# The Limits and Potentials of Local SGD for Distributed Heterogeneous Learning with Intermittent Communication

**Kumar Kshitij Patel** KKPATEL@TTIC.EDU
*Toyota Technological Institute, Chicago, IL, USA*

**Margalit Glasgow** MGLASGOW@STANFORD.EDU
*Stanford University, Stanford, CA, USA*

**Ali Zindari** ALI.ZINDARI@CISPA.DE
*CISPA Helmholtz Center, Saarbrücken, Germany*

**Lingxiao Wang** LINGXW@TTIC.EDU
*Toyota Technological Institute, Chicago, IL, USA*

**Sebastian U. Stich** STICH@CISPA.DE
*CISPA Helmholtz Center, Saarbrücken, Germany*

**Ziheng Cheng** ALEX-CZH@STU.PKU.EDU.CN
*Peking University, Beijing, China*

**Nirmit Joshi** NIRMIT@TTIC.EDU
*Toyota Technological Institute, Chicago, IL, USA*

**Nathan Srebro** NATI@TTIC.EDU
*Toyota Technological Institute, Chicago, IL, USA*

**Editors:** Shipra Agrawal and Aaron Roth

## Abstract

Local SGD is a popular optimization method in distributed learning, often outperforming mini-batch SGD. Despite this practical success, proving the efficiency of local SGD has been difficult, creating a significant gap between theory and practice. We provide new lower bounds for local SGD under existing first-order data heterogeneity assumptions, showing these assumptions can not capture local SGD's effectiveness. We also demonstrate the min-max optimality of accelerated mini-batch SGD under these assumptions. Our findings emphasize the need for improved modeling of data heterogeneity. Under higher-order assumptions, we provide new upper bounds that verify the dominance of local SGD over mini-batch SGD when data heterogeneity is low.

**Keywords:** Distributed optimization, Local SGD, intermittent communication, min-max optimal.

## 1. Introduction

We consider the following distributed optimization problem on $M$ machines,

$$\min_{x\in\mathbb{R}^d} \left( F(x) := \frac{1}{M} \sum_{m\in[M]} F_m(x) \right), \tag{1}$$

where $F_m := \mathbb{E}_{z_m\sim\mathcal{D}_m}[f(x; z_m)]$ is a stochastic objective on machine $m$, defined using a smooth, convex and differentiable loss function $f(\cdot; z \in \mathcal{Z})$ and a data distribution $\mathcal{D}_m \in \Delta(\mathcal{Z})$. Problem (1) is ubiquitous in machine learning—from training in a data center on multiple GPUs (Krizhevsky et al., 2012), to decentralized training on millions of devices (McMahan et al., 2016b,a). Perhaps the simplest, most basic, and most important distributed setting for solving Problem (1) is that of intermittent communication (IC) (Woodworth et al., 2018), where $M$ machines work in parallel over $R$ communication rounds to optimize objective (1), and during each round of communication, each machine may sequentially compute $K$ stochastic gradient estimates (c.f., Figure 3).

While several stochastic gradient descent (SGD) methods have been proposed for solving problem (1) in the IC setting, most of them (Kairouz et al., 2019; Wang et al., 2021) are variants of **local SGD** (update (5)) or **mini-batch SGD** (update (6)). In local SGD, each machine in parallel computes $K$ sequential stochastic gradients, and then the resulting updates are averaged across the machines at each round of communication. In practice, this simple algorithm often outperforms all other first-order optimization algorithms (Charles et al., 2021; Wang et al., 2022), including mini-batch SGD (Lin et al., 2018; Woodworth et al., 2020a). As such, there has been a concerted effort to understand the effectiveness of local SGD for more than a decade (Mcdonald et al., 2009; Zinkevich et al., 2010; Zhang et al., 2016; Stich, 2018; Dieuleveut and Patel, 2019; Khaled et al., 2020; Koloskova et al., 2020; Woodworth et al., 2020a; Karimireddy et al., 2020b; Woodworth et al., 2020b; Yuan and Ma, 2020; Woodworth et al., 2021; Glasgow et al., 2022; Wang et al., 2022).

In the homogeneous setting, when $\mathcal{D}_m = \mathcal{D}$ for each $m \in [M]$, Woodworth et al. (2021) showed that the min-max optimal algorithm for optimizing problem (1) with smooth and convex objectives, is the best of (accelerated) local and mini-batch SGD. Unfortunately, their results also implied that local SGD can outperform mini-batch SGD only when even SGD on a single machine (i.e., without any collaboration) also outperforms mini-batch SGD! This result was surprising because, in practice, local SGD almost always outperforms both single-machine SGD and mini-batch SGD (McMahan and Ramage, 2017; Charles et al., 2021). The most likely reason for this disparity between theory and practice is that the homogeneous setting is too simplistic. In practice, the data distributions of different machines are "similar" but not the same, making collaboration pivotal.

Towards this end, several works (Khaled et al., 2020; Karimireddy et al., 2020b; Koloskova et al., 2020; Woodworth et al., 2020b; Yuan and Ma, 2020; Glasgow et al., 2022; Wang et al., 2022) have tried to capture the similarity between machine's distributions by using *"data heterogeneity"* assumptions. However, most of these works either fail to show a theoretical advantage of local SGD over mini-batch SGD (Khaled et al., 2020; Karimireddy et al., 2020b; Koloskova et al., 2020) or they need to make restrictive assumptions that do not allow any interesting data heterogeneity (c.f., Assumption 2 of Woodworth et al. (2020b)). This inability to theoretically explain the effectiveness of local SGD presents a big gap between the theory and practice of distributed optimization. Recently Wang et al. (2022) called this the ***"unreasonable effectiveness of local SGD"***. In this paper, we further this discourse about the usefulness of local SGD by providing new lower and upper bounds for solving problem (1) in the IC setting. Our key theoretical insights are as follows:

**1. Existing first-order heterogeneity assumptions are insufficient for local SGD.** We provide a new lower bound for local SGD (Theorem 9) in the general convex setting. Our core finding is:

> ***There is a smooth, convex, and quadratic problem such that local SGD can not get arbitrarily close to the shared optimum of the clients with finite communication.***

Our lower bound precludes dominance of local SGD over mini-batch SGD under the following heterogeneity assumption considered by several existing works (Koloskova et al., 2020; Woodworth et al., 2020b; Glasgow et al., 2022).

**Assumption 1 (Bounded First-Order Heterogeneity at Optima)** *A set of objectives $\{F_m\}_{m\in[M]}$ satisfy $\zeta_\star$-first-order heterogeneity at the optima if for all minimizers of the average objective $x^\star \in \arg\min_{x\in\mathbb{R}^d} F(x)$,*

$$\frac{1}{M}\sum_{m\in[M]} \|\nabla F_m(x^\star)\|_2^2 \leq \zeta_\star^2.$$

We show that the above assumption, while a natural measure of data-heterogeneity, can not avoid some basic pathologies of distributed learning with convex objectives. Our lower bound also precludes any dominance of local SGD over mini-batch SGD under another similar assumption postulated by Wang et al. (2022) (Assumption 6 in Appendix C).

**2. Accelerated mini-batch SGD is min-max optimal when machines have shared optima.** We provide a new algorithm-independent lower bound (Theorem 11) under Assumption 1 (and Assumption 6), showing that accelerated mini-batch SGD is min-max optimal under this assumption! This conclusion is surprising, and while it further deepens the theory-practice disparity, it concludes a recent line of work on identifying the min-max optimal algorithm in the general convex setting under related first-order heterogeneity assumptions (Khaled et al., 2020; Koloskova et al., 2020; Woodworth et al., 2020b; Glasgow et al., 2022; Wang et al., 2022).

Together our lower bounds (Theorems 9 and 11) imply that existing first-order assumptions are insufficient to understand the effectiveness of local SGD, in that they do not characterize a measure of data heterogeneity, reducing which would allow local update steps to be useful. This motivates us to consider higher-order assumptions in addition to first-order ones.

**3. Local SGD shines under higher-order heterogeneity and smoothness assumptions.** We provide a new upper bound (Theorem 15) for local SGD improving the analysis of Woodworth et al. (2020b) by capturing the effect of second-order heterogeneity (Assumption 3) and third-order smoothness (Assumption 4). While our upper bound depends on the restrictive Assumption 2 of Woodworth et al. (2021), it relaxes local SGD's dependence on the assumption, i.e., our analysis implies a benefit of local updates with much larger data heterogeneity than that implied by the analysis of Woodworth et al. (2020b). This hints that higher-order assumptions indeed enhance the benefit of local updates. We explore this further by considering the special case when all the client objectives are strongly convex quadratics. In this setting, we provide a new upper bound (Theorem 22) for local SGD, underlining that Assumptions 1 and 3 control the discrepancy between the fixed point of local SGD and the minimizer of $F$ in problem (1).

**Notation.** We use $\cong$, $\preceq$, and $\succeq$ to refer to equality and inequality up to absolute numerical constants. We denote the set $\{i, i+1, \ldots, n\}$ by $[i, n]$ and when $i = 1$ by $[n]$. $\|.\|_2$ refers to the $L_2$ norm when applied to vectors in $\mathbb{R}^d$ and the spectral norm when applied to matrices in $\mathbb{R}^{d\times d}$.

