\exp(-\lambda R)$. Can this rate be improved? Now, we show that the answer depends on the level of heterogeneity—unless ζ_*^2 is sufficiently small, no algorithm can improve over Accelerated Minibatch SGD, which is optimal. To state the lower bound, we follow Carmon et al. [5] and define:

Definition 1 (Distributed zero-respecting algorithm)**.** *Let* $\operatorname{supp}(v) = \{i \in [d] : v_i \neq 0\}$ *and* $\pi_m(t, m')$ *be the last time before* t *when machines* m *and* m' *communicated. We say that an optimization algorithm is distributed zero-respecting if the* t*th query on the* m*th machine,* x_t^m *satisfies,*

$$\operatorname{supp}(x_t^m) \subseteq \bigcup_{s<t} \operatorname{supp}(\nabla f(x_s^m; z_s^m)) \cup \bigcup_{m' \neq m} \bigcup_{s \leq \pi_m(t,m')} \operatorname{supp}(\nabla f(x_s^{m'}; z_s^{m'})).$$

That is, the coordinates of a machine's iterates are non-zero only where gradients seen by that machine were non-zero. This encompasses many first-order optimization algorithms, including Minibatch SGD, Accelerated Minibatch SGD, Local SGD, coordinate descent methods, etc.

Theorem 4. *For any* M*,* K*, and* R*, there exist two quadratic objectives satisfying the convex and strongly convex assumptions (for* $H \geq 7\lambda$*) such that the output of any distributed zero-respecting algorithm will have suboptimality in the convex and strongly convex case respectively,*

$$F(\hat{x}) - F^* \geq c\left(\min\left\{\frac{\zeta_*^2}{HR^2}, \frac{HB^2}{R^2}\right\} + \frac{\sigma B}{\sqrt{MKR}}\right),$$
$$F(\hat{x}) - F^* \geq c\left(\min\left\{\frac{\lambda\zeta_*^2}{H^2}, \frac{\Delta\sqrt{\lambda}}{\sqrt{H}}\right\}\exp\left(-\frac{8R\sqrt{\lambda}}{\sqrt{H}}\right) + \frac{\sigma^2}{\lambda MKR}\right).$$

This is proven in Appendix G using techniques similar to those of Woodworth and Srebro [24] and Arjevani and Shamir [1]. However, care must be taken to control the parameter ζ_* for the hard instance and to see how this affects the final lower bound.

Corollary 1. *Accelerated Minibatch SGD is optimal when* $\zeta_* \geq HB$ *in the convex case, and is optimal up to log factors when* $\zeta_*^2 \geq H^{3/2}/\sqrt{\lambda}$ *in the strongly convex case.*

Thus, Accelerated Minibatch SGD is optimal for large ζ_*, but when ζ_* is smaller, the lower bound does not match, and it might be possible to improve. Indeed, focusing on the convex case, there is a substantial regime $HB/\sqrt{R} \leq \zeta_* \leq HB$ in which it may or may not be possible to improve over Accelerated Minibatch SGD. Is it possible to improve in this regime, and if so, with what algorithm? From the previous section, we know that Local SGD is certainly *not* such an algorithm. Thus, an important question for future work is **what can be done when** ζ_* **is bounded, but not insignificant,** and what new algorithms may be able to improve over Accelerated Minibatch SGD in this regime?

6 Inner and outer stepsizes

In understanding the relationship between Minibatch SGD and Local SGD, and thinking of how to improve over them, it is useful to consider a unified method that interpolates between them. This involves taking SGD steps locally with one stepsize, and then when the machines communicate, they take a step in the resulting direction with a second, different stepsize. Such a dual-stepsize approach was already presented and analyzed as FEDAVG by Karimireddy et al. [11]. We will refer to these two stepsizes as "inner" and "outer" stepsizes, respectively, and consider

$$\begin{aligned} x_{r,k}^m &= x_{r,k-1}^m - \eta_{\text{inner}} \nabla f(x_{r,k-1}^m; z_{r,k-1}^m) && \forall_{m\in[M],k\in[K]} \\ x_{r+1,0}^m &= x_{r,0}^m - \eta_{\text{outer}} \frac{1}{M}\sum_{n=1}^{M}\sum_{k=1}^{K} \nabla f(x_{r,k-1}^m; z_{r,k-1}^m) && \forall_{m\in[M],r\in[R]} \end{aligned} \tag{13}$$

Choosing $\eta_{\text{inner}} = 0$, this is equivalent to Minibatch SGD with stepsize η_{outer}, and choosing $\eta_{\text{inner}} = \eta_{\text{outer}}$ recovers Local SGD. Therefore, when the stepsizes are chosen optimally, this algorithm is always at least as good as both Minibatch and Local SGD achieving the minimum of Theorems 1 and 3—the former by choosing $\eta_{\text{inner}} = 0$ and the latter by choosing $\eta_{\text{inner}} = \eta_{\text{outer}}$. This analysis improves over Karimireddy et al.'s analysis of FEDAVG as we discuss in Appendix B. An important question is whether this method might be able to *improve* over both alternatives by choosing $0 \ll \eta_{\text{inner}} \ll \eta_{\text{outer}}$. Unfortunately, existing analysis does not resolve this question.

7 Using a Subset of Machines in Each Round

An interesting modification of our problem setting, recently formulated and studied by Karimireddy et al. [11], is a case in which only a random subset of $S \leq M$ machines are used in each round, as is the typical case in Federated Learning [10]. In the homogeneous case this makes no difference, since all machines are identical anyway. However, in a heterogeneous setting, having all machines participate at each round means that at each round we can obtain estimates of *all* components of the objective (1), which is very different from only seeing a different $S < M$ components each round, while still wanting to minimize the average of all M components.

We could still consider Minibatch SGD in this setting, but in each round we would thus use a gradient estimate based on SK stochastic gradients, K from each of S machines chosen uniformly at random (without replacement). This is still an unbiased estimator of the overall gradient $\nabla F(x_r)$, with variance $\frac{\sigma^2}{SK} + (1 - \frac{S}{M})\frac{\bar{\zeta}^2}{S}$ (and at x^* the variance is $\frac{\sigma_*^2}{SK} + (1 - \frac{S}{M})\frac{\zeta_*^2}{S}$), guaranteeing the following:

$$\mathbb{E}F(\hat{x}) - F^* \leq O\left(\frac{HB^2}{R} + \frac{\sigma_* B}{\sqrt{SKR}} + \sqrt{1 - \frac{S}{M}} \cdot \frac{\zeta_* B}{\sqrt{SR}}\right), \tag{14}$$

$$\mathbb{E}F(\hat{x}) - F^* \leq O\left(\lambda B^2 \exp\left(\frac{-\lambda R}{H}\right) + \frac{\sigma_*^2}{\lambda SKR} + \left(1 - \frac{S}{M}\right) \cdot \frac{\zeta_*^2}{\lambda SR}\right). \tag{15}$$

Similar guarantees are also available for Accelerated Minibatch SGD with σ and $\bar{\zeta}$ replacing σ_* and ζ_*. The guarantees (14) and (15) are valid also for Minibatch SGD (i.e. with $\eta_{\text{inner}} = 0$) and thus for the inner/outer stepsize updates (13). These guarantees improve over what was shown for the inner/outer stepsize variant by Karimireddy et al., but they also present SCAFFOLD which shows benefits over Minibatch SGD in some regimes under this setting (see discussion in Appendix B)

8 Discussion

Despite extensive efforts to analyze Local SGD both in homogeneous and heterogeneous settings, nearly all efforts have failed to show that Local SGD improves over Minibatch SGD in any situation. In fact, as far as we are aware, only Woodworth et al. [23] were able to show any improvement over Minibatch SGD in any regime, and their work is confined to the easier homogeneous setting. This raised the question of what happens in the heterogeneous case, which is substantially more difficult than the homogeneous setting, and where it is plausibly necessary to deviate from the quite naïve approach of (Accelerated) Minibatch SGD. In this paper, we conducted a careful analysis and comparison of Local and Minibatch SGD in the heterogeneous case and showed that moving from the homogeneous to heterogeneous setting significantly harms the performance of Local SGD. For instance, in the convex case Local SGD is strictly worse than Minibatch SGD unless $\zeta_*^2 < H^2B^2/R$. This indicates that any benefits of Local over Minibatch SGD actually lie in the homogeneous or near-homogenous setting, and not in highly hetrogenous settings (unless additional assumptions are made). We also show similar benefits for Local SGD as those shown by Woodworth et al., extending them to a near-homogeneous regime, whenever $\bar{\zeta}^2$ is bounded and less than $1/R$. This is the first analysis that shows any benefit for Local SGD over Minibatch SGD in any non-homogeneous regime. At the other extreme, with high heterogeneity $\zeta_* > HB$, we show that Accelerated Mini-Batch SGD is optimal. But this leaves open the regime of medium heterogeneity, of $HB/\sqrt{R} \ll \zeta_* \ll HB$, where it might be possible to improve over (Accelerated) Minibatch SGD, but *not* using Local SGD—can other methods be devised for this regime?

Broader impact

Distributed optimization is an important problem with significant implications for the success and feasibility of large scale machine learning. In recent years, the optimization community has expended significant efforts into trying to analyze and understand a seemingly simple and appealing algorithm—Local SGD. The recent work of Woodworth et al. [23] shows that much of these efforts were in vain, resulting in guarantees that do not show improvement over the familiar Minibatch SGD. However, their analysis was limited to the homogeneous case. Understanding distributed learning for heterogeneous data is important as the setting arises quite naturally, for example, when different servers or devices have data with different characteristics. Our analysis can provide clarity for the field by highlighting (i) the important regimes and important questions to focus on, (ii) what types of guarantees can help progress, and (iii) what regimes require new algorithms to be developed. Such improved focus can significantly improve and accelerate development, which in turn can have significant practical implications wherever distributed learning is used (which is increasingly everywhere).

As with many other distributed optimization papers, we focus here on reaching consensus between the different machines. However, in many, if not most, heterogeneous data settings, this could be undesirable, since insisting on a single solution (e.g. single predictor or single model) for different distributions may be sub-optimal, and may also perform very poorly on outlier distributions, which could disproportionately affect atypical users, minority groups or groups for which less training data is available. Therefore, it may sometimes be preferable to take a different, pluralistic approach to distributed learning, where different predictors are learned for each individual distribution, but where commonalities are leveraged to improve performance. This can be more fair and/or more efficient. Although in this paper we focus on achieving consensus, in part so as to align it with existing literature, the ideas developed here can also be applicable in studying pluralistic approaches.

Acknowledgments and Disclosure of Funding

This work is partially supported by NSF/BSF award 1718970, NSF-DMS 1547396, and a Google Faculty Research Award. BW is supported by a Google PhD Fellowship.

References

[1] Yossi Arjevani and Ohad Shamir. Communication complexity of distributed convex learning and optimization. In *Advances in neural information processing systems*, pages 1756–1764, 2015.

[2] Ahmed Khaled Ragab Bayoumi, Konstantin Mishchenko, and Peter Richtarik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529, 2020.

[3] Dimitri P Bertsekas and John N Tsitsiklis. *Parallel and distributed computation: numerical methods*, volume 23. Prentice hall Englewood Cliffs, NJ, 1989.

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[5] Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *arXiv preprint arXiv:1710.11606*, 2017. URL `https://arxiv.org/abs/1710.11606`.

[6] Andrew Cotter, Ohad Shamir, Nati Srebro, and Karthik Sridharan. Better mini-batch algorithms via accelerated gradient methods. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1647–1655. Curran Associates, Inc., 2011. URL `http://papers.nips.cc/paper/4432-[]better-[]mini-[]batch-[]algorithms-[]via-[]accelerated-[]gradient-[]methods.pdf`.

[7] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

[8] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[9] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, ii: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 23(4):2061–2089, 2013.

[10] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning, 2019.

[11] 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.

[12] Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian U Stich. A unified theory of decentralized sgd with changing topology and local updates. *arXiv preprint arXiv:2003.10422*, 2020.

[13] Tao Lin, Sebastian U Stich, Kumar Kshitij Patel, and Martin Jaggi. Don't use large mini-batches, use local sgd. *arXiv preprint arXiv:1808.07217*, 2018.

[14] H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, et al. Communication-efficient learning of deep networks from decentralized data. *arXiv preprint arXiv:1602.05629*, 2016.

[15] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

[16] Angelia Nedic, Asuman Ozdaglar, and Pablo A Parrilo. Constrained consensus and optimization in multi-agent networks. *IEEE Transactions on Automatic Control*, 55(4):922–938, 2010.

[17] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[18] Yurii Nesterov. Introductory lectures on convex optimization: a basic course. 2004.

[19] S Sundhar Ram, Angelia Nedić, and Venugopal V Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Journal of optimization theory and applications*, 147(3):516–545, 2010.

[20] Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018. URL `https://arxiv.org/abs/1805.09767`.

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

[22] 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.

[23] Blake Woodworth, Kumar Kshitij Patel, Sebastian U Stich, Zhen Dai, Brian Bullins, H Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? *arXiv preprint arXiv:2002.07839*, 2020.

[24] Blake E Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3639–3647. Curran Associates, Inc., 2016. URL `http://papers.nips.cc/paper/6058-[]tight-[]complexity-[]bounds-[]for-[]optimizing-[]composite-[]objectives.pdf`.

[25] Jian Zhang, Christopher De Sa, Ioannis Mitliagkas, and Christopher Ré. Parallel sgd: When does averaging help? *arXiv preprint arXiv:1606.07365*, 2016.