## 2. Setting and Preliminaries

Note that an instance of Problem (1) can be characterized by the client distributions $\{\mathcal{D}_m \in \Delta(\mathcal{Z})\}_{m\in[M]}$ and a differentible loss function $f(.; z \in \mathcal{Z}) : \mathbb{R}^d \to \mathbb{R}$. We will denote the set of all such problem instances by $\mathcal{P}$. To further restrict the problems we will study, we assume that for all $m \in [M]$, the objective function $F_m$ is convex, and $H$-smooth, i.e., for all $m \in [M],\ x, y \in \mathbb{R}^d$,

$$F_m(x) \leq F_m(y) + \langle \nabla F_m(y), x - y\rangle + \frac{H}{2}\|x - y\|_2^2 \ . \tag{2}$$

| Type of Result (Reference) | Convergence bound on $\mathbb{E}[F(\hat{x}) - F(x^\star)] \preceq / \succeq$ |
|---|---|
| Local SGD Upper Bound (Koloskova et al., 2020) | $\boldsymbol{\frac{HB^2}{R}} + \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}}$ |
| Local SGD Lower Bound (Glasgow et al., 2022) | $\frac{HB^2}{KR} + \min\left\{\frac{\sigma B}{\sqrt{KR}}, \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}}\right\} + \frac{\sigma B}{\sqrt{MKR}} + \min\left\{\frac{\zeta_\star^2}{H}, \frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}}\right\}$ |
| Local SGD Lower Bound (Theorem 9) | $\boldsymbol{\frac{HB^2}{R}} + \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}}$ |
| Mini batch SGD Upper Bound (Dekel et al., 2012) | $\boldsymbol{\frac{HB^2}{R}} + \frac{\sigma B}{\sqrt{MKR}}$ |
| Acc. MB SGD Upper Bound (Ghadimi and Lan, 2012) | $\boldsymbol{\frac{HB^2}{R^2}} + \frac{\sigma B}{\sqrt{MKR}}$ |
| Algo. Independent Lower Bound (Woodworth et al., 2020b) | $\min\left\{\boldsymbol{\frac{HB^2}{R^2}}, \frac{\zeta_\star^2}{HR^2}, \frac{\zeta_\star B}{R}\right\} + \frac{\sigma B}{\sqrt{MKR}}$ |
| Algo. Independent Lower Bound (Theorem 11) | $\boldsymbol{\frac{HB^2}{R^2}} + \frac{\sigma B}{\sqrt{MKR}}$ |

Table 1: Summary of old and new results (up to logarithmic factors) for problem instances in the class $\mathcal{P}_{\zeta_\star}^{H,B,\sigma}$, i.e., under Assumption 1. Throughout we assume $\frac{\sigma B}{\sqrt{MKR}} \leq HB^2$.

We also assume that each machine $m \in [M]$ computes stochastic gradients of $F_m$ in the IC setting by sampling from its distributions $z \sim \mathcal{D}_m$, and that for all $x \in \mathbb{R}^d$,

$$\mathbb{E}_{z\sim\mathcal{D}^m}\left[\nabla f(x;z)\right] = \nabla F_m(x), \text{ and } \mathbb{E}_{z\sim\mathcal{D}^m}\left[\|\nabla f(x;z) - \nabla F_m(x)\|_2^2\right] \leq \sigma^2 \ . \tag{3}$$

Finally, we assume that the average objective $F$ has bounded optima, i.e.,

$$\|x^\star\|_2 \leq B, \ \forall\, x^\star \in S^\star := \arg\min_{x\in\mathbb{R}^d} F(x). \tag{4}$$

We will denote the class of all the problems satisfying these assumptions[1] by $\mathcal{P}^{H,B,\sigma}$. With this, we are ready to describe our algorithms with intermittent communication. We can write the update for round $r \in [R]$ for local SGD (with $x_{r,0}^m = x_{r-1}, \ \forall\, m \in [M]$) and mini-batch SGD as follows,

**Local SGD**

$$\begin{aligned} &\forall\, m \in [M], k \in [0, K-1] : \\ &\qquad g_{r,k}^m = \nabla f(x_{r,k}; z_{r,k}^m), \ z_{r,k}^m \sim \mathcal{D}_m, \\ &\qquad x_{r,k+1}^m = x_{r,k}^m - \eta g_{r,k}^m, \\ &x_r = x_{r-1} + \frac{\beta}{M}\sum_{m\in[M]}\left(x_{r,K}^m - x_{r-1}\right), \end{aligned} \tag{5}$$

**Mini-batch SGD**

$$\begin{aligned} &\forall\, m \in [M], k \in [0, K-1] : \\ &\qquad g_{r,k}^m = \nabla f(x_{r-1}; z_{r,k}^m), \ z_{r,k}^m \sim \mathcal{D}_m, \\ &x_r = x_{r-1} - \frac{\beta}{M}\sum_{m\in[M],k\in[0,K-1]} g_{r,k}^m. \end{aligned} \tag{6}$$

For local SGD, $\eta$ is referred to as the inner step size, while $\beta$ is the outer step size. Setting $\beta = 1$ recovers ***"vanilla local SGD"*** with a single step size which has been analyzed in several earlier

1. Note that when $\sigma = 0$, i.e., in the noiseless setting, then the problem instance is characterized by $\{F_m\}_{m\in[M]}$.

works (Stich, 2018; Dieuleveut and Patel, 2019; Khaled et al., 2020; Woodworth et al., 2020a). The main difference in the mini-batch update compared to vanilla local SGD is that its local gradient is computed at the same point for the entire communication round. Due to this, mini-batch SGD is not impacted by data heterogeneity, as it optimizes $F$ without getting affected by the multi-task nature of problem (1) (for more discussion, see Woodworth et al. (2020b)). This, however, is also the reason why local SGD can intuitively outperform mini-batch SGD: because it has more effective updates compared to mini-batch SGD (for e.g., without noise, i.e., $\sigma = 0$, $KR$ v/s $R$ updates).

When the data heterogeneity is "low", the iterates on different machines would not go too far between communication rounds, a.k.a., local SGD will have a small "consensus error". In the extreme homogeneous setting, the iterates on different machines will be the same in expectation. We will denote the class of homogeneous problems by $\mathcal{P}_{hom}^{H,B,\sigma} \subset \mathcal{P}^{H,B,\sigma}$. Most heterogeneity assumptions are variants of first-order assumptions, which impose restrictions on the gradients of functions and thus aid the analysis of local SGD by controlling how far the machines' iterates can go between two communication rounds. These assumptions can be generalized as follows.

**Assumption 2 (Bounded First-Order Heterogeneity)** *A set of objectives $\{F_m\}_{m\in[M]}$ satisfy $\zeta$-first-order heterogeneity on the domain $\mathcal{X} \subseteq \mathbb{R}^d$ if,*

$$\sup_{x\in\mathcal{X}, m\in[M]} \|\nabla F_m(x) - \nabla F(x)\|_2^2 \leq \zeta(\mathcal{X})^2.$$

We will refer to problems satisfying Assumption 2 on $\mathcal{X}$ by $\mathcal{P}_{\zeta(\mathcal{X})}^{H,B,\sigma}$ and omit the dependence on $\mathcal{X}$ when it is clear from the context. Woodworth et al. (2020b) showed an advantage of vanilla local SGD over mini-batch SGD under Assumption 2 on the unbounded domain $\mathcal{X} = \mathbb{R}^d$ (see Table 2). Unfortunately, assuming $\mathcal{X}$ to be $\mathbb{R}^d$ makes the above assumption too restrictive, for e.g., quadratic functions in the class $\mathcal{P}_{\zeta(\mathbb{R}^d)}^{H,B,\sigma}$ must have the same Hessian.

**Proposition 1** *Let $F_m(x) = \frac{1}{2}x^T A_m x + b_m^T x + c_m$ for all $m \in [M]$. If $\{F_m\}_{m\in[M]}$ satisfy Assumption 2 for $\mathcal{X} = \mathbb{R}^d$ for $\zeta < \infty$ then for any two machines $m, n \in [M]$, $A_m = A_n$.*

**Remark 2 ($\mathcal{P}_{hom}^{H,B,\sigma=0} \approx \mathcal{P}_{\zeta(\mathbb{R}^d)}^{H,B,\sigma=0}$)** *Essentially, Assumption 2 on $\mathbb{R}^d$ only allows heterogeneity in linear terms. In the noiseless setting, i.e., when $\sigma = 0$, such heterogeneity can be adjusted for using one round of communication at the beginning of optimization by computing the gradient at $0$ on all machines and transmitting $\nabla F(0)$ to the machines. At each round, the machines add the correction $\nabla F(0) - \nabla F_m(0)$ to their gradients and behave as if the problem were homogeneous!*

The restrictiveness of Assumption 2 on $\mathbb{R}^d$ is precisely why several papers (Khaled et al., 2020; Karimireddy et al., 2020b; Koloskova et al., 2020) have considered relaxed versions of the assumption. A natural relaxation is to assume Assumption 2 on $\mathcal{X} = S^\star$, the set of optima of $F$, which is precisely Assumption 1. We will denote $\zeta_\star := \zeta(S^\star)$, and define the problem class of problems satisfying Assumption 1 by $\mathcal{P}_{\zeta_\star}^{H,B,\sigma} := \mathcal{P}_{\zeta(S^\star)}^{H,B,\sigma}$. We can make several interesting remarks related to interpretations of Assumption 1.

**Remark 3 ($\mathcal{P}_{hom}^{H,B,\sigma=0} \approx \mathcal{P}_{\zeta(\mathbb{R}^d)<\infty}^{H,B,\sigma=0} \subset \mathcal{P}_{\zeta_\star=0}^{H,B,\sigma=0}$)** *Note that $\zeta_\star = 0$ does not imply we are in the homogeneous regime—all it says is that the machines should have some shared optimum. This is in contrast to Assumption 2 with $\mathcal{X} = \mathbb{R}^d$, where $\zeta = 0$ implies that $F_m$'s have to be the same except for constants. Thus, assuming $\zeta_\star < \infty$ (or even $\zeta_\star = 0$) does allow for interesting data heterogeneity.*

**Remark 4 (Approximate Simultaneous Realizability)** *In a learning setting, Assumption 1 can be interpreted as an approximate* simultaneous realizability *assumption for $\mathcal{D}_m$'s, measuring how good the solution for the average objective $F$ for any individual machine. Besides literature on data heterogeneity in distributed optimization, the simultaneous realizability assumption also appears in the collaborative PAC learning literature (Blum et al., 2017; Nguyen and Zakynthinou, 2018; Haghtalab et al., 2022) and in the context of defections in federated learning (Han et al., 2023).*

**Remark 5 (Local SGD's Fixed Point)** *Denote the set of optima of machine $m \in [M]$ by $S_m^\star := \arg\min_{x\in\mathbb{R}^d} F_m(x)$. Assume the machines have some shared optimum $x^\star \in \cap_{m\in[M]} S_m^\star$, then in fact all their optima should be shared, i.e., $S^\star = \cap_{m\in[M]} S_m^\star$. This observation implies that all points in $S^\star = \cap_{m\in[M]} S_m^\star$ are fixed points of local gradient descent with any $K \geq 1$ (Pathak and Wainwright, 2020). Interestingly, beyond the setting of shared optima, for $K > 1$, there is no simple closed form for the fixed points of local gradient descent. Since $\zeta_\star = 0$ implies that the machines do have a shared optimum, Assumption 1 intimately "controls" the convergence behavior of local GD.*

Unfortunately, no known analysis has shown that local SGD improves over mini-batch SGD on the class $\mathcal{P}_{\zeta_\star}^{H,B,\sigma}$ (see Table 1). Based on the best known lower bound by Glasgow et al. (2022) (see Table 1), it is still *possible* that local SGD could improve upon (accelerated) mini-batch SGD when $\zeta_\star$ and $\sigma$ are small, but $K$ is large. For instance, in the extreme case when $\sigma = \zeta_\star = 0$, even if $K \to \infty$ but $R$ is small, accelerated large mini-batch SGD will not have zero function sub-optimality. In contrast, if the lower bound by Glasgow et al. (2022) were tight, local SGD would get zero function sub-optimality in such a regime. This raises the following question:

***Can local SGD dominate mini-batch SGD on the class $\mathcal{P}_{\zeta_\star}^{H,B,\sigma}$ for small $\zeta_\star$?***

We will answer this question negatively in the next section, and also establish the min-max optimality of mini-batch SGD amongst zero-respecting algorithms in the IC setting on the class $\mathcal{P}_{\zeta_\star}^{H,B,\sigma}$.

## 3. Middling Utility of First-order Heterogeneity Assumptions in the Convex Setting

Our lower bounds in this section are based on the key insight presented in Remark 3, i.e., even if the machines have a shared optimum, i.e., $S^\star \neq \emptyset$ (see Remark 4), the problem instance need not be homogeneous. As a motivating example, assume there are two machines, we do not have any noise, i.e., $\sigma = 0$ and the objectives of the machines for all $x = (x[1], x[2]) \in \mathbb{R}^2$ are given by,

$$F_1(x) = \frac{H}{2}\left(x[1] - x^\star[1]\right)^2 \text{ and } F_2(x) = \frac{H}{2}\left(x[2] - x^\star[2]\right)^2, \tag{7}$$

where $x^\star = (x^\star[1], x^\star[2]) \in \mathbb{R}^2$ is the unique optimum of $F(x) = \frac{H}{4}\left\|x - x^\star\right\|_2^2$. Note that each machine's objective is $H$-smooth and satisfies Assumption 1 with $S^\star = \{x^\star\}$ and $\zeta_\star = 0$. Assuming we run local SGD on both machines initialized at $(0,0)$, then the iterate after $R$ rounds is given by,

$$x_R = x^\star\left(1 - \left(1 - \frac{\beta}{2}\left(1 - (1-\eta H)^K\right)\right)^R\right). \tag{8}$$

Let us first consider vanilla local SGD, i.e., $\beta = 1$. The above expression (proved in Appendix B) implies that even if $K \to \infty$, $x_R$ **does not** converge to $x^\star$ for finite $R$. This implies that the lower bound from Glasgow et al. (2022), which goes to $0$ when $K \to \infty$, must be loose.

**Remark 6 (The Role of the Outer Step-size)** *Note that for Example 7, if we set $\beta = 2$, $x_R$ will converge to $x^\star$ in a single communication round when $K \to \infty$! Thus, at least in this example, we can not rule out the lower bound by Glasgow et al. (2022) if we use the correct outer step size. This example re-emphasizes the role of outer step-size (Charles and Konecny, 2020; Jhunjhunwala et al., 2023) and provides a separation between vanilla local SGD and local SGD with two step-sizes.*

In the following proposition, we extend this idea to obtain a lower bound for any outer step size $\beta$.

**Proposition 7** *For any $K \geq 2, R, M, H, B,$ and $\sigma = 0$ there exist $\{F_m\}_{m\in[M]}$ in the problem class $\mathcal{P}_{\zeta_\star=0}^{H,B,\sigma=0}$, s.t. the local GD iterate $x_R$ with any step-size $\eta, \beta > 0$, and initialized at zero has:*

$$\mathbb{E}\left[F(\hat{x}_R)\right] - F(\boldsymbol{x}^\star) \succeq \frac{HB^2}{R}\,.$$

The above proposition formalizes the idea presented in our motivating example (7), showing that local GD can not reach arbitrary function sub-optimality with finite $R$ but infinite $K$ for any $\beta$. In fact, we also use a quadratic construction to prove the proposition. This lower bound also precludes any domination over gradient descent on $F$ with $R$ updates, which is indeed the noiseless variant of mini-batch SGD (see update (6)). Finally, note that the proposition can not benefit from simultaneous realizability, i.e., non-empty $S^\star$. This highlights the hardness of general convex optimization in the intermittent communication model. To prove the above proposition for arbitrary step sizes, we reduce the local GD updates over a single communication round to a single gradient descent (GD) update on $F$ and then use the following lemma (proved in Appendix B) for GD.

**Lemma 8** *Let $F(x)$ be a convex quadratic function whose Hessian has top eigenvalue $H$, bottom eigenvalue $\mu$, and condition number $\kappa = \frac{H}{\mu} \geq 6$. Let $\hat{x}_R$ be the $R^{th}$ gradient descent iterate initialized at zero. Then for any $B$, there exists some $x^\star$ with $\|x^\star\|_2 \leq B$ such that for any step size $\eta$, $F(\hat{x}_R) - F(x^\star) \geq HB^2\frac{1}{\kappa}e^{-R/\kappa}$. In particular, if $\kappa = \Omega(R)$, then $F(\hat{x}_R) - F(x^\star) \geq \Omega\left(\frac{HB^2}{R}\right)$.*

To get a lower bound in terms of $\zeta_\star$ from Assumption 2 as well as understand the effect of noise level $\sigma$, we combine proposition 7 with the previous lower bound by Glasgow et al. (2022), to get the following lower bound result.

**Theorem 9** *For any $K \geq 2, R, M, H, B, \sigma, \zeta_\star$, there exists a problem instance in $\mathcal{P}_{\zeta_\star}^{H,B,\sigma}$ such that the final iterate $x_R$ of local SGD initialized at zero with any step size satisfies:*

$$\mathbb{E}\left[F(x_R)\right] - F(x^\star) \succeq \frac{HB^2}{R} + \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}}.$$

Combining Theorem 9 with the upper bound by Koloskova et al. (2020), we characterize the optimal convergence rate for local SGD for problems in class $\mathcal{P}_{\zeta_\star}^{H,B,\sigma}$. Thus, our lower bound concludes the line of work trying to analyze local SGD under the first-order heterogeneity assumption on the set $S^\star$ (Koloskova et al., 2020; Woodworth et al., 2020b; Glasgow et al., 2022).

**Remark 10 (Proposed Assumption of Wang et al. (2022))** *Proposition 7 also implies that the first-order assumption of Wang et al. (2022) (see Assumption 6 and discussion in Appendix C) can not be used to show a domination of local SGD over mini-batch SGD. This is because when $S^\star \neq \emptyset$, the data heterogeneity assumption of Wang et al. (2022) implies zero heterogeneity.*

Overall, the results of this section imply that none of the existing assumptions besides the Assumption 2 on the unbounded domain $\mathbb{R}^d$ provides the necessary control on data heterogeneity, for local SGD to beat mini-batch SGD in the general convex setting.

### 3.1. The Min-max Optimality of Mini-batch SGD

While we can not show the effectiveness of local SGD on the class $\mathcal{P}_{\zeta_\star}^{H,B,\sigma}$, can we say which algorithm is min-max optimal for this class of problems? It turns out that the answer is surprisingly simple: accelerated large mini-batch SGD (Ghadimi and Lan, 2012)! To show this, we prove the following new algorithm-independent lower bound, which does not improve with a small $\zeta_\star$.

**Theorem 11 (Algorithm independent lower bound)** *For any $K \geq 2, R, M, H, B, \sigma, \zeta_\star$, there exists a problem instance in the class $\mathcal{P}_{\zeta_\star}^{H,B,\sigma}$, s.t. the final iterate $\hat{x}$ of any distributed zero-respecting algorithm (see Appendix B.5) initialized at zero with $R$ rounds of communication and $K$ stochastic gradient computations per machine per round satisfies,*

$$\mathbb{E}\left[F(\hat{x})\right] - F(x^\star) \succeq \frac{HB^2}{R^2} + \frac{\sigma B}{\sqrt{MKR}}\,. \tag{9}$$

The above lower bound fully characterizes the min-max complexity of distributed optimization on the class $\mathcal{P}_{\zeta_\star}^{H,B,\sigma}$, showing that accelerated mini-batch SGD is the min-max optimal algorithm. This conclusion is surprising, but it closes a recent line of work investigating the intermittent communication setting under Assumption 2 (Khaled et al., 2020; Karimireddy et al., 2020b; Koloskova et al., 2020; Woodworth et al., 2020b; Glasgow et al., 2022; Wang et al., 2022). The above lower bound also implies the min-max optimality of large mini-batch SGD under Assumption 6 proposed by Wang et al. (2022) (see Appendix C).

Compared to Woodworth et al. (2021) who showed that for problem class $\mathcal{P}_{hom}^{H,B,\sigma}$ (and by implication for $\mathcal{P}_{\zeta=0}^{H,B,\sigma}$), either accelerated SGD on a single machine or accelerated local SGD can improve over accelerated mini-batch SGD, we have shown that for the problem class $\mathcal{P}_{\zeta_\star=0}^{H,B,\sigma}$ such an improvement is not possible. Thus, we can not hope to show the benefit of using (accelerated) local SGD on any super-class of $\mathcal{P}_{\zeta_\star=0}^{H,B,\sigma}$. In other words, we need additional assumptions on top of Assumption 1 to avoid the pathologies of our hard instances. In the remainder of this paper we will explore which additional assumptions can help improve local SGD's convergence.