[26] Fan Zhou and Guojing Cong. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3219–3227. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/447. URL `https://doi.org/10.24963/ijcai.2018/447`.

[27] Martin Zinkevich, Markus Weimer, Lihong Li, and Alex J Smola. Parallelized stochastic gradient descent. In *Advances in neural information processing systems*, pages 2595–2603, 2010.

Method/Analysis	Worst-Case Error (i.e. $\mathbb{E}F(\hat{x}) - F^* \lesssim$)
Minibatch SGD Theorem 1	$\frac{H\Delta}{\lambda}\exp\left(\frac{-\lambda R}{H}\right) + \frac{\sigma_*^2}{\lambda MKR}$
Accelerated Minibatch SGD Theorem 1	$\Delta\exp\left(\frac{-\sqrt{\lambda}R}{\sqrt{H}}\right) + \frac{\sigma^2}{\lambda MKR}$
Local SGD Koloskova et al. [12]	$\frac{\sigma_*^2}{\lambda MKR} + \frac{H\zeta_*^2}{\lambda^2R^2} + \frac{H\sigma_*^2}{\lambda^2KR^2}$ for $R \geq \tilde{\Omega}\left(\frac{H}{\lambda}\log\Delta\right)$
SCAFFOLD Karimireddy et al. [11]	$\left(H\Delta + \frac{\lambda\zeta_*^2}{H^2}\right)\exp\left(\frac{-\lambda R}{H}\right) + \frac{\sigma^2}{\lambda MKR}$
Local SGD Theorem 3	$\frac{HB^2}{HKR+\lambda K^2R^2} + \frac{\sigma_* B}{\sqrt{MKR}} + \frac{H\bar{\zeta}^2}{\lambda^2R^2} + \frac{H\sigma^2}{\lambda^2KR^2}$
Local SGD Lower Bound Theorem 2	$\min\left\{\Delta\exp\left(\frac{-\lambda R}{H}\right), \frac{H\zeta_*^2}{\lambda^2R^2}\right\} + \frac{\sigma^2}{\lambda MKR} + \min\left\{\Delta, \frac{H\sigma^2}{\lambda^2K^2R^2}\right\}$
Algorithm-Independent Lower Bound Theorem 2	$\min\left\{\frac{\Delta\sqrt{\lambda}}{\sqrt{H}}, \frac{\lambda\zeta_*^2}{H^2}\right\}\exp\left(-\frac{\sqrt{\lambda}R}{\sqrt{H}}\right) + \frac{\sigma^2}{\lambda MKR}$

Table 2: Guarantees under the strongly convex assumptions, with log factors omitted. See (12) for a definition and discussion of $\bar{\zeta}$.

A Discussion of Local SGD lower bound in strongly convex setting

In the strongly convex case, the lower bound from Theorem 2 nearly matches the upper bound of Koloskova et al.. Focusing on the case $H = B = \sigma = 1$ in order to emphasize the role of ζ_*, the only differences are between a term $1/(KR^2)$ versus $1/(K^2R^2)$—a gap which also exists in the homogeneous case [23]—and between $\exp(-\lambda R) + \zeta_*^2/(\lambda^2R^2)$ and $\min\{\exp(-\lambda R), \zeta_*^2/(\lambda^2R^2)\}$. The latter gap is somewhat more substantial than the convex case, but nevertheless indicates that the $\zeta_*^2/(\lambda^2R^2)$ rate cannot be improved until the number of rounds of communication is at least the condition number (or ζ_*^2 is very large).

B Discussion of Karimireddy et al. [11]

We compare our results to Karimireddy et al. [11], who presented an analysis of the inner/outer stepsize variant of Section 6 (as FEDAVG, with a different stepsize parametrization—see below) as well as the novel method SCAFFOLD which incorporates variance reduction.

Karimireddy et al. [11, Theorem V] show that for the inner/outer stepsize updates (13), with optimal choice of stepsizes, in the weakly convex case

$$\mathbb{E}F(\hat{x}) - F^* \leq O\left(\frac{HB^2}{R} + \frac{\sigma B}{\sqrt{SKR}} + \frac{\left(H\zeta_*^2B^4\right)^{1/3}}{R^{2/3}} + \sqrt{1-\frac{S}{M}}\cdot\frac{\zeta_* B}{\sqrt{SR}}\right) \tag{16}$$

and in the strongly convex case (for $R \geq \Omega\left(\frac{H}{\lambda}\right)$)

$$\mathbb{E}F(\hat{x}) - F^* \leq O\left(\lambda B^2\exp\left(\frac{-\lambda R}{H}\right) + \frac{\sigma^2}{\lambda SKR} + \frac{H\zeta_*^2}{\lambda^2R^2} + (1-\frac{S}{M})\cdot\frac{\zeta_*^2}{\lambda SR}\right). \tag{17}$$

But these are loose upper bounds: as discussed in Section 6, the Minibatch SGD guarantees also apply to the inner/outer variant (by using $\eta_{\text{inner}} = 0$). The Minibatch SGD guarantees (14) and (15) can therefor be viewed also as guarantees on the inner/outer variant (i.e. Karimireddy et al.'s FEDAVG) that improve over (16) and (17) in several ways: (a) they avoid the the third terms in (16) and (17); (b) they improve the fourth terms by a factor of $\sqrt{S}$ and S respectively; and (c) they avoid the requirement $R > H/\lambda$.

Karimireddy et al.'s presentation actually uses a different step-size parametrization that does not allow for $\eta_{\text{inner}} = 0$: they use $\eta_l = \eta_{\text{inner}}$ and $\eta_g = \eta_{\text{outer}}/\eta_{\text{inner}}$. We prefer the presentation using

η_{inner} and η_{outer} in order to emphasize the relationship with Minibatch SGD and in order to explicitly allow $\eta_{\text{inner}} = 0$. But in any case, even using their parametrization, $\eta_l = \eta_{\text{inner}}$ could be taken arbitrarily close to zero making the deviation from Minibatch SGD arbitrarily small. Indeed, the Karimireddy et al.'s bounds (16) and (17) are obtained when η_{inner} is already so small that the algorithm is essentially equivalent to Minibatch SGD.

But Karimireddy et al.'s main contribution was the presentation of SCAFFOLD, which incorporates machine-specific control iterates that reduce the inter-machine variances. For SCAFFOLD, [11, Theorem VII] show that in the weakly convex case[7]:

$$\mathbb{E}F(\hat{x}) - F^* \leq O\left(\frac{HB^2}{R} + \frac{\sigma B}{\sqrt{SKR}} + \frac{M\zeta_*^2}{HSR} + \frac{\sigma\zeta_*\sqrt{M}}{HS\sqrt{KR}}\right), \tag{18}$$

and in the strongly convex case (for $R \geq \max\{\frac{H}{\lambda}, \frac{M}{S}\}$)

$$\mathbb{E}F(\hat{x}) - F^* \leq O\left(\lambda\left(B^2 + \frac{M\zeta_*^2}{SH^2}\right)\exp\left(-\min\left\{\frac{\lambda}{H}, \frac{S}{M}\right\}R\right) + \frac{\sigma^2}{\lambda SKR}\right). \tag{19}$$

These guarantees are also obtained when η_{inner} is so close to zero that this is essentially a minibatch method, in this case "Minibatch SAGA" [cf. 7].

Although SCAFFOLD is aimed specifically at the setting where a subset of machines are used in each round (i.e. $S < M$), let us first check whether it provides benefits in our "standard" setting (introduced in Sections 2), where all machines are used each round (i.e. $S = M$). In this case, the SCAFFOLD bounds (18) and (19) may improve over the loose upper bounds (16) and (17), but this is only due to the looseness in these upper bounds. The SCAFFOLD upper bounds (for $S = M$) do not actually improve over the Minibatch SGD guarantees of Theorem 1, and only include additional terms (see Tables 1 and 2). That is, the SCAFFOLD upper bounds, when $S = M$, are valid also for FEDAVG (since as discussed above, Minibatch SGD gurantees are valid also for FEDAVG) and so do not show a benefit in the setting where all machines are used each round. This is perhaps not surprising, since in this setting there is no need to reduce inter-machine variance, and so no benefit from variance reduction.

Let us turn then to the setting of Section 7, where only a random subset of $S < M$ machines are used in each iteration, and for which SCAFFOLD was developed. In this setting, SCAFFOLD does show a benefit in some regimes. Let us focus on the weakly convex case and compare the SCAFFOLD guarantee (18) with the Minibatch SGD guarantee (14). We can verify that, e.g. when $\sigma = 0$ and $\zeta_* = HB$, SCAFFOLD improves over Minibatch SGD if $\frac{M}{R} \ll \frac{S}{M} \ll \frac{R}{M}$, and $S < M$, but the SCAFFOLD guarantee (18) is worse then Minibatch SGD if $\frac{S}{M} \ll \frac{M}{R}$. More generally, the SCAFFOLD guarantee is worse than Minibatch SGD if $\sigma = 0$ and $\frac{S}{M} \ll \frac{\zeta_*^2}{H^2B^2}\min(1, \frac{M}{R})$. Also for strongly convex objectives, SCAFFOLD improves over Minibatch SGD in some regimes but the guarantee (19) is worse than Minibatch SGD in other regimes. Care is required to map out the precise regimes and how they depend on the various problem parameters.

C Proof of Theorem 1

In this Appendix, we prove Theorem 1, starting with an analysis of Minibatch SGD, and proceeding to analyze Accelerated Minibatch SGD.

C.1 Minibatch SGD for heterogeneous objectives

For the proof, we will use the following standard property of convex functions:

Lemma 1 (Co-Coercivity of the Gradient)**.** *For any H-smooth and convex F, and any x, and y,*

$$\|\nabla F(x) - \nabla F(y)\|^2 \leq H\left\langle \nabla F(x) - \nabla F(y),\, x - y\right\rangle,$$

and

$$\|\nabla F(x) - \nabla F(y)\|^2 \leq 2H(F(x) - F(y) - \langle \nabla F(y),\, x - y\rangle).$$

[7] This is different than the bound stated as [11, Theorem III], but is what was proven [11, Theorem VII, Appendix E].

Also note the following result from Stich [21], which is useful for optimizing the step-sizes.

Lemma 2 (Stich [21], Lemma 3)**.** *Consider non-negative sequences$\{r_t\}_{t\geq 0}$ and $\{s_t\}_{t\geq 0}$, which satisfy:*

$$r_{t+1} \leq (1 - a\eta_t)r_t - b\eta_t s_t + c\eta_t^2, \tag{20}$$

for non-negative step-sizes $\eta_t \leq \frac{1}{d}, \forall t$, for a parameter $d \geq a$, $d > 0$. Then there exist a choice of step-sizes η_t and averaging weights w_t, such that:

$$\frac{b}{W_T}\sum_{t=0}^{T} sw_t + ar_{T+1} \leq 32dr_0 \exp\left[-\frac{aT}{2d}\right] + \frac{36c}{aT}, \tag{21}$$

for $W_T := \sum_{t=0}^{T} w_t$.

Finally we can prove the following result for Minibatch SGD in this setting.

Theorem 5. *Under the convex assumptions, the average of the iterates of Minibatch SGD guarantees for a universal constant c,*

$$\mathbb{E}F(\hat{x}) - F^* \leq c \cdot \frac{HB^2}{R} + c \cdot \frac{\sigma_* B}{\sqrt{MKR}}.$$

Under the strongly convex assumptions, a weighted average of its iterates guarantees

$$\mathbb{E}F(\hat{x}) - F^* \leq c \cdot \frac{H\Delta}{\lambda} \exp\left[-\frac{\lambda R}{8H}\right] + c \cdot \frac{\sigma_*^2}{\lambda MKR}.$$

Proof. By the λ-strong convexity of F, where λ might be equal to zero:

$$\mathbb{E}\|x_{t+1} - x^*\|^2 = \mathbb{E}\left\|x_t - \eta_t \frac{1}{KM}\sum_{m=1}^{M}\sum_{k=1}^{K}\nabla f(x_t; z_t^{m,k}) - x^*\right\|^2, \tag{22}$$

$$= \mathbb{E}\|x_t - x^*\|^2 - 2\eta_t \mathbb{E}\left\langle \nabla F(x_t),\, x_t - x^*\right\rangle + \eta_t^2 \mathbb{E}\left\|\frac{1}{KM}\sum_{m=1}^{M}\sum_{k=1}^{K}\nabla f(x_t; z_t^{m,k})\right\|^2, \tag{23}$$

$$\leq (1 - \lambda\eta_t)\mathbb{E}\|x_t - x^*\|^2 - 2\eta_t \mathbb{E}[F(x_t) - F^*] + \eta_t^2 \mathbb{E}\left\|\frac{1}{KM}\sum_{m=1}^{M}\sum_{k=1}^{K}\nabla f(x_t; z_t^{m,k})\right\|^2. \tag{24}$$