## 4. Beating Mini-batch SGD with Higher-order Heterogeneity and Smoothness

In the previous section, we showed that we can not prove the effectiveness of local SGD over mini-batch SGD if we consider the entire class $\mathcal{P}_{\zeta(S^\star)}^{H,B,\sigma} = \mathcal{P}_{\zeta_\star}^{H,B,\sigma}$. On the other hand, recall that Woodworth et al. (2020b) did show such a result over the "nearly homogeneous" class $\mathcal{P}_{\zeta(\mathbb{R}^d)}^{H,B,\sigma}$ (see Table 2). This prompts us to ask the following question:

> ***Can local SGD dominate mini-batch SGD on the class $\mathcal{P}_{\zeta(\mathbb{B}_2(D))}^{H,B,\sigma}$, where $\mathbb{B}_2(D)$ is the $L_2$ ball of radius $D$ around $0$ for some sufficiently large $D$, and small $\zeta(\mathbb{B}_2(D))$?***

Intuitively, if our problem has a low (but meaningful) heterogeneity, and we can ensure that the local SGD iterates never leave the ball $\mathbb{B}_2(D)$, for some large enough $D$, then we can recover the result of Woodworth et al. (2020b), to show the effectiveness of local SGD on a much larger problem class $\mathcal{P}_{\zeta(\mathbb{B}_2(D))}^{H,B,\sigma}$. To enable this, we would need the following second-order heterogeneity assumption, which controls how far the hessians on each machine can be from the average hessian:

| Type of Result (Reference) | Convergence bound on $\mathbb{E}[F(\hat{x}) - F(x^\star)] \preceq / \succeq$ |
|---|---|
| | $\mu$-Strongly Convex $F$ |
| Local SGD Upper Bound (Woodworth et al., 2020b) | $\frac{HB^2}{HKR+\lambda K^2R^2} + \frac{\sigma^2}{\mu MKR} + \frac{H\zeta^2}{\mu^2R^2} + \frac{H\sigma^2}{\mu^2KR^2}$ |
| Local SGD Upper Bound (Yuan and Ma, 2020) ($\star$) | $\exp\left(-\frac{\mu KR}{H}\right)HB^2 + \frac{\sigma^2}{\mu MKR} + \frac{Q^2\sigma^4}{\mu^5K^2R^4}$ |
| Local SGD Upper Bound (Theorem 27, Appendix D) | $\exp\left(-\frac{\mu KR}{H}\right)HB^2 + \frac{\sigma^2}{\mu MKR} + \frac{Q^2\sigma^4}{\mu^5K^2R^4} + \frac{Q^2\zeta^4}{\mu^5R^4} + \frac{\tau^2\sigma^2}{\mu^3KR^2} + \frac{\tau^2\zeta^2}{\mu^3R^2}$ |
| | General Convex $F$ |
| Local SGD Upper Bound (Woodworth et al., 2020b) | $\frac{HB^2}{KR} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(H\zeta^2B^4)^{1/3}}{R^{2/3}} + \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}}$ |
| Local SGD Upper Bound (Yuan and Ma, 2020) ($\star$) | $\frac{HB^2}{KR} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(Q\sigma^2B^5)^{1/3}}{K^{1/3}R^{2/3}}$ |
| Local SGD Upper Bound (Theorem 15) | $\frac{HB^2}{KR} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(\tau\sigma B^3)^{1/2}}{K^{1/4}R^{1/2}} + \frac{(\tau\zeta B^3)^{1/2}}{R^{1/2}} + \frac{(Q\sigma^2B^5)^{1/3}}{K^{1/3}R^{2/3}} + \frac{(Q\zeta^2B^5)^{1/3}}{R^{2/3}}$ |
| Alg. Independent Lower Bound (Woodworth et al., 2021) ($\star$) | $\frac{HB^2}{K^2R^2} + \min\left\{\frac{\sigma B}{\sqrt{MKR}}, HB^2\right\} + \min\left\{\frac{HB^2}{R^2\log^2 M}, \frac{\sqrt{Q\sigma}B^2}{K^{1/4}R^2\log^{7/4}M}, \frac{\sigma B}{\sqrt{KR}}\right\}$ |

Table 2: Summary of old and new results for the problem class $\mathcal{P}_\tau^{H,Q,B,\sigma} \cap \mathcal{P}_{\zeta(\mathbb{B}_2(D))}^{H,B,\sigma}$ for $\mu$-strongly-convex and convex average objective $F$. Rows with ($\star$) include results for the homogeneous setting.

**Assumption 3 (Bounded Second-order Heterogeneity)** *A set of doubly-differentiable objectives $\{F_m\}_{m\in[M]}$ satisfy $\tau$-second-order heterogeneity if for all $x \in \mathbb{R}^d$,*

$$\sup_{m,n\in[M]} \left\|\nabla^2 F_m(x) - \nabla^2 F_n(x)\right\|_2^2 \leq \tau^2.$$

**Remark 12** *Assumption 3 has recently received attention in the non-convex setting (Karimireddy et al., 2020a; Murata and Suzuki, 2021; Patel et al., 2022). Patel et al. (2022) show that it is possible to show the dominance of local update algorithms, essentially local SGD with variance reduction over mini-batch algorithms when $\tau$ is much smaller than $H$.*

Assumption 3 can always be satisfied by choosing $\tau \geq 2H$ for smooth functions. We will denote the problems satisfying Assumption 3 by $\mathcal{P}_\tau^{H,B,\sigma}$. Given Assumption 3, we note that on the set $\mathbb{B}_2(D)$, we can bound $\zeta$ in terms of $\tau, \zeta_\star$ and $D$ using the following upper bound (proved in Appendix D).

**Proposition 13** *Given a problem instance in the class $\mathcal{P}_{\zeta_\star}^{H,B,\sigma} \cap \mathcal{P}_\tau^{H,B,\sigma}$ with objectives $\{F_m\}_{m\in[M]}$,*

$$\sup_{x\in\mathbb{B}_2(D), m\in[M]} \|\nabla F_m(x) - \nabla F(x)\|_2^2 \leq (\zeta_\star + \tau D)^2.$$

The above proposition implies that for some level of first-order heterogeneity $\zeta$ when $\tau$ is small, we can allow for a much larger $D$. Alternatively, if we know our algorithm will be inside a ball $\mathbb{B}_2(D)$, the smaller the second-order heterogeneity of our problem, the smaller the bound on its first-order heterogeneity. In the extreme case of $\tau = 0$, we can replace $\zeta$ with $\zeta_\star$ (i.e., the converse of Proposition 1)! And as a sanity check, in the homogeneous setting the right hand side is zero. Finally, we will also define a third-order smoothness assumption in synergy with Assumption 3.

**Assumption 4 (Third-order Smoothness)** *The average function $F$ has a $Q$ Lipschitz hessian if*

$$\forall\, x, y \in \mathbb{R}^d, \quad \left\|\nabla^2 F(x) - \nabla^2 F(y)\right\|_2 \leq Q \left\|x - y\right\|_2 \ .$$

The above assumption intuitively upper bounds the gap between a function and its second-order approximation and is meaningful for smooth function only when for a given $x, y \in \mathbb{R}^d$, $H \gg Q\left\|x-y\right\|_2$, i.e., the bound implied by smoothness is loose. For sanity check, note that when $Q = 0$, then $F$ must be a quadratic function. We will denote the problems satisfying Assumptions 3 and 4 by $\mathcal{P}_\tau^{H,Q,B,\sigma}$.

**Remark 14** *For quadratic problems in the class $\mathcal{P}_{hom}^{H,B,\sigma}$, Woodworth et al. (2020a) showed that accelerated local SGD is min-max optimal. In particular, their result implies that local SGD can achieve an arbitrary accuracy for a fixed number of communication rounds $R$ by increasing the number of local steps $K$ per round. Given this result, several works (Yuan and Ma, 2020; Bullins et al., 2021) have tried to use higher-order smoothness assumptions to interpolate between the min-max optimal bounds for the quadratic and convex homogeneous setting (Table 2).*

We can obtain the following informal convergence result for the problem class $\mathcal{P}_\tau^{H,Q,B,\sigma} \cap \mathcal{P}_{\zeta_\star}^{H,B,\sigma}$. We state the full version of the above bound, and the step-size used to obtain it, in Appendix D.

**Theorem 15 (Informal)** *For any $K, R, M \geq 1$ and $H, B, Q, \sigma, \tau \geq 0$ consider a problem instance in the class $\mathcal{P}_\tau^{H,Q,B,\sigma} \cap \left(\mathcal{P}_{\zeta(\mathbb{B}_2(D))}^{H,B,\sigma} \cup \mathcal{P}_{\zeta_\star}^{H,B,\sigma}\right)$, where we assume $D$ is large-enough such that the local SGD iterates stay inside the ball $\mathbb{B}_2(D)$. Then, for a fixed inner stepsize of $\eta \leq \frac{1}{2H}$, and outer stepsize of $\beta = 1$, we have the following convergence rate (up to logarithmic factors) for local SGD (initialized at zero):*

$$\mathbb{E}\left[F\left(\frac{1}{MKR}\sum_{m=1}^{M}\sum_{r=1}^{R}\sum_{k=0}^{K-1} x_{r,k}^m\right) - F(x^\star)\right] \preceq \frac{HB^2}{KR} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(\tau\sigma B^3)^{1/2}}{K^{1/4}R^{1/2}} + \frac{(Q\sigma^2 B^5)^{1/3}}{K^{1/3}R^{2/3}}$$
$$+ \min\left\{\frac{(\tau\zeta B^3)^{1/2}}{R^{1/2}} + \frac{(Q\zeta^2 B^5)^{1/3}}{R^{2/3}}, \frac{(\tau\zeta_\star B^3)^{1/2}}{R^{1/2}} + \frac{(Q\zeta_\star^2 B^5)^{1/3}}{R^{2/3}} + \frac{(\tau^2 D B^3)^{1/2}}{R^{1/2}} + \frac{(Q\tau^2 D^2 B^5)^{1/3}}{R^{2/3}}\right\} .$$

**Remark 16** *Note that the bound implied by Proposition 13 is a worst-case bound for problems in the class $\mathcal{P}_{\zeta_\star}^{H,B,\sigma} \cap \mathcal{P}_\tau^{H,B,\sigma}$. And in general, there might exist an upper bound $\zeta(\mathbb{B}_2(D)) \ll \zeta_\star + \tau D$. For this reason, we have stated our upper bounds in terms of $\zeta$ as well as only in terms of $\tau, \zeta_\star, D$.*

To prove Theorem 15 we first obtain a convergence rate in the strongly convex setting (Theorem 27) and then use the standard convex to strongly convex reduction by adding appropriate regularization to our original problem. To the best of our knowledge, Theorem 15 is the only upper bound for local SGD for problems in the class $\mathcal{P}_\tau^{H,Q,B,\sigma} \cap \mathcal{P}_{\zeta(\mathbb{B}_2(D))}^{H,B,\sigma}$ even for $D = \infty$, i.e., $\mathcal{X} = \mathbb{R}^d$ in Assumption 2. Note that the upper bound **can** benefit from second-order heterogeneity, i.e., Assumption 3. In the extreme case, i.e., the homogeneous setting where $\tau, \zeta = 0$, our rate recovers the convergence rate by Yuan and Ma (2020) (see Table 2). To compare our result to existing results in the heterogeneous setting by Woodworth et al. (2020b), and to highlight the benefit of Assumptions 3 and 4, we will now compare requirements on $\zeta(\mathbb{B}_2(D))$ in order to reach some target sub-optimality $\epsilon$. For simplicity, we will assume that $K$ is large enough so that we can ignore all the terms in the convergence bound with $K$ in the denominator. Then the requirements on $\zeta$ are,

$$\zeta_{old} \preceq \frac{\epsilon^{3/2}R}{H^{1/2}B^2} \qquad \text{v/s} \qquad \zeta_{our} \preceq \min\left\{\frac{\epsilon^{3/2}R}{(QB)^{1/2}B^2}, \frac{\epsilon^2 R}{\tau B^3}\right\}. \tag{10}$$

In the regime when $Q, \tau$ are small, this highlights that our requirements are much less stringent on $\zeta$. The main limitation of our upper bound, however, is that we assume there exists a $D$ such that

$$\sup_{m\in[M], r\in[R], k\in[0,K-1]} \left\|x_{r,K}^m\right\|_2 \le D,$$

and prove the upper bound in terms of such a $D$. In fact, we require this for all (random) sequences of local SGD iterates. This is a strong assumption, but we believe it can be avoided by choosing appropriate $\eta, \beta$ in terms of problem-dependent quantities such as $\zeta_\star, \tau, Q$, so that we can control the norm of local SGD iterates (with high probability). We conjecture the following result.

**Conjecture 17 (Effectiveness of Local SGD)** *Local SGD can dominate mini-batch SGD over the problem class $\mathcal{P}_\tau^{H,Q,B,\sigma} \cap \mathcal{P}_{\zeta_\star}^{H,B,\sigma}$ in the regime when $\tau, QB \ll H$ and $\sigma, \zeta_\star$ are small.*

While we do not yet know how to prove this conjecture for the entire class $\mathcal{P}_\tau^{H,Q,B,\sigma} \cap \mathcal{P}_{\zeta_\star}^{H,B,\sigma}$, we will prove a very special case of this conjecture in the next section for strongly convex quadratic objectives on each machine which lie in the problem class $\mathcal{P}_\tau^{H,Q=0,B,\sigma} \cap \mathcal{P}_{\zeta_\star}^{H,B,\sigma}$. Quadratic objectives are analytically simpler, and we rely on a new fixed-point perspective to understand the behavior of local SGD.

**Remark 18 (Mini-batch SGD's indifference to our assumptions)** *In our setting, mini-batch SGD can not improve from additional assumptions on top of convexity and smoothness (at least in the worst case). To see this, we assume all the machines have the same objective $F$, a quadratic function (which will be determined later). In this simple setup, note that all data heterogeneity measures are zero, and since the objective is quadratic, $Q = 0$ also. Essentially, mini-batch SGD reduces to optimizing $F$ with $R$ mini-batch updates, each with batch size $MK$. Using Lemma 8 and standard sample complexity lower bounds for mean-estimation (also a quadratic problem), we can argue that the convergence rate for mini-batch SGD in this setting is $\frac{HB^2}{R} + \frac{\sigma B}{\sqrt{MKR}}$, which does not benefit from $\tau, \zeta_\star, Q = 0$. On the other hand, we conjecture above that local SGD can significantly benefit from these low measures of heterogeneity and sharpness. Thus, if the conjecture is true, we would identify a regime of low data heterogeneity where local SGD dominates mini-batch SGD.*

## 5. Fixed Point of Local SGD for Strongly Convex Objectives

Throughout this section, we will assume the objectives of each machine are of the following form:

$$F_m(x) = \frac{1}{2}(x - x_m^\star)^T A_m (x - x_m^\star),\ \forall m \in [M], \tag{11}$$

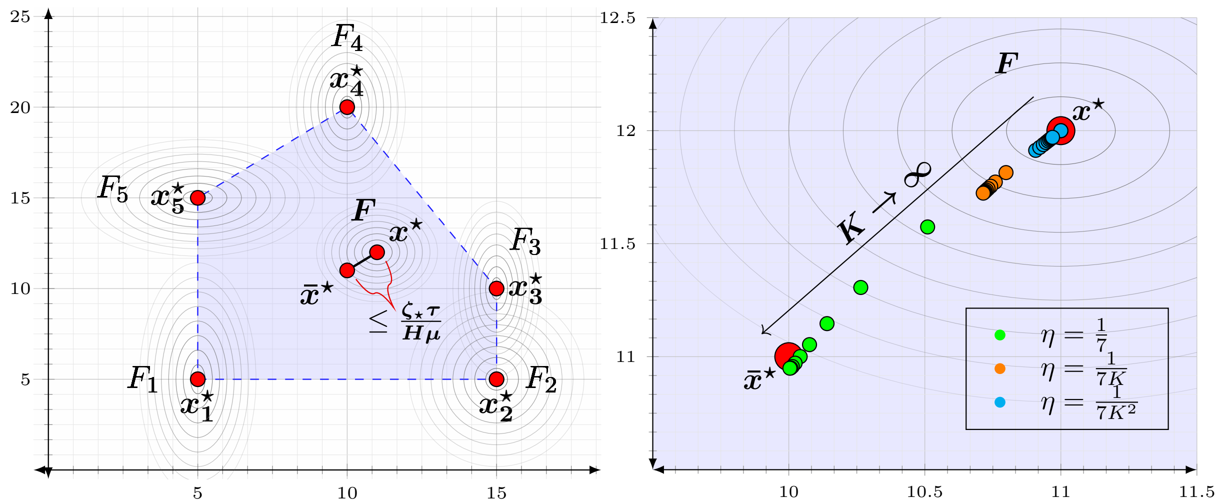


Figure 1: Illustration of a two-dimensional optimization problem with $M = 5$ machines, each with a 1-strongly convex, and 6-smooth objective. On the left figure, we draw the contour lines for each of the machine's objective as well as for the average objective. We also indicate the two relevant solution concepts $\bar{x}^\star$ and $x^\star$ in the same figure, noting that their distance is bounded by Proposition 21. On the right figure, we zoom into the convex hull of the machines' optima, noting the sequence of fixed points for local GD as a function of $\eta$ and increasing $K \in [10]$. We plot the fixed points for three different choices of $\eta$ each demonstrating a different trend for $\lim_{K\to\infty} x_\infty(K, \eta, \beta)$.

where $A_m$ has minimum and maximum eigenvalues $\mu, H > 0$ respectively and $x_m^\star$ is the unique minimizer of $F_m$. We will denote the optimum of the average objective $x^\star$ and the average of the optima of the machines $\bar{x}^\star$ as follows where $A := \frac{1}{M}\sum_{m\in[M]} A_m$ is the average hessian:

$$x^\star := \frac{1}{M}\sum_{m\in[M]} A^{-1}A_m x_m^\star, \quad \text{and} \quad \bar{x}^\star := \frac{1}{M}\sum_{m\in[M]} x_m^\star, \tag{12}$$

We will only consider the noiseless setting in this section. Given the quadratic form of the machines' objectives, we can also find a closed-form expression for the fixed point of local gradient descent.

**Proposition 19** *Consider an instance of the problem of the form (11) with strongly convex and smooth machines' objectives. Let $x_R(K, \eta, \beta)$ be the local gradient descent iterate after $R$ communication rounds, using exact gradient calls (i.e., $\sigma = 0$), $K$ local updates, and step-sizes, $\eta, \beta$. Then the fixed point for local GD (whenever it exists) is given by,*

$$x_\infty(K, \eta, \beta) := \lim_{R\to\infty} x_R(K, \eta, \beta) = \frac{1}{M}\sum_{m\in[M]} C^{-1}C_m x_m^\star,$$

*where $C_m = I - (I - \eta A_m)^K$ and $C = \frac{1}{M}\sum_{m\in[M]} C_m$.*

**Remark 20** *Note that the fixed point might not be defined for all choices of $\eta, \beta, K$. Intuitively, it should be well defined when $\eta, \beta$ are "small enough" relative to how $R$ increases. Also note that*

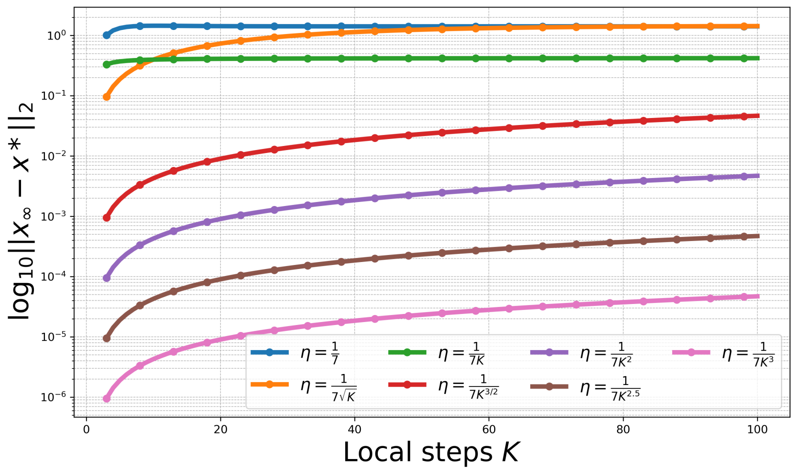


Figure 2: Illustration of the same distributed problem as Figure 1 to understand where the fixed point converges as $K$ grows. We consider 7 different choices of $\eta$ (as a function of $K$) and plot $\log \|x_\infty(K, \eta, 1) - x^\star\|_2$ as a function of $K \in [100]$. We notice that for $\eta > \frac{1}{HK}$, the fixed point goes to $\bar{x}^\star$ as $K$ increases, while for $\eta < \frac{1}{HK}$, the fixed point gets progressively closer to $x^\star$.