By the H-smoothness of $f(\cdot\,; z)$, we can bound the final term with

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{KM}\sum_{m=1}^{M}\sum_{k=1}^{K}\nabla f(x_t; z_t^{m,k})\right\|^2 \\
&= \mathbb{E}\left\|\frac{1}{KM}\sum_{m=1}^{M}\sum_{k=1}^{K}\Big[\nabla f(x_t; z_t^{m,k}) - \nabla f(x^*; z_t^{m,k}) + \nabla f(x^*; z_t^{m,k})\Big]\right\|^2, && (25)\\
&\leq 2\mathbb{E}\left\|\frac{1}{KM}\sum_{m=1}^{M}\sum_{k=1}^{K}\Big[\nabla f(x_t; z_t^{m,k}) - \nabla f(x^*; z_t^{m,k})\Big]\right\|^2 \\
&\quad + 2\mathbb{E}\left\|\frac{1}{KM}\sum_{m=1}^{M}\sum_{k=1}^{K}\nabla f(x^*; z_t^{m,k})\right\|^2, && (26)\\
&\leq \frac{2}{KM}\sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla f(x_t; z_t^{m,k}) - \nabla f(x^*; z_t^{m,k})\right\|^2 \\
&\quad + 2\mathbb{E}\left\|\frac{1}{KM}\sum_{m=1}^{M}\sum_{k=1}^{K}\nabla f(x^*; z_t^{m,k}) - \nabla F_m(x^*)\right\|^2, && (27)\\
&\leq \frac{4H}{KM}\sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{E}\Big[f(x_t; z_t^{m,k}) - f(x^*; z_t^{m,k}) - \Big\langle\nabla f(x^*; z_t^{m,k}),\, x_t - x^*\Big\rangle\Big] + \frac{2\sigma_*^2}{MK}, && (28)\\
&= 4H\mathbb{E}[F(x_t) - F^*] + \frac{2\sigma_*^2}{MK}. && (29)
\end{aligned}$$

Where, for the third inequality we used Lemma 1. Plugging this back into (24), then for $\eta_t \leq \frac{1}{4H}$,

$$\begin{aligned}
\mathbb{E}\|x_{t+1} - x^*\|^2 &\leq (1 - \lambda\eta_t)\mathbb{E}\|x_t - x^*\|^2 - 2\eta_t(1 - 2H\eta_t)\mathbb{E}[F(x_t) - F^*] + \frac{2\eta_t^2\sigma_*^2}{MK}, && (30)\\
&\leq (1 - \lambda\eta_t)\mathbb{E}\|x_t - x^*\|^2 - \eta_t\mathbb{E}[F(x_t) - F^*] + \frac{2\eta_t^2\sigma_*^2}{MK}, && (31)\\
\mathbb{E}[F(x_t) - F^*] &\leq \left(\frac{1}{\eta_t} - \lambda\right)\mathbb{E}\|x_t - x^*\|^2 - \frac{1}{\eta_t}\mathbb{E}\|x_{t+1} - x^*\|^2 + \frac{2\eta_t\sigma_*^2}{MK}. && (32)
\end{aligned}$$

Now we look at $\lambda = 0$ and $\lambda > 0$ separately.

Convex case ($\lambda = 0$): Choose a constant step-size,

$$\eta_t = \eta = \min\left\{\frac{1}{4H},\, \frac{B\sqrt{MK}}{\sigma_*\sqrt{T}}\right\}. \tag{33}$$

Then the averaged iterate $\bar{x}_R = \frac{1}{R}\sum_{t=1}^{R} x_t$ satisfies:

$$\begin{aligned}
\mathbb{E}F(\bar{x}_R) - F^* &\leq \frac{1}{R}\sum_{t=1}^{R}\mathbb{E}F(x_t) - F^*, && (34)\\
&\leq \frac{1}{R}\sum_{t=1}^{R}\frac{1}{\eta}\mathbb{E}\|x_t - x^*\|^2 - \frac{1}{\eta}\mathbb{E}\|x_{t+1} - x^*\|^2 + \frac{2\eta\sigma_*^2}{MK}, && (35)\\
&\leq \frac{\|x_0 - x^*\|^2}{\eta R} + \frac{2\eta\sigma_*^2}{MK}, && (36)\\
&\leq \max\left\{\frac{4HB^2}{R},\, \frac{\sigma_* B}{\sqrt{MKR}}\right\} + \frac{2\sigma_* B}{\sqrt{MKR}}, && (37)\\
&\leq \frac{4HB^2}{R} + \frac{3\sigma_* B}{\sqrt{MKR}}. && (38)
\end{aligned}$$

Strongly convex case ($\lambda > 0$): Rewriting (31),

$$\mathbb{E}\|x_{t+1} - x^*\|^2 \leq (1 - \lambda\eta_t)\mathbb{E}\|x_t - x^*\|^2 - \eta_t\mathbb{E}[F(x_t) - F^*] + \frac{2\eta_t^2\sigma_*^2}{MK}, \tag{39}$$

we note that it satisfies the conditions for Lemma 2 for the specific assignment:

$$r_t = \mathbb{E}\|x_t - x^*\|^2,\ s_t = \mathbb{E}[F(x_t) - F^*], \tag{40}$$

$$a = \lambda,\ b = 1,\ c = \frac{2\sigma_*^2}{MK},\ d = 4H,\ T = R. \tag{41}$$

Thus using Lemma 2, and applying Jensen's inequality we can guarantee the following convergence rate for the averaged iterate $\hat{x}_R = \frac{1}{W_R}\sum_{t=1}^R w_t x_t$,

$$\mathbb{E}[F(\hat{x}_R) - F^*] \leq 128H\|x_0 - x^*\|^2 \exp\left[-\frac{\lambda R}{8H}\right] + \frac{72\sigma_*^2}{\lambda MKR}, \tag{42}$$

using step-size η_t and w_t given by,

$$\begin{aligned}
&if\ R \leq \frac{4H}{\lambda}, && \eta_t = \frac{1}{4H}, && w_t = (1 - \lambda\eta_t)^{-(t+1)},\\
&if\ R > \frac{4H}{\lambda}\ and\ t < t_0, && \eta_t = \frac{1}{4H}, && w_t = 0,\\
&if\ R > \frac{4H}{\lambda}\ and\ t \geq t_0, && \eta_t = \frac{2}{\lambda(\kappa + t - t_0)}, && w_t = (\kappa + t - t_0)^2,
\end{aligned}$$

where $\kappa = \frac{8H}{\lambda}$ and $t_0 = \lceil \frac{R}{2} \rceil$. We conclude by observing that $HB^2 \leq \frac{H\Delta}{\lambda}$. □

C.2 Accelerated Minibatch SGD for heterogeneous objectives

We first recall some classical results from Ghadimi and Lan [8, 9] for accelerated variants of minibatch SGD. These results are for minimizing $F(x) := \mathbb{E}_{z\sim\mathcal{D}} f(x, z)$ where F is H-smooth and $\lambda(\geq 0)$-strongly convex. The algorithms use unbiased stochastic gradients $\{g_t\}_{t\in[T]}$, i.e., for all t, $\mathbb{E}\left[g_t(x)\right] = \nabla F(x)$ which have bounded variance for all x, i.e., $\mathbb{E}\|g_t(x) - \nabla F(x)\|^2 \leq \sigma^2$.

First consider the AC-SA algorithm Ghadimi and Lan [c.f., Sec 3.1, 8]), with step-size parameters $\{\alpha_t\}_{t\geq 1}$ and $\{\gamma_t\}_{t\geq 1}$ s.t. $\alpha_1 = 1$, $\alpha_t \in (0, 1)$ for any $t \geq 2$ and $\gamma_t > 0$ for any $t \geq 1$. The algorithm maintains three intertwined sequences $\{x_t\}$, $\{x_t^{ag}\}$, and $\{x_t^{md}\}$, updated as follows:

1. Set the initial points $x_0^{ag} = x_0 \in X$ and $t = 1$;
2. Set $x_t^{md} = \frac{(1-\alpha_t)(\lambda+\gamma_t)}{\gamma_t+(1-\alpha_t^2)\lambda} x_{t-1}^{ag} + \frac{\alpha_t[(1-\alpha_t)\mu+\gamma_t]}{\gamma_t+(1-\alpha_t^2)\lambda} x_{t-1}$;
3. Call the stochastic oracle to get the gradient g_t at the point x_t^{md};
4. Set $x_t = \arg\min_{x\in X}\left\{\alpha_t[\langle g_t,\, x\rangle + \frac{\lambda}{2}\left\|x_t^{md} - x\right\|^2] + [(1-\alpha_t)\frac{\lambda}{2} + \frac{\gamma_t}{2}]\|x_{t-1} - x\|^2\right\}$;
5. Set $x_t^{ag} = \alpha_t x_t + (1 - \alpha_t)x_{t-1}^{ag}$;
6. Set $t \leftarrow t + 1$ and go to step 1.

We have the following (almost optimal) convergence rate for strongly convex functions using AC-SA (see, Sec 3.1 in [8]).

Lemma 3. *(Ghadimi and Lan [8], Proposition 9) Let $\hat{x}^{ag}$ be computed by T steps of AC-SA using stochastic gradients of variance σ^2, then for a universal constant c,*

$$\mathbb{E}\left[F(x^{ag}) - F(x^\star)\right] \leq c \cdot \frac{H\|x_0 - x^\star\|^2}{T^2} + c \cdot \frac{\sigma^2}{\lambda T}.$$

It can be adapted to the weakly convex case by noting that, if $\tilde{F}(x) := F(x) + \frac{\lambda}{2}\|x_0 - x\|^2$ for any λ, x_0, and $x^\star = \arg\min F(x)$ then,

$$\begin{aligned}
&\min_y \mathbb{E}\left[\tilde{F}(y)\right] \leq \mathbb{E}\left[F(x^\star) + \frac{\lambda}{2}\|x_0 - x^\star\|^2\right], \\
&\Rightarrow -\mathbb{E}\left[F(x^\star)\right] \leq -\mathbb{E}\left[\min_y \tilde{F}(y)\right] + \frac{\lambda}{2}\|x_0 - x^\star\|^2, \\
&\Rightarrow \mathbb{E}\left[F(x) - F(x^\star)\right] \leq \mathbb{E}\left[\tilde{F}(x) - \min_y \tilde{F}(y)\right] + \frac{\lambda}{2}\|x_0 - x^\star\|^2, \forall x.
\end{aligned}$$

This also holds if we optimize the right hand side w.r.t. λ. In other words, a guarantee for the strongly convex case, can be converted to the weakly convex case, by regularizing with $\frac{\lambda}{2}\|x_0 - x^\star\|^2$ with optimal value of λ. This gives the following result,

Lemma 4. *Let $\hat{x}^{ag}$ be computed by T steps of AC-SA on the regularized objective $\tilde{F}(x) = F(x) + \frac{\sigma}{2\|x_0 - x^*\|\sqrt{T}}\|x - x_0\|^2$, where the stochastic gradients have variance σ^2, then for a constant c*

$$\mathbb{E}\left[F(\hat{x}^{ag}) - F(x^\star)\right] \leq c \cdot \frac{H\|x_0 - x^\star\|^2}{T^2} + c \cdot \frac{\sigma\|x_0 - x^\star\|}{\sqrt{T}}.$$

This is minimax optimal for weakly convex functions. Next we consider the multi-stage accelerated SGD algorithm Ghadimi and Lan [c.f., Sec 3, 9] which uses the above AC-SA algorithm. Let $p_0 \in X$, have bounded sub-optimality $F(p_0) - F(x^\star) \leq \Delta$, then for $k = 1, 2, \ldots,$

1. Run N_k iterations of the generic AC-SA by using $x_0 = p_{k-1}$, $\{\alpha_t\}_{t\geq 1}$, and $\{\gamma_t\}_{t\geq 1}$, with relevant definitions as follows,

$$\begin{aligned}
N_k &= \lceil \max\left\{4\sqrt{\frac{2H}{\lambda}}, \frac{128\sigma^2}{3\lambda\Delta 2^{-(k+1)}}\right\}\rceil, \\
\alpha_t &= \frac{2}{t+1}, \gamma_t = \frac{4\phi_k}{t(t+1)}, \\
\phi_k &= \max\left\{2H, \left[\frac{\lambda\sigma^2}{3\Delta 2^{-(k-1)}N_k(N_k+1)(N_k+2)}\right]^{1/2}\right\};
\end{aligned}$$

2. Set $p_k = x^{ag}_{N_k}$ where $x^{ag}_{N_k}$ is the solution obtained in the previous step.

We have following optimal rate for strongly convex functions for this algorithm,

Lemma 5. *(Ghadimi and Lan [9], Proposition 7) Let $\hat{x}^{ag}$ be computed by T steps of multi-stage AC-SA using stochastic gradients of variance σ^2, then for a universal constant c,*

$$\mathbb{E}\left[F(\hat{x}^{ag}) - F(x^\star)\right] \leq c \cdot \Delta \exp\left[-\sqrt{\frac{\lambda}{H}}T\right] + c \cdot \frac{\sigma^2}{\lambda T}.$$

Theorem 6. *Under the convex assumptions, performing AC-SA on the regularized objective $\tilde{F}(x) = F(x) + \frac{\sigma}{2B\sqrt{MKR}}\|x\|^2$ guarantees for a universal constant c*

$$\mathbb{E}F(\hat{x}) - F^* \leq c \cdot \frac{HB^2}{R^2} + c \cdot \frac{\sigma B}{\sqrt{MKR}}.$$

Under the strongly convex assumptions, the multi-stage AC-SA algorithm guarantees

$$\mathbb{E}F(\hat{x}) - F^* \leq c \cdot \Delta \exp\left[-\sqrt{\frac{\lambda}{H}}R\right] + c \cdot \frac{\sigma^2}{\lambda MKR}.$$