*since $\sigma = 0$, the fixed point is non-random. When $\sigma > 0$, we instead need to look at the expected local SGD iterate. Given the closed form for the fixed point, we can ask the question of how it is related to $x^\star$, the optimizer of $F$.*

Using the closed form of $x_\infty$ from Proposition 19 we can make two observations (c.f., Figure 1):

**1. The correct fixed point.** When $K = 1$, for any choice of $\eta, \beta > 0$ that ensures convergence of local GD, $x_\infty = x^\star$. This is unsurprising because with $K = 1$, local GD reduces to GD on the averaged objective, so the fixed point must be a stationary point of $F$, i.e., $x^\star$. We also observe in Figure 2 that as $\eta$ decreases as a function of $K$, and $K \to \infty$ we recover the correct fixed point. This is also expected, as the effect of local updates diminishes with smaller $\eta$.

**2. The incorrect fixed point.** On the other hand when $K \to \infty$ and we pick $\eta = \omega\left(\frac{1}{KH}\right)$ but $\eta < \frac{1}{H}$ we get that $\lim_{K\to\infty} x_\infty(K, \eta, \beta) = \bar{x}^\star$. This makes intuitive sense (at least for $\beta = 1$) because for a large step-size $\eta$, that also ensures convergence of GD on each machine, local GD between communication rounds would converge to machines' optima, and then averaging the optima would yield $\bar{x}^\star$. In fact, using Assumptions 3 and 7 (a strongly convex variant of Assumption 1), we can show that the distance between $\bar{x}^\star$, $x^\star$ and $x_\infty$ are bounded as follows (see Figure 1).

**Proposition 21** *Consider a problem instance of the form (11) with $\mu$-strongly convex and $H$-smooth machines' objectives satisfying Assumptions 3 and 7, then for $K > 1$,*

$$\|x^\star - \bar{x}^\star\|_2 \le \frac{\zeta_\star \tau}{H\mu} \quad \textit{and} \quad \|x_\infty(K, \eta, \beta) - \bar{x}^\star\|_2 \le \frac{\zeta_\star \tau}{H\mu} \cdot \frac{\eta\mu K (1 - \eta\mu)^{K-1}}{1 - (1 - \eta\mu)^K} \preceq \frac{\zeta_\star \tau}{H\mu} \ .$$

Using triangle inequality Proposition 21 bounds the fixed point discrepancy for local SGD. Combining it with a convergence guarantee for local SGD to its fixed point, we get the following convergence guarantee in the noise-less setting (compared to Wang et al. (2022) in Appendix C).

**Theorem 22 (Local GD with Strongly Convex Objectives)** *Consider a problem instance of the form (11) with $\mu$-strongly convex and $H$-smooth machines' objectives satisfying Assumptions 3 and 7, then for $\beta = 1$ and $\eta = \frac{1}{2H}$ we get the following convergence for local GD initialized at zero,*

$$\|x_R - x^\star\|_2 \preceq e^{-\frac{KR}{\kappa}} \cdot \frac{1-\left(\frac{1}{2}\right)^K}{1-\left(1-\frac{1}{2\kappa}\right)^K} B + \frac{\zeta_\star \tau}{H\mu} \ .$$

*Assuming $K \geq \frac{1}{-\log_2\left(1-\frac{1}{2\kappa}\right)}$ this simplifies to,*

$$\|x_R - x^\star\|_2 \preceq Be^{-\frac{KR}{\kappa}} + \frac{\zeta_\star \tau}{H\mu}.$$

Comparing this bound to the rate for GD (i.e., the noiseless variant of mini-batch SGD), we see that when $\zeta_\star, \tau$ are small local GD has a much better optimization term v/s the optimization term $Be^{-R/\kappa}$ of GD. This implies that we have identified a $\mu$-strongly sub-class of $\mathcal{P}_\tau^{H,Q,B,\sigma=0} \cap \mathcal{P}_{\zeta_\star}^{H,B,\sigma=0}$, where local GD dominates GD. This further motivates us to believe that Conjecture 17 is correct. We discuss our analysis' connections to two-stage optimization algorithms in Section E.6.

### 5.1. Interpreting the Heterogeneity Assumptions

We have introduced several heterogeneity assumptions in this paper. We end this discussion by considering a simple linear regression model to highlight how these assumptions might appear in real-world problems relating to actual distribution shifts between clients. We will consider a linear regression problem, where each data point is a co-variate label pair $z_m := (\beta_m, y_m) \sim \mathcal{D}_m$ on each client such that: (i) the covariates are sampled from a Gaussian distribution, i.e., $\beta_m \sim \mathcal{N}(\mu_m, \Sigma_m) \in \mathbb{R}^d$; and (ii) the labels are generated as $y_m \sim \langle x_m^\star, \beta_m \rangle + \mathcal{N}(0, \sigma_{noise}^2) \in \mathbb{R}$ for some ground truth model $x_m^\star \in \mathbb{R}^d$ on machine $m$. The goal of each client is to minimize the mean squared error, i.e., $f(x; (\beta_m, y_m)) = \frac{1}{2}\left(y_m - \langle x, \beta_m \rangle\right)^2$ and the expected loss function is given by,

$$F_m(x) = \frac{1}{2}(x - x_m^\star)^T \left(\mu_m \mu_m^T + \Sigma_m\right)(x - x_m^\star) + \frac{1}{2}\sigma_{noise}^2 \ ,$$

which clearly fits in the model of smooth and strongly convex quadratic functions we have considered in this section, as long as $\mu_m\mu_m^T + \Sigma_m$ is positive semi-definite and $\left\|\mu_m\mu_m^T + \Sigma_m\right\|_2 \leq H$. Interestingly, Assumptions 3 and 7 have a very nice interpretation in this setting,

- Note that $\forall m, n \in [M]$, $\|A_m - A_n\|_2 \leq (\|\mu_m\|_2 + \|\mu_n\|_2)\|\mu_m - \mu_n\|_2 + \|\Sigma_m - \Sigma_n\|_2 =: \tau_{m,n}$. Thus this problem satisfies Assumption 3 with $\tau^2 := \sup_{m,n} \tau_{m,n}^2$ which in turn can be thought of as a measure of the ***"co-variate shift"*** between the machines.
- Similarly, note that $\forall m, n \in [M]$, $\|x_m^\star - x_n^\star\|_2^2 \leq 2\|x_m^\star - x^\star\|_2^2 + 2\|x_m^\star - x^\star\|_2^2$, for $x \in S^\star$. Thus Assumption 7 controls the average ***"concept-shift"*** between the clients as,

$$\frac{1}{M^2}\sum_{m,n\in[M]} \|x_m^\star - x_n^\star\|_2^2 \leq \frac{4}{M}\sum_{m\in[M]} \|x_m^\star - x^\star\|_2 \leq 4\frac{\zeta_\star^2}{H^2}.$$

In light of this interpretation, note that Theorem 22 implies that the local SGD can approach a fast convergence rate, i.e., not suffer due to the fixed point discrepancy, as long as either the co-variate shift ($\tau$) or the concept shift ($\zeta_\star$) is low. We believe this provides a useful interpretation for these heterogeneity assumptions, and this simple learning problem might pave the way to a better understanding of distributed learning algorithms.

## Acknowledgments

Preliminary versions of this work also appeared at the Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities workshop at ICML 2023 (Patel et al., 2023) and the 15th International OPT Workshop on Optimization for Machine Learning at NeurIPS 2023 (Zindari et al., 2023). We thank the anonymous reviewers at COLT, ICML, and NeurIPS, who helped significantly improve the paper. We also thank Blake Woodworth, Anastasia Koloskova, and Brian Bullins for several useful discussions. KKP was supported through the NSF TRIPOD Institute on Data, Economics, Algorithms and Learning (IDEAL) and other awards from DARPA and NSF. AZ and SUS acknowledge partial funding from a Google Scholar Research Award.

## References

Avrim Blum, Nika Haghtalab, Ariel D Procaccia, and Mingda Qiao. Collaborative pac learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Brian Bullins, Kumar Kshitij Patel, Ohad Shamir, Nathan Srebro, and Blake E Woodworth. A stochastic newton algorithm for distributed convex optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *Mathematical Programming*, 184(1):71–120, 2020.

Zachary Charles and Jakub Konecny. On the outsized importance of learning rates in local update methods. *arXiv preprint arXiv:2007.00878*, 2020.

Zachary Charles, Zachary Garrett, Zhouyuan Huo, Sergei Shmulyian, and Virginia Smith. On large-cohort training for federated learning. *Advances in neural information processing systems*, 34: 20461–20475, 2021.

Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1), 2012.

Aymeric Dieuleveut and Kumar Kshitij Patel. Communication trade-offs for local-sgd with large step size. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Margalit R Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (local sgd) and continuous perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 9050–9090. PMLR, 2022.

Gene H Golub. Cme 302: Numerical linear algebra fall 2005/06 lecture 10. 2005.

Nika Haghtalab, Michael Jordan, and Eric Zhao. On-demand sampling: Learning optimally from multiple distributions. *Advances in Neural Information Processing Systems*, 35:406–419, 2022.