Proof. In order to use the previous lemmas, first note that the stochastic gradient at time t at point x is given by $\frac{1}{MK}\sum_{m=1}^M \sum_{k=1}^K \nabla f(x; z^t_{m,k})$, where $z^t_{m,k} \sim^{i.i.d.} \mathcal{D}^m$ for all machines. Fortunately,

its still an unbiased gradient estimate, i.e., $\mathbb{E}\left[\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}\nabla f(x;z_{m,k}^t)\right]=\nabla F(x)$ since the iterates on each machine are sampled i.i.d. Its variance is given by,

$$\mathbb{E}\left\|\frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}\nabla f(x;z_{m,k}^t)-\nabla F(x)\right\|^2 \tag{43}$$

$$=\frac{1}{M^2K^2}\sum_{m=1}^{M}\sum_{k=1}^{K}\mathbb{E}\left\|\nabla f(x;z_{m,k}^t)-\nabla F_m(x)\right\|^2 \tag{44}$$

$$\leq\frac{1}{M^2K^2}\sum_{m=1}^{M}\sum_{k=1}^{K}\sigma^2 \tag{45}$$

$$=\frac{\sigma^2}{MK} \tag{46}$$

Plugging this into Lemmas 4 and 5 completes the proof. □

D Proof of Theorem 2

Consider the following function $F:\mathbb{R}^4\to\mathbb{R}$:

$$F(x)=\frac{1}{2}(F_1(x)+F_2(x))=\frac{1}{2}\left(\mathbb{E}_{z^1\sim\mathcal{D}^1}f(x;z^1)+\mathbb{E}_{z^2\sim\mathcal{D}^2}f(x;z^2)\right) \tag{47}$$

The distribution $z^1\sim\mathcal{D}^1$ is described by $z^1=(1,z)$ for $z\sim\mathcal{N}(0,\sigma^2)$. Similarly, $z^2\sim\mathcal{D}^2$ is specified by $z^2=(2,z)$ for $z\sim\mathcal{N}(0,\sigma^2)$. The lower bound construction will be based on just two functions. For $M>2$ machines, we simply assign the first $\lfloor M/2\rfloor$ machines F_1 and the next $\lfloor M/2\rfloor$ machines F_2. This diminishes the lower bound by at most a $(M-1)/M$ factor. Therefore, we continue with the case $M=2$.

Following Woodworth et al. [23], we define the local functions F_1 and F_2 via the auxiliary function

$$g(x_1,x_2,x_3,z)=\frac{\mu}{2}(x_1-c)^2+\frac{H}{2}\left(x_2-\frac{\sqrt{\mu}c}{\sqrt{H}}\right)^2+\frac{H}{8}\left(x_3^2+[x_3]_+^2\right)+z^\top x_3 \tag{48}$$

$$G(x_1,x_2,x_3)=\mathbb{E}_z g(x_1,x_2,x_3,z) \tag{49}$$

where $c>0$ and $\mu\in\left[\lambda,\frac{H}{16}\right]$ are parameters to be determined later, and where $[x]_+:=\max\{x,0\}$. Then, we define

$$f(x;(1,z))=g(x_1,x_2,x_3,z)+\frac{Lx_4^2}{2}+\zeta_* x_4 \tag{50}$$

$$f(x;(2,z))=g(x_1,x_2,x_3,z)+\frac{\lambda x_4^2}{2}-\zeta_* x_4 \tag{51}$$

for a parameter $L\in[\lambda,H]$ to be determined later. Therefore,

$$F_1(x)=\mathbb{E}_{z^1\sim\mathcal{D}^1}f(x;z^1)=G(x_1,x_2,x_3)+\frac{Lx_4^2}{2}+\zeta_* x_4 \tag{52}$$

$$F_2(x)=\mathbb{E}_{z^2\sim\mathcal{D}^2}f(x;z^2)=G(x_1,x_2,x_3)+\frac{\lambda x_4^2}{2}-\zeta_* x_4 \tag{53}$$

It is clear from inspection that both F_1 and F_2, and consequently F, are H-smooth and λ-strongly convex. Furthermore, the variance of the gradients is bounded by σ^2 for both $\mathcal{D}^1$ and $\mathcal{D}^2$.

It is clear that G attains its minimum of zero at $\left[c,\frac{\sqrt{\mu}c}{\sqrt{H}},0\right]$ so $\nabla G\left(c,\frac{\sqrt{\mu}c}{\sqrt{H}},0\right)=0$, and thus

$$\nabla F\left(c,\frac{\sqrt{\mu}c}{\sqrt{H}},0,0\right)=\nabla G\left(c,\frac{\sqrt{\mu}c}{\sqrt{H}},0\right)+\left(\frac{\zeta_*}{2}-\frac{\zeta_*}{2}\right)e_4=0 \tag{54}$$

From now on, we use $x^*=\left[c,\frac{\sqrt{\mu}c}{\sqrt{H}},0,0\right]$ to denote the minimizer of F, which has norm

$$\|x^*\|^2=\left(1+\frac{\mu}{H}\right)c^2\leq 2c^2 \tag{55}$$

We can therefore ensure $\|x^*\|^2 \leq B^2$ by choosing $c^2 \leq \frac{B^2}{2}$. Furthermore, the initial suboptimality

$$F(0,0,0,0) - F^* = \mu c^2 \tag{56}$$

Therefore, we can ensure $F(0,0,0,0) - F^* \leq \Delta$ by choosing $c^2 \leq \frac{\Delta}{\mu}$. We conclude by showing that for this objective, ζ_*^2 bounded by

$$\frac{1}{2}\sum_{m=1}^{2} \|\nabla F_m(x^*)\|^2 = \|\nabla F_2(x^*)\|^2 = \|\nabla F_1(x^*)\|^2 = \zeta_*^2 \tag{57}$$

Therefore, this objective has the desired level of heterogeneity.

Therefore, we have shown that the objective satisfies all of the necessary conditions for the lower bound. All that remains is to lower bound the error of Local SGD with a constant stepsize η applied to this function.

Lemma 6. *For $\mu \leq 2L$, then Local SGD with any constant stepsize $\eta \leq \frac{1}{L}$ applied to F_1 and F_2 after being initialized at zero results in $\hat{x}_4$ such that*

$$\frac{(L+\mu)\hat{x}_4^2}{4} \geq \frac{\zeta_*^2(L+\mu)}{16\mu^2}\left(\frac{L-\mu}{L} - (1-\mu\eta)^K\right)^2 \mathbb{1}_{\left\{\eta \leq \frac{1}{L}\right\}} \mathbb{1}_{\left\{(1-\mu\eta)^K \leq \frac{L-\mu}{L}\right\}}$$

Proof. Since the coordinates of F_1 and F_2 are completely decoupled, the behavior of the fourth coordinate of the iterates can be analyzed separately from the others.

Let $x_{k,r}^{(1)}$ denote the fourth coordinate of machine 1's iterate at the kth iteration of round r, and similarly for $x_{k,r}^{(2)}$. The local SGD dynamics give

$$x_{k+1,r}^{(1)} = x_{k,r}^{(1)} - \eta\left(Lx_{k,r}^{(1)} + \zeta_*\right) = -\frac{\zeta_*}{L} + (1-L\eta)\left(x_{k,r}^{(1)} + \frac{\zeta_*}{L}\right) \tag{58}$$

$$x_{k,r}^{(2)} = x_{k,r}^{(2)} - \eta\left(-\zeta_* + \mu x_{k,r}^{(2)}\right) = \frac{\zeta_*}{\mu} + (1-\mu\eta)\left(x_{k,r}^{(2)} - \frac{\zeta_*}{\mu}\right) \tag{59}$$

and $\hat{x}_4 = \frac{1}{2}\left(x_{K,R}^{(1)} + x_{K,R}^{(2)}\right) = x_{0,R+1}$. Unravelling this recursion, we have that

$$x_{0,r+1} = x_{0,r+1}^{(1)} = x_{0,r+1}^{(2)} = \frac{1}{2}\left(\frac{\zeta_*}{\mu} - \frac{\zeta_*}{L} + (1-\mu\eta)^K\left(x_{0,r} - \frac{\zeta_*}{\mu}\right) + (1-L\eta)^K\left(x_{0,r} + \frac{\zeta_*}{L}\right)\right) \tag{60}$$

Furthermore, if $\eta \leq \frac{1}{L}$ then $(1-L\eta) \geq 0$, so if $x_{0,r} \geq 0$ then

$$x_{0,r+1} \geq \frac{\zeta_*}{2\mu} - \frac{\zeta_*}{2L} + (1-\mu\eta)^K\left(\frac{x_{0,r}}{2} - \frac{\zeta_*}{2\mu}\right) \geq \frac{\zeta_*}{2\mu}\left(\frac{L-\mu}{L} - (1-\mu\eta)^K\right) \tag{61}$$

Finally, since $x_{0,0} = 0 \geq 0$, the condition $x_{0,r} \geq 0$ will hold throughout optimization, so

$$\hat{x}_4 \geq \frac{\zeta_*}{2\mu}\left(\frac{L-\mu}{L} - (1-\mu\eta)^K\right) \tag{62}$$

Therefore, if $\eta \leq \frac{1}{L}$ and $(1-\mu\eta)^K \leq \frac{L-\mu}{L}$ then

$$\frac{(L+\mu)\hat{x}_4^2}{4} \geq \frac{\zeta_*^2(L+\mu)}{16\mu^2}\left(\frac{L-\mu}{L} - (1-\mu\eta)^K\right)^2 \tag{63}$$

This completes the proof. □

We now prove the theorem:

Theorem 2. *For any M, K, and R there exist objectives in four dimensions such that Local SGD initialized at zero and using any fixed stepsize η will have suboptimality at least*

$$\mathbb{E}F(\hat{x}) - F^* \geq c \cdot \left(\min\left\{\frac{HB^2}{R}, \frac{\left(H\zeta_*^2B^4\right)^{1/3}}{R^{2/3}}\right\} + \frac{\left(H\sigma^2B^4\right)^{1/3}}{K^{2/3}R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}}\right)$$

$$\mathbb{E}F(\hat{x}) - F^* \geq c \cdot \left(\min\left\{\Delta\exp\left(-\frac{6\lambda R}{H}\right), \frac{H\zeta_*^2}{\lambda^2R^2}\right\} + \min\left\{\Delta, \frac{H\sigma^2}{\lambda^2K^2R^2}\right\} + \frac{\sigma^2}{\lambda MKR}\right)$$

under the convex and strongly convex assumptions (for $H \geq 16\lambda$), respectively.

Proof. Since the four different coordinates are completely decoupled from each other, it suffices to analyze each coordinate separately.

In the course of proving [Theorem 3 23], Woodworth et al. prove that

$$\mathbb{E}G(\hat{x}_1,\hat{x}_2,\hat{x}_3) - G\left(c,\frac{\sqrt{\mu}c}{\sqrt{H}},0\right) \geq \frac{\mu c^2(1-\mu\eta)^{KR}}{2} + \frac{\mu c^2}{2}\mathbb{1}_{\{\eta>\frac{2}{H}\}} + \frac{H\eta^2\sigma^2}{18432}\mathbb{1}_{\{\eta\leq\frac{2}{H}\}}\mathbb{1}_{\{\eta\geq\frac{8}{HKR}\}} \tag{64}$$

Furthermore, by Lemma 6

$$\frac{(L+\lambda)\hat{x}_4^2}{4} \geq \frac{\zeta_*^2(L+\mu)}{16\mu^2}\left(\frac{L-\mu}{L} - (1-\mu\eta)^K\right)^2 \mathbb{1}_{\{\eta\leq\frac{1}{L}\}}\mathbb{1}_{\{(1-\mu\eta)^K\leq\frac{L-\mu}{L}\}} \tag{65}$$

Therefore, choosing $L = \frac{H}{2}$

$$\mathbb{E}F(\hat{x}) - F^* = \mathbb{E}G(\hat{x}_1,\hat{x}_2,\hat{x}_3) - G\left(c,\frac{\sqrt{\mu}c}{\sqrt{H}},0\right) + \frac{H+2\lambda}{8}\hat{x}_4^2 \tag{66}$$

$$\geq \frac{\mu c^2(1-\mu\eta)^{KR}}{2} + \frac{\mu c^2}{2}\mathbb{1}_{\{\eta>\frac{2}{H}\}} + \frac{H\eta^2\sigma^2}{18432}\mathbb{1}_{\{\eta\leq\frac{2}{H}\}}\mathbb{1}_{\{\eta\geq\frac{8}{HKR}\}} + \frac{\zeta_*^2(H+2\mu)}{32\mu^2}\left(\frac{H-2\mu}{H} - (1-\mu\eta)^K\right)^2 \mathbb{1}_{\{\eta\leq\frac{2}{H}\}}\mathbb{1}_{\{(1-\mu\eta)^K\leq\frac{H-2\mu}{H}\}} \tag{67}$$

Stochastic terms

First, we will show a lower bound in terms of σ^2 using solely the first three terms of (67). Consider three cases:

Case 1 $\eta \geq \frac{2}{H}$**:** In this case, from the second term of (67) we see that

$$\mathbb{E}F(\hat{x}) - F^* \geq \frac{\mu c^2}{2} \tag{68}$$

Case 2 $\frac{1}{2\mu KR} \leq \eta \leq \frac{2}{H}$**:** In this case, the third term of (67) shows

$$\mathbb{E}F(\hat{x}) - F^* \geq \frac{H\eta^2\sigma^2}{18432} \tag{69}$$

where we recalled that $\mu \leq \frac{H}{16}$, so $\eta \geq \frac{1}{2\mu KR} \geq \frac{8}{HKR}$. This is non-decreasing in η, so for any η

$$\mathbb{E}F(\hat{x}) - F^* \geq \frac{H\sigma^2}{73728\mu^2K^2R^2} \tag{70}$$

Case 3 $\eta \leq \frac{2}{H}$ **and** $\eta \leq \frac{1}{2\mu KR}$**:** In this case, from the first term of (67),

$$\mathbb{E}F(\hat{x}) - F^* \geq \frac{\mu c^2(1-\mu\eta)^{KR}}{2} \geq \frac{\mu c^2\left(1-\frac{1}{2KR}\right)^{KR}}{2} \geq \frac{\mu c^2}{4} \tag{71}$$

Combination: Combining these three cases, we conclude that for any η

$$\mathbb{E}F(\hat{x}) - F^* \geq \min\left\{\frac{\mu c^2}{2}, \frac{H\sigma^2}{73728\mu^2K^2R^2}, \frac{\mu c^2}{4}\right\} = \min\left\{\frac{\mu c^2}{3}, \frac{H\sigma^2}{73728\mu^2K^2R^2}\right\} \tag{72}$$

This lower bound holds for any stepsize, and any $\mu \in \left[\lambda, \frac{H}{16}\right]$ and regardless of ζ_*. In the strongly convex case, we recall that $F(0) - F(x^*) = \mu c^2$, therefore, we choose $\mu = \lambda$, and $c^2 = \frac{\Delta}{\lambda}$ so the lower bound reads (for a universal constant β)

$$\mathbb{E}F(\hat{x}) - F^* \geq \beta \cdot \min\left\{\Delta, \frac{H\sigma^2}{\lambda^2K^2R^2}\right\} \tag{73}$$

To conclude, it is well known that any first-order method which accesses at most MKR stochastic gradients with variance σ^2 for a λ-strongly convex objective will suffer error at least $\beta\frac{\sigma^2}{\lambda MKR}$ in the worst case [17]. Therefore, the strongly convex lower bound is

$$\mathbb{E}F(\hat{x}) - F^* \geq \beta \cdot \min\left\{\Delta,\ \frac{H\sigma^2}{\lambda^2 K^2 R^2}\right\} + \beta \cdot \frac{\sigma^2}{\lambda MKR} \tag{74}$$

In the convex case, we recall that $\|x^*\|^2 \leq 2c^2$, so we choose $c^2 = \frac{B^2}{2}$, and set $\mu = \left(\frac{H\sigma^2}{B^2K^2R^2}\right)^{1/3}$ so the lower bound reads

$$\mathbb{E}F(\hat{x}) - F^* \geq \beta \cdot \frac{\left(H\sigma^2 B^4\right)}{K^{2/3}R^{2/3}} \tag{75}$$

To conclude, it is well known that any first-order method which accesses at most MKR stochastic gradients with variance σ^2 for a convex objective with $\|x^*\| \leq B$ will suffer error at least $\beta\frac{\sigma B}{\sqrt{MKR}}$ in the worst case [17]. Therefore, the convex lower bound is

$$\mathbb{E}F(\hat{x}) - F^* \geq \beta \cdot \frac{\left(H\sigma^2 B^4\right)}{K^{2/3}R^{2/3}} + \beta \cdot \frac{\sigma B}{\sqrt{MKR}} \tag{76}$$

Heterogeneity terms

Next, we consider solely the first, second, and fourth terms of (67) in order to show a lower bound with respect to ζ_*. Again, we consider three cases:

Case 1 $\eta \geq \frac{2}{H}$**:** Again, in this case, from the second term of (67) we see that

$$\mathbb{E}F(\hat{x}) - F^* \geq \frac{\mu c^2}{2} \tag{77}$$

Case 2 $\eta \leq \frac{2}{H}$ **and** $(1-\mu\eta)^K > \frac{H-2\mu}{H}$**:** In this case, from the first term of (67), we have

$$\begin{aligned}
\mathbb{E}F(\hat{x}) - F^* &\geq \frac{\mu c^2 (1-\mu\eta)^{KR}}{2} && (78)\\
&\geq \frac{\mu c^2}{2}\left(1 - \frac{2\mu}{H}\right)^R && (79)\\
&\geq \frac{\mu c^2}{2}\left(\left(1 - \frac{4\mu}{H}\left(1 - \frac{1}{e}\right)\right)^{\frac{H}{4\mu}}\right)^{\frac{4\mu R}{H}} && (80)\\
&\geq \frac{\mu c^2}{2}\exp\left(-\frac{4\mu R}{H}\right) && (81)
\end{aligned}$$

Case 3 $\eta \leq \frac{2}{H}$ **and** $(1-\mu\eta)^K \leq \frac{H-2\mu}{H}$**:** In this case, from the first and fourth terms of (67), we have

$$\mathbb{E}F(\hat{x}) - F^* \geq \frac{\mu c^2}{2}(1-\mu\eta)^{KR} + \frac{\zeta_*^2(H+2\mu)}{32\mu^2}\left(\frac{H-2\mu}{H} - (1-\mu\eta)^K\right)^2 \tag{82}$$

Suppose that $(1-\mu\eta)^K \geq \frac{H-2\mu}{H} - \frac{1}{4R}$, then

$$\frac{\mu c^2}{2}(1-\mu\eta)^{KR} \geq \frac{\mu c^2}{2}\left(1 - \frac{2\mu}{H} - \frac{1}{4R}\right)^R \tag{83}$$

Then, if $R \geq \frac{H}{4\mu}$, then

$$\frac{\mu c^2}{2}(1-\mu\eta)^{KR} \geq \frac{\mu c^2}{2}\left(1 - \frac{3\mu}{H}\right)^R \geq \frac{\mu c^2}{2}\left(\left(1 - \frac{6\mu}{H}\left(1 - \frac{1}{e}\right)\right)^{\frac{H}{6\mu}}\right)^{\frac{6\mu R}{H}} \geq \frac{\mu c^2}{2}\exp\left(-\frac{6\mu R}{H}\right) \tag{84}$$

Otherwise, if $R \leq \frac{H}{4\mu}$, then

$$\frac{\mu c^2}{2}(1-\mu\eta)^{KR} \geq \frac{\mu c^2}{2}\left(1-\frac{1}{2R}\right)^R \geq \frac{\mu c^2}{4} \geq \frac{\mu c^2}{4}\exp\left(-\frac{6\mu R}{H}\right) \tag{85}$$

Therefore, when $(1-\mu\eta)^K \geq \frac{H-2\mu}{H} - \frac{1}{4R}$,

$$\mathbb{E}F(\hat{x}) - F^* \geq \frac{\mu c^2}{4}\exp\left(-\frac{6\mu R}{H}\right) \tag{86}$$

On the other hand, if $(1-\mu\eta)^K \leq \frac{H-2\mu}{H} - \frac{1}{4R}$, then

$$\mathbb{E}F(\hat{x}) - F^* \geq \frac{\zeta_*^2(H+2\mu)}{32\mu^2}\left(\frac{H-2\mu}{H} - (1-\mu\eta)^K\right)^2 \tag{87}$$

$$\geq \frac{\zeta_*^2(H+2\mu)}{32\mu^2}\left(\frac{1}{4R}\right)^2 \tag{88}$$

$$\geq \frac{H\zeta_*^2}{512\mu^2R^2} \tag{89}$$

Combination: Combining these three cases, we conclude that

$$\mathbb{E}F(\hat{x}) - F^* \geq \min\left\{\frac{\mu c^2}{4}\exp\left(-\frac{6\mu R}{H}\right),\ \frac{H\zeta_*^2}{512\mu^2R^2}\right\} \tag{90}$$

In the strongly convex case, we recall that $F(0) - F(x^*) = \mu c^2$, so we choose $\mu = \lambda$ and $c^2 = \frac{\Delta}{\lambda}$ so that the objective satisfies the strongly convex assumptions. Now, the lower bound reads (for a universal constant β)

$$\mathbb{E}F(\hat{x}) - F^* \geq \beta \cdot \min\left\{\Delta\exp\left(-\frac{6\lambda R}{H}\right),\ \frac{H\zeta_*^2}{512\lambda^2R^2}\right\} \tag{91}$$

In the convex case, we recall that $\|x^*\|^2 \leq 2c^2$, so we choose $c^2 = \frac{B}{2}$ so that the convex assumptions are satisfied. We now have two options, if $R \leq \frac{H^2B^2}{\zeta_*^2}$, then we pick $\mu = \left(\frac{H\zeta_*^2}{B^2R^2}\right)^{1/3}$ so that the lower bound reads

$$\mathbb{E}F(\hat{x}) - F^* \geq \beta \cdot \frac{\left(H\zeta_*^2B^4\right)^{1/3}}{R^{2/3}}\exp\left(-\frac{6\zeta_*^{2/3}R^{1/3}}{H^{2/3}B^{2/3}}\right) \tag{92}$$

$$\geq \beta \cdot \frac{\left(H\zeta_*^2B^4\right)^{1/3}}{R^{2/3}}\exp(-6) \tag{93}$$

$$\geq \beta' \cdot \frac{\left(H\zeta_*^2B^4\right)^{1/3}}{R^{2/3}} \tag{94}$$

On the other hand, if $R \geq \frac{H^2B^2}{\zeta_*^2}$, then we pick $\mu = \frac{H}{6R}$ so the lower bound reads

$$\mathbb{E}F(\hat{x}) - F^* \geq \beta \cdot \min\left\{\frac{HB^2}{R},\ \frac{\zeta_*^2}{H}\right\} = \beta \cdot \frac{HB^2}{R} \tag{95}$$

Consequently,

$$\mathbb{E}F(\hat{x}) - F^* \geq \beta \cdot \min\left\{\frac{HB^2}{R},\ \frac{\left(H\zeta_*^2B^4\right)^{1/3}}{R^{2/3}}\right\} \tag{96}$$

Combining these with the stochastic terms completes the proof. □

E Proof of Theorem 3

We prove the theorem with the help of several technical lemmas.

Lemma 7. *For any stepsize* $\eta_t \leq \frac{1}{10H}$

$$\mathbb{E}[F(\bar{x}_t) - F^*] \leq \left(\frac{1}{\eta_t} - \lambda\right)\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{1}{\eta_t}\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 + \frac{3\sigma_*^2\eta_t}{M} + \frac{2H}{M}\sum_{m=1}^{M}\mathbb{E}\|\bar{x}_t - x_t^m\|^2$$

Proof. This lemma and its proof are nearly identical to [Lemma 8 12]. We include a proof here in order to keep the paper self-contained.

Let $\bar{x}_{t+1} = \frac{1}{M}\sum_{m=1}^{M} x_t^m$ be the average of the machines' local iterates at time t. Then,

$$\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 = \mathbb{E}\left\|\bar{x}_t - \frac{\eta_t}{M}\sum_{m=1}^{M}\nabla F_m(x_t^m) - x^*\right\|^2 + \eta_t^2\mathbb{E}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla f(x_t^m; z_t^m) - \nabla F_m(x_t^m)\right\|^2 \tag{97}$$

Beginning with the first term of (97):

$$\begin{aligned}
&\mathbb{E}\left\|\bar{x}_t - \frac{\eta_t}{M}\sum_{m=1}^{M}\nabla F_m(x_t^m) - x^*\right\|^2 \\
&\quad= \mathbb{E}\|\bar{x}_t - x^*\|^2 + \eta_t^2\mathbb{E}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla F_m(x_t^m)\right\|^2 - \frac{2\eta_t}{M}\sum_{m=1}^{M}\mathbb{E}\left\langle \bar{x}_t - x^*,\, \nabla F_m(x_t^m)\right\rangle
\end{aligned} \tag{98}$$

We can bound the second term of (98) with:

$$\begin{aligned}
&\eta_t^2\mathbb{E}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla F_m(x_t^m)\right\|^2 \\
&\quad\leq 2\eta_t^2\mathbb{E}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla F_m(x_t^m) - \nabla F_m(\bar{x}_t)\right\|^2 + 2\eta_t^2\mathbb{E}\left\|\frac{1}{M}\sum_{m=1}^{M}\nabla F_m(\bar{x}_t) - \nabla F_m(x^*)\right\|^2 && (99)\\
&\quad\leq \frac{2\eta_t^2}{M}\sum_{m=1}^{M}\mathbb{E}\|\nabla F_m(x_t^m) - \nabla F_m(\bar{x}_t)\|^2 + 2\eta_t^2\mathbb{E}\|\nabla F(\bar{x}_t) - \nabla F(x^*)\|^2 && (100)\\
&\quad\leq \frac{2H^2\eta_t^2}{M}\sum_{m=1}^{M}\mathbb{E}\|x_t^m - \bar{x}_t\|^2 + 4H\eta_t^2\mathbb{E}[F(\bar{x}_t) - F(x^*)] && (101)
\end{aligned}$$