Minbiao Han, Kumar Kshitij Patel, Han Shao, and Lingxiao Wang. On the effect of defections in federated learning and how to prevent them. *arXiv preprint arXiv:2311.16459*, 2023.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Charlie Hou, Kiran K Thekumparampil, Giulia Fanti, and Sewoong Oh. Fedchain: Chained algorithms for near-optimal communication cost in federated learning. *arXiv preprint arXiv:2108.06869*, 2021.

Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. Fedexp: Speeding up federated averaging via extrapolation. *arXiv preprint arXiv:2301.09604*, 2023.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. corr. *arXiv preprint arXiv:1912.04977*, 2019.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020a.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020b.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020.

Prashant Khanduri, Pranay Sharma, Haibo Yang, Mingyi Hong, Jia Liu, Ketan Rajawat, and Pramod Varshney. Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.

Dmitry Kovalev, Konstantin Mishchenko, and Peter Richtárik. Stochastic newton and cubic newton methods with simple local linear-quadratic rates. *arXiv preprint arXiv:1912.01597*, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

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.

Ryan Mcdonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon Mann. Efficient large-scale distributed training of conditional maximum entropy models. *Advances in neural information processing systems*, 22, 2009.

Brendan McMahan and Daniel Ramage. Federated learning: Collaborative machine learning without centralized training data, Apr 2017. URL https://ai.googleblog.com/2017/04/federated-[]learning-[]collaborative.html.

H Brendan McMahan, Eider Moore, Daniel Ramage, S Hampson, and B Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data (2016). *arXiv preprint arXiv:1602.05629*, 2016a.

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

Tomoya Murata and Taiji Suzuki. Bias-variance reduced local sgd for less heterogeneous federated learning. *arXiv preprint arXiv:2102.03198*, 2021.

Huy Nguyen and Lydia Zakynthinou. Improved algorithms for collaborative pac learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Kumar Kshitij Patel, Lingxiao Wang, Blake Woodworth, Brian Bullins, and Nathan Srebro. Towards optimal communication complexity in distributed non-convex optimization. In *Advances in Neural Information Processing Systems*, 2022.

Kumar Kshitij Patel, Margalit Glasgow, Lingxiao Wang, Nirmit Joshi, and Nathan Srebro. On the still unreasonable effectiveness of federated averaging for heterogeneous distributed learning. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023.

Reese Pathak and Martin J Wainwright. Fedsplit: An algorithmic framework for fast federated optimization. *Advances in neural information processing systems*, 33:7057–7066, 2020.

Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

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

Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

Jianyu Wang, Rudrajit Das, Gauri Joshi, Satyen Kale, Zheng Xu, and Tong Zhang. On the unreasonable effectiveness of federated averaging with heterogeneous data. *arXiv preprint arXiv:2206.04723*, 2022.

Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pages 10334–10343. PMLR, 2020a.

Blake E Woodworth, Jialei Wang, Adam Smith, Brendan McMahan, and Nati Srebro. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. *Advances in neural information processing systems*, 31, 2018.

Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020b.

Blake E Woodworth, Brian Bullins, Ohad Shamir, and Nathan Srebro. The min-max complexity of distributed stochastic convex optimization with intermittent communication. In *Conference on Learning Theory*, pages 4386–4437. PMLR, 2021.

Honglin Yuan and Tengyu Ma. Federated accelerated stochastic gradient descent. *Advances in Neural Information Processing Systems*, 33:5332–5344, 2020.

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

Haoyu Zhao, Zhize Li, and Peter Richtárik. Fedpage: A fast local stochastic gradient method for communication-efficient federated learning. *arXiv preprint arXiv:2108.04755*, 2021.

Ali Zindari, Ruichen Luo, and Sebastian U Stich. On the convergence of local sgd under third-order smoothness and hessian similarity. In *OPT 2023: Optimization for Machine Learning*, 2023.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010.

## Appendix A. Missing Details from Section 2

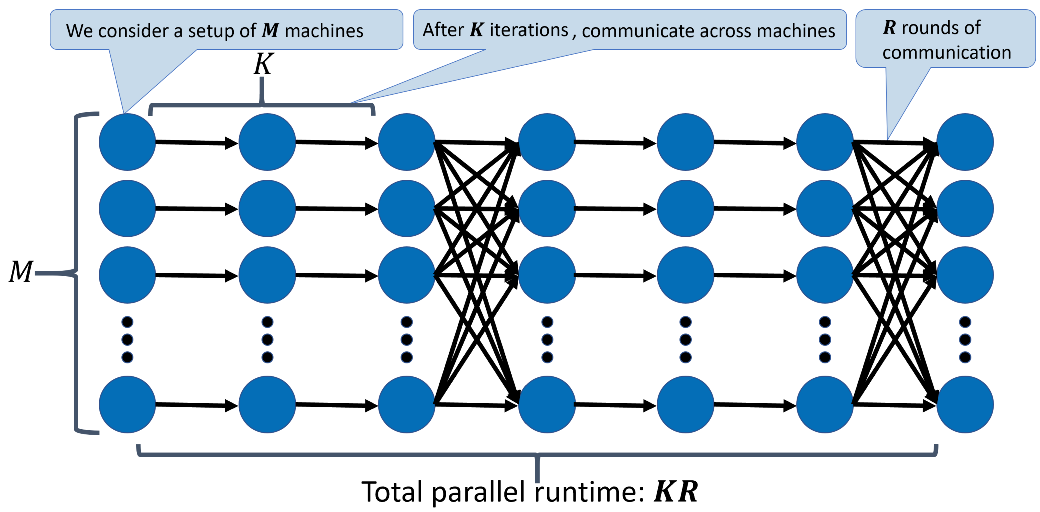


Figure 3: An Illustration of the intermittent communication setting.

### A.1. Another First-order Heterogeneity Assumption

Karimireddy et al. (2020b) looked at a following relaxed first-order heterogeneity assumption, but they were **not** able to show that local SGD dominates large mini-batch SGD under this assumption.

**Assumption 5 (Relaxed First-Order Heterogeneity Everywhere)** *A set of objectives* $\{F_m\}_{m\in[M]}$ *satisfy* $(\zeta_\star, D)$*-first-order heterogeneity everywhere if for all* $x \in \mathbb{R}^d$,

$$\frac{1}{M}\sum_{m\in[M]} \|\nabla F_m(x)\|_2^2 \le \zeta_\star^2 + D^2 \|\nabla F(x)\|_2^2 .$$

This assumption *"interpolates"* between Assumption 2 over the unbounded domain $\mathcal{X} = \mathbb{R}^d$ v/s at the optima $\mathcal{X} = S^\star$ for different values of $D$. Furthermore, the restrictiveness of Assumption 2 pointed out in Proposition 1 does not extend to this assumption as long as $D > 0$. While we did not look at this assumption in this paper, it is also promising for future work.

### A.2. Proof of Proposition 1

**Proof** Note the following for any $m \in [M]$ using triangle inequality,

$$\begin{aligned} sup_{x\in\mathbb{R}^d} \|\nabla F_m(x) - \nabla F(x)\|_2 &= sup_{x\in\mathbb{R}^d} \|(A_m - A)x + b_m - b\|_2 ,\\ &\ge sup_{x\in\mathbb{R}^d} \|(A_m - A)x\|_2 - \|b_m - b\|_2 . \end{aligned}$$

Denote the matrix $C_m := A_m - A = [c_{m,1}, \ldots, c_{m,d}]$ using its column vectors. Then take $x = \delta e_i$ where $e_i$ is the $i$-th standard basis vector to note in the above inequality,

$$sup_{x\in\mathbb{R}^d} \|\nabla F_m(x) - \nabla F(x)\|_2 \ge \delta \|(A_m - A)e_i\|_2 - \|b_m - b\|_2 ,$$

$$\geq \delta \left\|c_{m,i}\right\|_2 - \left\|b_m\right\|_2 - \left\|b\right\|_2.$$

Assuming $\left\|b_m\right\|_2, \left\|b\right\|_2$ are finite, since we can take $\delta \to \infty$ we must have $\left\|c_{m,i}\right\|_2 = 0$ for all $i \in [d]$ if $\zeta < \infty$. This implies that $c_{m,i} = 0$ for all $i \in [d]$, or in other words $A_m = A$. Since this is true for all $m \in [M]$, the machines must have the same Hessians, and thus they can only differ upto linear terms. ■

## Appendix B. Missing Details from Section 3

### B.1. Proof of Equation 7

**Proof** Note that the update on machine $m$ leading up to communication round $r$ is as follows for $k \in [0, K-1]$ and $m = 1$,

$$\begin{aligned}
&x^1_{r,k+1}[1] = x^1_{r,k}[1] - \eta H(x^1_{r,k}[1] - x^\star[1]),\\
&\Rightarrow x^1_{r,k+1}[1] = x^\star[1] + (1-\eta H)^{k+1}(x^1_{r,0}[1] - x^\star[1]),\\
&\Rightarrow x^1_{r,K}[1] = x^\star[1] + (1-\eta H)^{K}(x_{r-1}[1] - x^\star[1]),\\
&\Rightarrow x^1_{r,K}[1] - x_{r-1}[1] = (1-(1-\eta H)^{K})(x^\star[1] - x_{r-1}[1]).
\end{aligned}$$

On the second dimension, the iterates don't move at all for $m = 1$,

$$x^1_{r,K}[2] - x_{r-1}[2] = 0.$$

Writing a similar expression for the second machine and averaging these updates we get,

$$\frac{1}{2}\sum_{m\in[2]}(x^m_{r,K} - x_{r-1}) = \frac{1}{2}(1-(1-\eta H)^K)(x^\star - x_{r-1}).$$

This gives the update for communication round $r$ as follows,

$$\begin{aligned}
&x_r = x_{r-1} + \frac{\beta}{2}(1-(1-\eta H)^K)(x^\star - x_{r-1}),\\
&\Rightarrow x_r - x^\star = \left(1 - \frac{\beta}{2}(1-(1-\eta H)^K)\right)(x_{r-1} - x^\star),\\
&\Rightarrow x_R = x^\star + \left(1 - \frac{\beta}{2}(1-(1-\eta H)^K)\right)^R(x_0 - x^\star),\\
&\Rightarrow x_R = \left(1 - \left(1 - \frac{\beta}{2}(1-(1-\eta H)^K)\right)^R\right)x^\star,
\end{aligned}$$

which finishes the proof. ■

### B.2. Proof of Lemma 8

**Proof** Let $A$ be the Hessian of $F$. Observe that we have $F(x) - F(x^\star) = \frac{1}{2}(x - x^\star)^T A(x - x^\star)$.