For the third term of (98):

$$\begin{aligned}
&-\frac{2\eta_t}{M}\sum_{m=1}^{M}\mathbb{E}\left\langle \bar{x}_t - x^*,\, \nabla F_m(x_t^m)\right\rangle \\
&\quad= -\frac{2\eta_t}{M}\sum_{m=1}^{M}\mathbb{E}\left\langle x_t^m - x^*,\, \nabla F_m(x_t^m)\right\rangle + \frac{2\eta_t}{M}\sum_{m=1}^{M}\mathbb{E}\left\langle x_t^m - \bar{x}_t,\, \nabla F_m(x_t^m)\right\rangle && (102)\\
&\quad\leq -\frac{2\eta_t}{M}\sum_{m=1}^{M}\mathbb{E}\left[F_m(x_t^m) - F_m(x^*) + \frac{\lambda}{2}\|x_t^m - x^*\|^2\right] \\
&\qquad + \frac{2\eta_t}{M}\sum_{m=1}^{M}\mathbb{E}\left[F_m(x_t^m) - F_m(\bar{x}_t) + \frac{H}{2}\|x_t^m - \bar{x}_t\|^2\right] && (103)\\
&\quad\leq -2\eta_t\mathbb{E}\left[F(\bar{x}_t) - F(x^*) + \frac{\lambda}{2}\|\bar{x}_t - x^*\|^2\right] + \frac{H\eta_t}{M}\sum_{m=1}^{M}\|x_t^m - \bar{x}_t\|^2 && (104)
\end{aligned}$$

Finally, for the second term of (97)

$$\begin{aligned}
&\eta_t^2 \mathbb{E}\left\| \frac{1}{M}\sum_{m=1}^{M} \nabla f(x_t^m; z_t^m) - \nabla F_m(x_t^m) \right\|^2 \\
&= \frac{\eta_t^2}{M^2}\sum_{m=1}^{M} \mathbb{E}\|\nabla f(x_t^m; z_t^m) - \nabla F_m(x_t^m)\|^2 && (105)\\
&\leq \frac{3\eta_t^2}{M^2}\sum_{m=1}^{M} \Big[\mathbb{E}\|\nabla f(x_t^m; z_t^m) - \nabla f(\bar{x}_t; z_t^m)\|^2 + \mathbb{E}\|\nabla f(\bar{x}_t; z_t^m) - \nabla f(x^*; z_t^m)\|^2 \\
&\qquad + \mathbb{E}\|\nabla f(x^*; z_t^m) - \nabla F_m(x^*)\|^2\Big] && (106)\\
&\leq \frac{3\eta_t^2}{M^2}\sum_{m=1}^{M} \Big[H^2\mathbb{E}\|x_t^m - \bar{x}_t\|^2 + 2H\mathbb{E}[F_m(\bar{x}_t) - F_m(x^*)] + \sigma_{*,m}^2\Big] && (107)\\
&\leq \frac{3\eta_t^2}{M^2}\sum_{m=1}^{M} \Big[H^2\mathbb{E}\|x_t^m - \bar{x}_t\|^2 + 2H\mathbb{E}[F(\bar{x}_t) - F(x^*)] + \sigma_{*,m}^2\Big] && (108)
\end{aligned}$$

Combining all these results back into (97), we have

$$\begin{aligned}
\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 &\leq (1 - \lambda\eta_t)\mathbb{E}\|\bar{x}_t - x^*\|^2 + \frac{H\eta_t + 5H^2\eta_t^2}{M}\sum_{m=1}^{M}\mathbb{E}\|x_t^m - \bar{x}_t\|^2 \\
&\qquad + (10H\eta_t^2 - 2\eta_t)\mathbb{E}[F(\bar{x}_t) - F(x^*)] + \frac{3\eta_t^2\sigma_*^2}{M} && (109)\\
&\leq (1 - \lambda\eta_t)\mathbb{E}\|\bar{x}_t - x^*\|^2 + \frac{2H\eta_t}{M}\sum_{m=1}^{M}\mathbb{E}\|x_t^m - \bar{x}_t\|^2 \\
&\qquad - \eta_t\mathbb{E}[F(\bar{x}_t) - F(x^*)] + \frac{3\eta_t^2\sigma_*^2}{M} && (110)
\end{aligned}$$

where for the final line we used that $\eta_t \leq \frac{1}{10H}$. Rearranging completes the proof. □

Lemma 8. *If* $\sup_{x,m}\|\nabla F_m(x) - \nabla F(x)\|^2 \leq \bar{\zeta}^2$*, then for any fixed stepsize* η

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|x_t^m - \bar{x}_t\|^2 \leq 3K\sigma^2\eta^2 + 6K^2\eta^2\bar{\zeta}^2$$

Similarly, the decreasing stepsize $\eta_t = \frac{2}{\lambda(a+t+1)}$ *for any* a

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|x_t^m - \bar{x}_t\|^2 \leq 3K\sigma^2\eta_{t-1}^2 + 6K^2\bar{\zeta}^2\eta_{t-1}^2$$

Proof. By Jensen's inequality

$$\mathbb{E}\|x_t^m - \bar{x}_t\|^2 \leq \frac{1}{M}\sum_{n=1}^{M}\mathbb{E}\|x_t^m - x_t^n\|^2 \tag{111}$$

Therefore, it suffices to bound $\mathbb{E}\|x_t^m - x_t^n\|^2$, which we do now:

$$\begin{aligned}
&\mathbb{E}\|x_t^m - x_t^n\|^2 \\
&\leq \mathbb{E}\left\|x_{t-1}^m - x_{t-1}^n - \eta_{t-1}\big(\nabla F(x_{t-1}^m) - \nabla F(x_{t-1}^n)\big)\right. \\
&+ \left.\eta_{t-1}\big(\nabla F(x_{t-1}^m) - \nabla F_m(x_{t-1}^m) - \nabla F(x_{t-1}^n) + \nabla F_n(x_{t-1}^n)\big)\right\|^2 + \eta_{t-1}^2\sigma_m^2 \qquad (112)\\
&\leq \inf_{\gamma>0}\left(1+\frac{1}{\gamma}\right)\mathbb{E}\left\|x_{t-1}^m - x_{t-1}^n - \eta_{t-1}\big(\nabla F(x_{t-1}^m) - \nabla F(x_{t-1}^n)\big)\right\|^2 \\
&\qquad + (1+\gamma)\eta_{t-1}^2\mathbb{E}\left\|\nabla F(x_{t-1}^m) - \nabla F_m(x_{t-1}^m) - \nabla F(x_{t-1}^n) + \nabla F_n(x_{t-1}^n)\right\|^2 + \eta_{t-1}^2\sigma_m^2 \qquad (113)\\
&\leq \inf_{\gamma>0}\left(1+\frac{1}{\gamma}\right)(1-\lambda\eta_{t-1})\mathbb{E}\left\|x_{t-1}^m - x_{t-1}^n\right\|^2 + \eta_{t-1}^2\sigma_m^2 \\
&+ (1+\gamma)\eta_{t-1}^2\mathbb{E}\left\|\nabla F(x_{t-1}^m) - \nabla F_m(x_{t-1}^m)\right\|^2 \\
&+ (1+\gamma)\eta_{t-1}^2\mathbb{E}\left\|\nabla F(x_{t-1}^n) - \nabla F_n(x_{t-1}^n)\right\|^2 \\
&- 2(1+\gamma)\eta_{t-1}^2\mathbb{E}\left\langle\nabla F(x_{t-1}^m) - \nabla F_m(x_{t-1}^m),\, \nabla F(x_{t-1}^n) - \nabla F_n(x_{t-1}^n)\right\rangle \qquad (114)
\end{aligned}$$

For the third inequality we used Lemma 1. Therefore,

$$\begin{aligned}
&\frac{1}{M^2}\sum_{m=1}^M\sum_{n=1}^M\mathbb{E}\|x_t^m - x_t^n\|^2 \\
&\qquad\leq \frac{1}{M^2}\sum_{m=1}^M\inf_{\gamma>0}\left(1+\frac{1}{\gamma}\right)(1-\lambda\eta_{t-1})\mathbb{E}\left\|x_{t-1}^m - x_{t-1}^n\right\|^2 + \eta_{t-1}^2\sigma_m^2 + 2(1+\gamma)\eta_{t-1}^2\bar{\zeta}^2
\end{aligned} \qquad (115)$$

We will unroll this recurrence, using that $x_{t_0}^m = x_{t_0}^n$ for all m, n where t_0 is the most recent time that the iterates were synchronized, so $t - t_0 \leq K - 1$. Taking $\gamma = K - 1$, we have

$$\begin{aligned}
\frac{1}{M^2}\sum_{m=1}^M\sum_{n=1}^M\mathbb{E}\|x_t^m - x_t^n\|^2 &= \sum_{i=t_0}^{t-1}\big(\eta_i^2\sigma^2 + 2(1+\gamma)\eta_i^2\bar{\zeta}^2\big)\prod_{j=i+1}^{t-1}\left(1+\frac{1}{\gamma}\right)(1-\lambda\eta_j) &(116)\\
&\leq \sum_{i=t_0}^{t-1}\big(\eta_i^2\sigma^2 + 2K\eta_i^2\bar{\zeta}^2\big)\prod_{j=i+1}^{t-1}\left(1+\frac{1}{K-1}\right)(1-\lambda\eta_j) &(117)\\
&\leq \sum_{i=t_0}^{t-1}\big(\eta_i^2\sigma^2 + 2K\eta_i^2\bar{\zeta}^2\big)\left(1+\frac{1}{K-1}\right)^{K-1}\prod_{j=i+1}^{t-1}(1-\lambda\eta_j) &(118)\\
&\leq 3\big(\sigma^2 + 2K\bar{\zeta}^2\big)\sum_{i=t_0}^{t-1}\eta_i^2\prod_{j=i+1}^{t-1}(1-\lambda\eta_j) &(119)
\end{aligned}$$

For a constant stepsize η,

$$\begin{aligned}
\frac{1}{M^2}\sum_{m=1}^M\sum_{n=1}^M\mathbb{E}\|x_t^m - x_t^n\|^2 &\leq 3\big(\sigma^2 + 2K\bar{\zeta}^2\big)\sum_{i=t_0}^{t-1}\eta^2 &(120)\\
&\leq 3K\big(\sigma^2 + 2K\bar{\zeta}^2\big)\eta^2 &(121)
\end{aligned}$$

For decreasing stepsize $\eta_t = \frac{2}{\lambda(a+t+1)}$

$$\frac{1}{M^2}\sum_{m=1}^{M}\sum_{n=1}^{M}\mathbb{E}\|x_t^m - x_t^n\|^2 \le 3\big(\sigma^2 + 2K\bar{\zeta}^2\big)\sum_{i=t_0}^{t-1}\eta_i^2\prod_{j=i+1}^{t-1}\frac{a+j-1}{a+j+1} \tag{122}$$

$$= 3\big(\sigma^2 + 2K\bar{\zeta}^2\big)\sum_{i=t_0}^{t-1}\eta_i^2\frac{(a+i)(a+i+1)}{(a+t)(a+t+1)} \tag{123}$$

$$= 3\big(\sigma^2 + 2K\bar{\zeta}^2\big)\sum_{i=t_0}^{t-1}\eta_i^2\frac{\eta_{t-1}\eta_t}{\eta_{i-1}\eta_i} \tag{124}$$

$$\le 3\big(\sigma^2 + 2K\bar{\zeta}^2\big)\sum_{i=t_0}^{t-1}\eta_i^2\frac{\eta_{t-1}^2}{\eta_i^2} \tag{125}$$

$$= 3K\big(\sigma^2 + 2K\bar{\zeta}^2\big)\eta_{t-1}^2 \tag{126}$$

□

Theorem 3. *When* $\sup_x \frac{1}{M}\sum_{m=1}^{M}\|\nabla F_m(x) - \nabla F(x)\|^2 \le \bar{\zeta}^2$*, an average of the Local SGD iterates guarantees under the convex and strongly convex assumptions, respectively*

$$\mathbb{E}F(\hat{x}) - F^* \le c\cdot\left(\frac{HB^2}{KR} + \frac{\big(H\bar{\zeta}^2B^4\big)^{1/3}}{R^{2/3}} + \frac{\big(H\sigma^2B^4\big)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\sigma_* B}{\sqrt{MKR}}\right),$$

$$\mathbb{E}F(\hat{x}) - F^* \le c\cdot\left(\frac{H^2B^2}{HKR + \lambda K^2R^2} + \left(\frac{H\bar{\zeta}^2}{\lambda^2R^2} + \frac{H\sigma^2}{\lambda^2KR^2}\right)\log\left(\frac{H}{\lambda} + KR\right) + \frac{\sigma_*^2}{\lambda MKR}\right).$$

Proof. By Lemma 7, for any $\eta_t \le \frac{1}{10H}$

$$\mathbb{E}[F(\bar{x}_t) - F^*] \le \left(\frac{1}{\eta_t} - \lambda\right)\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{1}{\eta_t}\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 + \frac{3\sigma_*^2\eta_t}{M} + \frac{2H}{M}\sum_{m=1}^{M}\mathbb{E}\|\bar{x}_t - x_t^m\|^2 \tag{127}$$

By Lemma 8, when $\eta_t = \eta$ is constant then

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|x_t^m - \bar{x}_t\|^2 \le 3K\sigma^2\eta^2 + 6K^2\eta^2\bar{\zeta}^2 \tag{128}$$

and when $\eta_t = \frac{2}{\lambda(a+t+1)}$

$$\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}\|x_t^m - \bar{x}_t\|^2 \le 3K\sigma^2\eta_{t-1}^2 + 6K^2\bar{\zeta}^2\eta_{t-1}^2 \tag{129}$$

We now consider the convex and strongly convex cases separately:

Convex case: In the convex case, we use a constant stepsize η, so

$$\mathbb{E}[F(\bar{x}_t) - F^*]$$

$$\le \frac{1}{\eta}\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{1}{\eta}\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 + \frac{3\sigma_*^2\eta}{M} + \frac{2H}{M}\sum_{m=1}^{M}\mathbb{E}\|\bar{x}_t - x_t^m\|^2 \tag{130}$$

$$\le \frac{1}{\eta}\mathbb{E}\|\bar{x}_t - x^*\|^2 - \frac{1}{\eta}\mathbb{E}\|\bar{x}_{t+1} - x^*\|^2 + \frac{3\sigma_*^2\eta}{M} + 6HK\sigma^2\eta^2 + 12HK^2\eta^2\bar{\zeta}^2 \tag{131}$$

Therefore, by the convexity of F

$$\mathbb{E}\left[F\left(\frac{1}{KR}\sum_{t=1}^{KR}\bar{x}_t\right) - F^*\right] \le \frac{1}{KR}\sum_{t=1}^{KR}\mathbb{E}[F(\bar{x}_t) - F^*] \tag{132}$$

$$\le \frac{B^2}{\eta KR} + \frac{3\sigma_*^2\eta}{M} + 6HK\sigma^2\eta^2 + 12HK^2\eta^2\bar{\zeta}^2 \tag{133}$$

Choosing

$$\eta = \min\left\{\frac{1}{10H}, \frac{B\sqrt{M}}{\sigma_*\sqrt{KR}}, \left(\frac{B^2}{HK^2\sigma^2}\right)^{1/3}, \left(\frac{B^2}{HK^2\bar{\zeta}^2}\right)^{1/3}\right\} \tag{134}$$

then ensures

$$\mathbb{E}\left[F\left(\frac{1}{KR}\sum_{t=1}^{KR}\bar{x}_t\right) - F^*\right] \leq \frac{10HB^2}{KR} \frac{13\left(H\bar{\zeta}^2B^4\right)^{1/3}}{R^{2/3}} + \frac{7\left(H\sigma^2B^4\right)^{1/3}}{K^{1/3}R^{2/3}} + \frac{4\sigma_*B}{\sqrt{MKR}} \tag{135}$$

Strongly convex case: In the strongly convex case, we take the stepsize $\eta_t = \frac{2}{\lambda(a+t+1)}$ for $a = 20H/\lambda$ which ensures $\eta_t \leq \frac{1}{10H}$. In addition, we define weights $w_t = (a+t)$ and define

$$\bar{x} = \frac{1}{W}\sum_{t=1}^{KR} w_t\bar{x}_t \tag{136}$$

where $W = \sum_{t=1}^{KR} w_t \geq \frac{1}{2}KR(a+KR)$. By the convexity of F,

$$\begin{aligned}
&\mathbb{E}F(\bar{x}) - F^* \\
&\leq \frac{1}{W}\sum_{t=1}^{KR}(a+t)\mathbb{E}F(\bar{x}_t) - F^* && (137)\\
&\leq \frac{\lambda(a+1)(a+2)B^2}{2W} + \frac{1}{W}\sum_{t=1}^{KR}\left[\frac{6\sigma_*^2}{\lambda M} + \frac{2H(a+t)}{M}\sum_{m=1}^{M}\mathbb{E}\|\bar{x}_t - x_t^m\|^2\right] && (138)\\
&\leq \frac{\lambda(a+1)(a+2)B^2}{2W} + \frac{6\sigma_*^2KR}{W\lambda M} + \frac{6HK\sigma^2 + 12HK^2\bar{\zeta}^2}{W}\sum_{t=1}^{KR}(a+t)\eta_{t-1}^2 && (139)\\
&\leq \frac{\lambda(a+1)(a+2)B^2}{2W} + \frac{6\sigma_*^2KR}{W\lambda M} + \frac{6HK\sigma^2 + 12HK^2\bar{\zeta}^2}{\lambda^2 W}(1+\log(a+KR)) && (140)\\
&\leq \frac{132H^2B^2}{\lambda KR(10H/\lambda + KR)} + \left(\frac{12H\bar{\zeta}^2}{\lambda^2R^2} + \frac{6H\sigma^2}{\lambda^2KR^2}\right)\log\left(\frac{13H}{\lambda} + KR\right) + \frac{6\sigma_*^2}{\lambda MKR} && (141)
\end{aligned}$$

Note that in the strongly convex case, it is likely possible to achieve a first term scaling with $\exp(-KR)$ using a method similar to Lemma 2. However, the recurrence we derived here has a different form, and it is difficult to determine the correct stepsize and weighting schedule to achieve linear convergence. □

F Details of Experiments

The training set of MNIST (60,000 examples) was divided by digit into ten groups of equal size $n \approx 6,000$ (which required discarding some examples from the more common digits). PCA was used to reduce the dimensionality to 100, but no other preprocessing was used.

Then, for each of the 25 combinations (i,j) for even i and odd j, a binary classification "task" was created, i.e. classifying even ($+1$) versus odd (-1). These tasks were arbitrarily labelled task $1, 2, \ldots, 25$.

For each $p \in [0.0, 0.2, 0.4, 0.6, 0.8, 1.0]$, machine m was assigned data composed of $p \cdot 2n$ random examples from task m, and $(1-p)\cdot 2n$ random examples from a mixture of all the tasks.

Local and Minibatch SGD were then used to optimize the logistic loss for each of the six described local datasets. The constant stepsize was tuned (from a log-scale grid of 10 points ranging from $e^{-6}, \ldots, e^0$ for Minibatch SGD, and a log-scale grid of 10 points ranging from $e^{-8}, \ldots, e^{-1}$ for Local SGD) for each value of p, K, and R individually, and the average loss over four runs is reported for the best stepsize for each point in the plot. That is, each point in the plot represents the best possible performance of the algorithm for that p, K, and R specifically.

Finally, we computed the value of ζ_*^2 as a function of p by using Newton's method to compute a very accurate estimate of the minimizer, and then explicitly calculating $\zeta_*^2(p)$ at that point.

G Proof of Theorem 4

For this lower bound, the gradients will always be noiseless, so we simply define the expectation of the local functions. Furthermore, we will construct just two local functions F_1 and F_2. For the case $M > 2$, F_1 will be assigned to the first $\lfloor M/2 \rfloor$ machines, and F_2 to the next $\lfloor M/2 \rfloor$ machines. If there is an odd number of machines, we simply assign the last machine $F_3(x) = \frac{\lambda}{2}\|x\|^2$, which will reduce the lower bound by a factor of at most $\frac{M-1}{M}$. Therefore, we proceed by focusing on the case $M = 2$.

We define the following H-smooth and λ-strongly convex functions on $\mathbb{R}^d$ for even d:

$$F(x) = \frac{1}{2}(F_1(x) + F_2(x)) \tag{142}$$

$$F_1(x) = \frac{H-\lambda}{8}\left(x_1^2 - 2Cx_1 + \beta x_d^2 + \sum_{i=1}^{d/2-1}\left(x_{2i+1} - x_{2i}\right)^2\right) + \frac{\lambda}{2}\|x\|^2 \tag{143}$$

$$F_2(x) = \frac{H-\lambda}{8}\left(\sum_{i=1}^{d/2}(x_{2i} - x_{2i-1})^2\right) + \frac{\lambda}{2}\|x\|^2 \tag{144}$$

Here, β and C are constants which will be chosen later.

These functions are identical to ones used by Woodworth and Srebro [24] to prove lower bounds for finite sum optimization, and are very similar both to classic work by Nesterov [18] on lower bounds and to more closely related work by Arjevani and Shamir [1]. Arjevani and Shamir also prove lower bounds for distributed optimization algorithms, but their slightly different construction made it more difficult to tune ζ_*^2, which is necessary for our lower bound.

These functions have the following important property: let $E_k = \operatorname{span}\{e_1, \ldots, e_k\}$ be the set of vectors whose $k+1, \ldots, d$ coordinates are all zero, then for all $x_k \in E_k$ for even k

$$\nabla F_1(x_k) \in E_{k+1} \qquad \text{and} \qquad \nabla F_2(x_k) \in E_k \tag{145}$$

and for $x_k \in E_k$ for odd k

$$\nabla F_1(x_k) \in E_k \qquad \text{and} \qquad \nabla F_2(x_k) \in E_{k+1} \tag{146}$$

For algorithms whose iterates, for example, remain in the span of previous gradients, the only way to access the next coordinate is to query the gradient of one of the two functions—F_1 if the next coordinate is odd, and F_2 if the next coordinate is even. Since each machine will only have access to one of the two functions throughout each round of communication, this means that each round of communication can only unlock a single new coordinate. We now formalize this.

Following Carmon et al. [5], we define:

Definition 2 (Distributed zero-respecting algorithm)**.** *For a vector v, let $\operatorname{supp}(v) = \{i \in \{1, \ldots, d\} : v_i \neq 0\}$. We say that an optimization algorithm is distributed zero-respecting if for all t and m, the tth query on the mth machine, x_t^m satisfies*

$$\operatorname{supp}(x_t^m) \subseteq \bigcup_{s<t} \operatorname{supp}(\nabla f(x_s^m; z_s^m)) \cup \bigcup_{m' \neq m} \bigcup_{s \leq \pi_m(t,m')} \operatorname{supp}(\nabla f(x_s^{m'}; z_s^{m'}))$$

where $\pi_m(t, m')$ is the most recent time before t when machines m and m' communicated with each other.

This definition captures a very wide variety of distributed optimization algorithms, including minibatch SGD, accelerated minibatch SGD, local SGD, coordinate descent methods, and many more. Algorithms which are *not* distributed zero-respecting are those whose iterates have components in directions about which the algorithm has no information, meaning that in some sense, it is just "wild guessing." Using techniques similar to Woodworth and Srebro [Theorem 7 24] and Carmon et al. [5], it should be possible to extend this lower bound beyond distributed zero-respecting algorithms to arbitrary randomized algorithms.

We now argue that the progress of distributed zero-respecting algorithms is controlled by the number of rounds of communication, R, regardless of K:

Lemma 9. *Let $\hat{x}$ be the output after R rounds of communication of a distributed zero-respecting algorithm optimizing $F = \frac{1}{2}(F_1 + F_2)$ as defined in* (142)*. Then,*

$$\mathrm{supp}(x_t^m) \in E_R$$

Proof. The definition of a zero-respecting algorithm requires that every machine's initial iterate $x_0^m = 0$. We will now prove the Lemma by induction on the round of communication.

As a base case, for the first iteration of the first round of communication:

$$\nabla F_1(x_0^1) = \nabla F_1(0) = \frac{(\lambda - H)C}{4} e_1 \in E_1 \qquad \text{and} \qquad \nabla F_2(x_1^2) = \nabla F_2(0) = 0 \in E_0 \quad (147)$$

Therefore, by the distributed zero-respecting property, $x_2^1 \in E_1$ and $x_2^2 \in E_0$. Furthermore, for all $y_1 \in E_1$, $\nabla F_1(y_1) \in E_1$ and for all $y_0 \in E_0$, $\nabla F_2(y_0) \in E_0$. Therefore, further gradient queries on each machine will not change the set of coordinates that the distributed zero-respecting property allows to be non-zero. We conclude that $x_t^1 \in E_1$ and $x_t^2 \in E_0$ for all t until machines 1 and 2 communicate with each other.

Now, suppose that after $r-1$ rounds of communication, $x_t^1, x_t^2 \in E_{r-1}$. If r is even, then $\nabla F_1(x_t^1) \in E_{r-1}$ and $\nabla F_2(x_t^2) \in E_r$. Furthermore, additional gradient computations within the rth round of communication will not expand the set of coordinates that the distributed zero-respecting property will allow to be non-zero. Therefore, both machines' coordinates will remain in E_r until the end of the rth round of communication. A similar argument can be made for odd r. □

Now, we will compute the minimizer of F. We note that by the definition of F_1 and F_2,

$$F(x) = \frac{H - \lambda}{16}\left(x_1^2 - 2Cx_1 + \beta x_d^2 + \sum_{i=2}^{d}(x_i - x_{i-1})^2\right) + \frac{\lambda}{2}\|x\|^2 \quad (148)$$

Calculating the gradient of F, we see that $x^* = \arg\min_x F(x)$ must satisfy

$$\begin{aligned} C &= \left(2 + \frac{8\lambda}{H - \lambda}\right)x_1^* - x_2^* \\ 0 &= \left(2 + \frac{8\lambda}{H - \lambda}\right)x_i^* - x_{i+1}^* - x_{i-1}^* \qquad \forall_{i \in \{2,\ldots,d-1\}} \\ 0 &= \left(1 + \beta + \frac{8\lambda}{H - \lambda}\right)x_d^* - x_{d-1}^* \end{aligned} \quad (149)$$

Let q be the smaller solution of the quadratic equation

$$1 - \left(2 + \frac{8\lambda}{H - \lambda}\right)q + q^2 = 0 \quad (150)$$

That is,

$$q = 1 + \frac{4\lambda}{H - \lambda} - \sqrt{\frac{16\lambda^2}{(H - \lambda)^2} + \frac{8\lambda}{H - \lambda}} \quad (151)$$

$$= 1 + \frac{4\lambda}{H - \lambda}\left(1 - \sqrt{1 + \frac{H - \lambda}{2\lambda}}\right) \quad (152)$$

$$= 1 - \frac{2}{\left(1 - \sqrt{1 + \frac{H-\lambda}{2\lambda}}\right)\left(1 + \sqrt{1 + \frac{H-\lambda}{2\lambda}}\right)}\left(1 - \sqrt{1 + \frac{H - \lambda}{2\lambda}}\right) \quad (153)$$

$$= \frac{\sqrt{1 + \frac{H-\lambda}{2\lambda}} - 1}{\sqrt{1 + \frac{H-\lambda}{2\lambda}} + 1} \quad (154)$$

Let $\alpha = \sqrt{1 + \frac{H-\lambda}{2\lambda}}$ so that $q = \frac{\alpha-1}{\alpha+1}$, and define $\beta = 1 - q$. Then it is straightforward to confirm that

$$x^* = C \sum_{i=1}^{d} q^i e_i \tag{155}$$

satisfies all of the conditions (149), and is thus the minimizer of F. This point has value

$$F(x^*) = \frac{C^2(H-\lambda)}{16}\left(q^2 - 2q + \beta q^{2d} + (1-q)^2 \sum_{i=2}^{d} q^{2i-2} + \frac{8\lambda}{H-\lambda}\sum_{i=1}^{d} q^{2i}\right) \tag{156}$$

$$= \frac{C^2(H-\lambda)}{16}\left(-1 + \beta q^{2d} + (1-q)^2 \sum_{i=1}^{d} q^{2i-2} + \frac{8\lambda}{H-\lambda}\sum_{i=1}^{d} q^{2i}\right) \tag{157}$$

$$= \frac{C^2(H-\lambda)}{16}\left(-1 + (1-q) q^{2d} + \frac{8\lambda}{H-\lambda}\sum_{i=1}^{d} q^{2i-1} + q^{2i}\right) \tag{158}$$

$$= \frac{C^2(H-\lambda)}{16}\left(-1 + (1-q) q^{2d} + \frac{8\lambda}{H-\lambda}\left(\frac{q(1-q^{2d})}{1-q^2} + \frac{q^2(1-q^{2d})}{1-q^2}\right)\right) \tag{159}$$

$$= \frac{C^2(H-\lambda)}{16}\left(-1 + (1-q) q^{2d} + \frac{(1-q)^2}{q}\left(\frac{q(1-q^{2d})}{1-q^2} + \frac{q^2(1-q^{2d})}{1-q^2}\right)\right) \tag{160}$$

$$= \frac{C^2(H-\lambda)}{16}\left(-1 + (1-q) q^{2d} + (1-q)(1-q^{2d})\right) \tag{161}$$

$$= \frac{-qC^2(H-\lambda)}{16} \tag{162}$$

For the third equality, we used that (150) implies $(1-q)^2 = \frac{8\lambda q}{H-\lambda}$. For the fifth inequality, we used that $\frac{8\lambda}{H-\lambda} = \frac{(1-q)^2}{q}$. This solution has norm

$$\|x^*\|^2 = C^2 \sum_{i=1}^{d} q^{2i} = C^2 \frac{q^2(1-q^{2d})}{1-q^2} \leq \frac{q^2 C^2}{1-q^2} = \frac{C^2(\alpha-1)^2}{4\alpha} \leq \frac{\alpha C^2}{4} \tag{163}$$

Furthermore,

$$F(0) - F(x^*) = -F(x^*) = \frac{qC^2(H-\lambda)}{16} \tag{164}$$

Finally, we evaluate the degree of heterogeneity:

$$\zeta_*^2 = \frac{1}{2}\sum_{m=1}^{2} \|\nabla F_m(x^*)\|^2 = \|\nabla F_1(x^*)\|^2 = \|\nabla F_2(x^*)\|^2 \tag{165}$$

$$= \frac{(H-\lambda)^2}{64}\left\|2\sum_{i=1}^{d/2}(x^*_{2i} - x^*_{2i-1})(e_{2i} - e_{2i-1}) + \frac{4\lambda}{H-\lambda}x^*\right\|^2 \tag{166}$$

$$= \frac{(H-\lambda)^2}{16}\sum_{i=1}^{d/2}\left[\left(x^*_{2i-1}\left(-1 + \frac{4\lambda}{H-\lambda}\right)\right)^2 + \left(x^*_{2i}\left(1 + \frac{4\lambda}{H-\lambda}\right)\right)^2\right] \tag{167}$$

$$= \frac{C^2(H-\lambda)^2}{16}\sum_{i=1}^{d/2}\left[q^{4i-2}\frac{(H-5\lambda)^2}{(H-\lambda)^2} + q^{4i}\frac{(H+3\lambda)^2}{(H-\lambda)^2}\right] \tag{168}$$

$$\leq \frac{(H+3\lambda)^2}{16}\|x^*\|^2 \tag{169}$$

$$\leq \frac{\alpha C^2(H+3\lambda)^2}{64} \tag{170}$$

With this, we are ready to prove the lower bound.

Theorem 4. *For any M, K, and R, there exist two quadratic objectives satisfying the convex and strongly convex assumptions (for $H \geq 7\lambda$) such that the output of any distributed zero-respecting algorithm will have suboptimality in the convex and strongly convex case respectively,*

$$F(\hat{x}) - F^* \geq c\left(\min\left\{\frac{\zeta_*^2}{HR^2}, \frac{HB^2}{R^2}\right\} + \frac{\sigma B}{\sqrt{MKR}}\right),$$
$$F(\hat{x}) - F^* \geq c\left(\min\left\{\frac{\lambda\zeta_*^2}{H^2}, \frac{\Delta\sqrt{\lambda}}{\sqrt{H}}\right\}\exp\left(-\frac{8R\sqrt{\lambda}}{\sqrt{H}}\right) + \frac{\sigma^2}{\lambda MKR}\right).$$

Proof. By Lemma 9, the output of the algorithm $\hat{x} \in E_R$. Furthermore, since F is λ-strongly convex, $F(\hat{x}) - F^* \geq \frac{\lambda}{2}\|\hat{x} - x^*\|^2$. Therefore,

$$\frac{F(\hat{x}) - F^*}{F(0) - F^*} \geq \frac{\frac{\lambda}{2}\|\hat{x} - x^*\|^2}{\frac{qC^2(H-\lambda)}{16}} \tag{171}$$
$$\geq \frac{8\lambda}{q(H-\lambda)}\sum_{i=R+1}^{d} q^{2i} \tag{172}$$
$$= \frac{8\lambda q(q^{2R} - q^{2d})}{(H-\lambda)(1-q^2)} \tag{173}$$
$$= \frac{(1-q)^2(q^{2R} - q^{2d})}{1-q^2} \tag{174}$$
$$= \frac{(1-q)(q^{2R} - q^{2d})}{1+q} \tag{175}$$
$$= \frac{q^{2R} - q^{2d}}{\alpha} \tag{176}$$

For the third equality we used that (150) implies $\frac{8\lambda q}{H-\lambda} = (1-q)^2$. For the final equality, we used that $q = \frac{\alpha-1}{\alpha+1}$. Taking $d \geq R + \frac{1}{2\ln(1/q)}$ ensures that $q^{2d} \leq \frac{q^{2R}}{2}$ so

$$\frac{F(\hat{x}) - F^*}{F(0) - F^*} \geq \frac{q^{2R}}{2\alpha} = \frac{\left(1 - \frac{2}{\alpha+1}\right)^{2R}}{2\alpha} \tag{177}$$

Therefore,

$$R \leq \frac{\ln\left(\frac{F(0)-F^*}{2\alpha\epsilon}\right)}{\ln\left(1 + \frac{2}{\alpha-1}\right)} \implies F(\hat{x}) - F^* \geq \epsilon \tag{178}$$

Using the fact that $\ln(1+x) \leq x$ and solving the above inequality on R for ϵ, we conclude that

$$F(\hat{x}) - F^* \geq \frac{F(0) - F^*}{2\alpha}\exp\left(-\frac{2R}{\alpha - 1}\right) \tag{179}$$

In order to satisfy the strongly convex assumptions, we recall from (170) and (164) that we must choose C such that

$$\frac{\alpha C^2(H+3\lambda)^2}{64} \leq \frac{\alpha C^2 H^2}{16} \leq \zeta_*^2 \tag{180}$$
$$\frac{qC^2(H-\lambda)}{16} \leq \frac{C^2 H}{16} \leq \Delta \tag{181}$$

Therefore, we choose $C^2 = 16\min\left\{\frac{\zeta_*^2}{\alpha H^2},\ \frac{\Delta}{H}\right\}$ meaning that

$$F(\hat{x}) - F^* \geq \frac{F(0) - F^*}{2\alpha}\exp\left(-\frac{2R}{\alpha - 1}\right) \tag{182}$$

$$\geq \frac{\min\left\{\frac{\zeta_*^2}{\alpha H},\ \Delta\right\}}{2\alpha}\exp\left(-\frac{2R}{\alpha - 1}\right) \tag{183}$$

$$\geq \min\left\{\frac{\lambda\zeta_*^2}{H^2},\ \frac{\sqrt{\lambda}\Delta}{2\sqrt{H}}\right\}\exp\left(-\frac{8\sqrt{\lambda}R}{\sqrt{H}}\right) \tag{184}$$

For the convex case, we note that in order to satisfy the convex assumptions, we must choose C such that

$$\frac{\alpha C^2(H + 3\lambda)^2}{64} \leq \frac{\alpha C^2 H^2}{16} \leq \zeta_*^2 \tag{185}$$

$$\frac{\alpha C^2}{4} \leq B^2 \tag{186}$$

We therefore choose $C^2 = 4\min\left\{\frac{\zeta_*^2}{\alpha H^2},\ \frac{B^2}{\alpha}\right\}$. Returning to (179), this means

$$F(\hat{x}) - F^* \geq \frac{F(0) - F^*}{2\alpha}\exp\left(-\frac{2R}{\alpha - 1}\right) \tag{187}$$

$$= \frac{qC^2(H - \lambda)}{32\alpha}\exp\left(-\frac{2R}{\alpha - 1}\right) \tag{188}$$

$$\geq \frac{q(H - \lambda)\min\left\{\frac{\zeta_*^2}{\alpha H^2}\ \frac{B^2}{\alpha}\right\}}{8\alpha}\exp\left(-\frac{8\sqrt{\lambda}R}{\sqrt{H}}\right) \tag{189}$$

$$\geq q\min\left\{\frac{\zeta_*^2}{16\alpha^2 H},\ \frac{HB^2}{16\alpha^2}\right\}\exp\left(-\frac{8\sqrt{\lambda}R}{\sqrt{H}}\right) \tag{190}$$

From here, we use that $H \geq 7\lambda$ implies $\alpha \geq 2$ so $q \geq 1/3$, so

$$F(\hat{x}) - F^* \geq \min\left\{\frac{\lambda\zeta_*^2}{48H^2},\ \frac{\lambda B^2}{48}\right\}\exp\left(-\frac{8\sqrt{\lambda}R}{\sqrt{H}}\right) \tag{191}$$

Finally, this holds for any $\lambda \geq 0$, so it holds, in particular, for $\lambda = \frac{H}{64R^2}$ thus

$$F(\hat{x}) - F^* \geq c \cdot \min\left\{\frac{\zeta_*^2}{HR^2},\ \frac{HB^2}{R^2}\right\} \tag{192}$$

Finally, it is well known that any first-order method which accesses at most MKR stochastic gradients with variance σ^2 for a λ-strongly convex objective will suffer error at least $\beta\frac{\sigma^2}{\lambda MKR}$ in the worst case for a universal constant β [17]. Similarly, any first-order method which accesses at most MKR stochastic gradients with variance σ^2 for a convex objective with $\|x^*\| \leq B$ will suffer error at least $\beta\frac{\sigma B}{\sqrt{MKR}}$ in the worst case for a universal constant β [17]. □

H Proof of Corollary 1

Corollary 1. *Accelerated Minibatch SGD is optimal when $\zeta_* \geq HB$ in the convex case, and is optimal up to log factors when $\zeta_*^2 \geq H^{3/2}/\sqrt{\lambda}$ in the strongly convex case.*

Proof. In the convex case, Theorem 1 ensures Accelerated Minibatch SGD converges at a rate proportional to

$$\frac{HB^2}{R^2} + \frac{\sigma B}{\sqrt{MKR}} \tag{193}$$

The lower bound for convex functions in Theorem 4 precisely matches this whenever

$$\frac{HB^2}{R^2} = \min\left\{\frac{\zeta_*^2}{HR^2}, \frac{HB^2}{R^2}\right\} \implies \zeta_* \geq HB \tag{194}$$

For the strongly convex case, Theorem 1 ensures convergence at a rate porportional to

$$\Delta \exp\left(-\frac{\sqrt{\lambda}R}{c_3\sqrt{H}}\right) + \frac{\sigma^2}{\lambda MKR} \tag{195}$$

The lower bound is given by

$$\min\left\{\frac{\lambda\zeta_*^2}{H^2}, \frac{\Delta\sqrt{\lambda}}{\sqrt{H}}\right\} \exp\left(-\frac{8R\sqrt{\lambda}}{\sqrt{H}}\right) + \frac{\sigma^2}{\lambda MKR} \tag{196}$$

When $\zeta_*^2 \geq H^{3/2}/\sqrt{\lambda}$, this reduces to

$$\frac{\Delta\sqrt{\lambda}}{\sqrt{H}} \exp\left(-\frac{8R\sqrt{\lambda}}{\sqrt{H}}\right) + \frac{\sigma^2}{\lambda MKR} = \Delta \exp\left(-\frac{8R\sqrt{\lambda}}{\sqrt{H}} - \log\frac{\sqrt{\lambda}}{\sqrt{H}}\right) + \frac{\sigma^2}{\lambda MKR} \tag{197}$$

Comparing this with (195), we see that the R needed to guarantee error ϵ using Theorem 1 is larger than the minimum possible R, as lower bounded by (197), by at most a log factor. □