Let $v_1$ and $v_2$ be the eigenvectors of norm 1 of $A$ with the greatest and least eigenvalues, respectively. Assume $x^\star := -B\left(\frac{v_1+v_2}{\sqrt{2}}\right)$, which ensures $\|x^\star\|_2 = B$. Then, solving for the GD iterates in closed form, we have

$$x_R - x^* = \frac{B}{\sqrt{2}}\left(1 - \eta H\right)^R v_1 + \frac{B}{\sqrt{2}}\left(1 - \eta\frac{H}{\kappa}\right)^R v_2. \tag{13}$$

Observe that if $\eta \geq \frac{3}{H}$, then the iterates explode and we have $F(x_R) \geq F(x_0) \geq \Omega\left(HB^2\right)$.

If $\eta \leq \frac{3}{H}$, then using the fact that $\kappa \geq 6$, we have

$$\begin{aligned} F(x_R) - F(x^\star) &\geq \frac{1}{2}\left(\frac{B}{\sqrt{2}}\left(1 - \frac{3}{\kappa}\right)^R v_2\right)^T A\left(\frac{B}{\sqrt{2}}\left(1 - \frac{3}{\kappa}\right)^R v_2\right) && (14)\\ &= \frac{B^2}{4}\left(1 - \frac{3}{\kappa}\right)^{2R} v_2^T A v_2 && (15)\\ &= \frac{B^2}{4}\left(1 - \frac{3}{\kappa}\right)^{2R} \frac{H}{\kappa} && (16)\\ &\geq \frac{HB^2}{4\kappa} e^{-12R/\kappa}. && (17) \end{aligned}$$

The result follows. ∎

### B.3. Proof of Proposition 7

**Proof** We will consider a two-dimensional problem in the noiseless setting for this proof, as we do not want to understand the dependence on $\sigma$ or $d$. Define $A_1 := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ and $A_2 := vv^T$, where $v = (\alpha, \sqrt{1-\alpha^2})$ and $\alpha \in (0,1)$. For even $m$, let

$$F_m(x) := \frac{H}{2}(x - x^*)^T A_1 (x - x^*) \ .$$

For odd $m$, let

$$F_m(x) := \frac{H}{2}(x - x^*)^T A_2 (x - x^*) \ .$$

Note that $A_1$ and $A_2$ are rank-1 and have eigenvalues 0 and 1, and thus they satisfy Assumption 2. Furthermore, both the functions have a shared optimizer $x^\star$. It is easy to verify that,

$$(I - \eta H A_i)^K = I - (1 - (1 - \eta H)^K) A_i =: I - \widetilde{\eta} H A_i,$$

where $\widetilde{\eta} := (1 - (1 - \eta H)^K)/H$. Note that the above property will be crucial for our construction, and we can not satisfy this property if our matrices are not ranked one. For any $x$, let us denote the centered iterate by $\tilde{x} := x - x^\star$. Then, for any $r \in [R]$ and $m \in [M]$ we have

$$\tilde{x}^m_{r,K} = (I - \eta H A_m)^K \tilde{x}_{r-1} = (I - \widetilde{\eta} H A_m)\tilde{x}_{r-1}.$$

Using this, we can write the updates between two communication rounds as,

$$\begin{aligned}
\tilde{x}_r &= \tilde{x}_{r-1} + \frac{\beta}{M} \sum_{m \in [M]} \left( \tilde{x}_{r,K}^m - \tilde{x}_{r-1} \right), \\
&= \tilde{x}_{r-1} - \frac{\beta}{M} \sum_{m \in [M]} \widetilde{\eta} H A_m \tilde{x}_{r-1}, \\
&= \left( I - \beta \widetilde{\eta} H A \right) \tilde{x}_{r-1}, \\
&= \tilde{x}_{r-1} - \beta \widetilde{\eta} \nabla F(x_{r-1}),
\end{aligned}$$

where we used that $F(x) = \frac{H}{2}(x - x^\star)^T A (x - x^\star)$, for $A = (1-a)A_1 + aA_2$ and

$$a := \begin{cases} 1/2 & \text{if } M \text{ is even}, \\ (M+1)/2M & \text{otherwise}. \end{cases}$$

This implies the iterates of local GD across communication rounds are equivalent to GD on $F(x)$ with step size $\beta(1-(1-\eta H)^K)/H$. Combining this observation with Lemma 8 about the function sub-optimality of gradient descent updates will finish the proof. To use the lemma, however, we need to verify our average function $F$ has condition number $\Omega(R)$. We can explicitly compute the eigenvalues of $A$ as follows,

$$\begin{aligned}
\lambda_1 &= \frac{1}{2} + \sqrt{\frac{1}{4} - (a - a^2)(1 - \alpha^2)}, \\
\lambda_2 &= \frac{1}{2} - \sqrt{\frac{1}{4} - (a - a^2)(1 - \alpha^2)}.
\end{aligned}$$

Note that $\lim_{\alpha \to 1} \lambda_1 = 1$, and $\lim_{\alpha \to 1} \lambda_2 = 0$ and thus $\lim_{\alpha \to 1} \lambda_1/\lambda_2 = \infty$. Since $\lambda_1/\lambda_2 = 1$ when $\alpha = 0$, by the intermediate value theorem, we can choose $\alpha$ to get $\kappa = \Omega(R)$ for the average objective $F$. Thus, we can use Lemma 8 and finish the proof. ■

Note that in the construction used in the above proof, the machines share an optimum point $x^\star$, which ensures that $\zeta_\star = 0$.

### B.4. Proof of Theorem 9

**Proof** To combine the hard instance in Proposition 7 with the previous hard instance of Glasgow et al. (2022) we simply place the two instances on disjoint co-ordinates increasing the dimensionality of our hard instance. This is a standard technique to combine lower bounds. To get rid of the terms with the minimum function in the lower bound of Glasgow et al. (2022), we note the following,

- $\frac{\sigma B}{\sqrt{KR}} \geq \frac{(H\sigma^2 B^4)^{1/3}}{K^{1/3} R^{2/3}}$ implies that $\frac{\sigma B}{\sqrt{KR}} \leq \frac{HB^2}{R}$, and
- $\frac{\zeta_\star^2}{H} \geq \frac{(H\zeta_\star^2 B^4)^{1/3}}{R^{2/3}}$ implies that $\frac{\zeta_\star^2}{H} \leq \frac{HB^2}{R}$.

These observations allow us to avoid the minimum operations, thus concluding the proof. ■

### B.5. Distributed Zero-respecting Algorithms

Intermittent communication is motivated by the sizeable gap between the wall-clock time $\mathcal{C}$ required for a single synchronous communication and the time required per unit of computation $\mathcal{T}$, say a single oracle call (McMahan et al., 2016a; Kairouz et al., 2019). For an efficient implementation, typically, we want our local computation budget $K$ to be comparable to $\mathcal{C}/\mathcal{T}$, i.e., we want to increase our computation load per communication to match the time required for a single communication round. We consider a generalization of zero respecting algorithms (Carmon et al., 2020) denoted by $\mathcal{A}_{ZR}$ in the intermittent communication (IC) setting defined as follows.

**Definition 23 (Distributed Zero-respecting Algorithms)** *Consider $M$ machines in the IC setting, each endowed with an oracle $\mathcal{O}_m : \mathcal{I}\times\mathcal{Z} \to \mathcal{V}$ and a distribution $\mathcal{D}_m$ on $\mathcal{Z}$. Let $I_{r,k}^m$ denote the input to the $k^{th}$ oracle call, leading up to the $r^{th}$ communication round on machine $m$. An optimization algorithm initialized at $0$ is distributed zero-respecting if:*

1. *for all $r \in [R], k \in [K], m \in [M]$, $I_{r,k}^m$ is in*

$$\left\{\bigcup_{l\in[k-1]} supp\left(\mathcal{O}_{F_m}(I_{r,l}^m; z_{r,l}^m \sim \mathcal{D}_m)\right)\right\} \cup \left\{\bigcup_{n\in[M], s\in[r-1], l\in[K]} supp\left(\mathcal{O}_{F_n}(I_{s,l}^n; z_{s,l}^n \sim \mathcal{D}_n)\right)\right\},$$

2. *for all $r \in [R], k \in [K], m \in [M]$, $I_{r,k}^m$ is a deterministic function (which is same across all the machines) of*

$$\left\{\bigcup_{l\in[k-1]} \mathcal{O}_{F_m}(I_{r,l}^m; z_{r,l}^m \sim \mathcal{D}_m)\right\} \cup \left\{\bigcup_{n\in[M], s\in[r-1], l\in[K]} \mathcal{O}_{F_n}(I_{s,l}^n; z_{s,l}^n \sim \mathcal{D}_n)\right\},$$

3. *at the $r^{th}$ communication round, the machines only communicate vectors in*

$$\left\{\bigcup_{n\in[M], s\in[r], l\in[K]} supp\left(\mathcal{O}_{F_n}(I_{s,l}^n; z_{s,l}^n \sim \mathcal{D}_n)\right)\right\},$$

*where $supp\,(.)$ denotes the co-ordinate support/span of its arguments. We denote this class of algorithms by $\boldsymbol{\mathcal{A}_{ZR}}$. Furthermore, if all the oracle inputs are the same between two communication rounds, i.e., $I_{r,k}^m = I_r \in \mathcal{I}$ for all $m \in [M], k \in [K], r \in [R]$, then we say that the algorithm is centralized, and denote this class of algorithms by $\boldsymbol{\mathcal{A}_{ZR}^{cent}} \subset \mathcal{A}_{ZR}$.*

This class captures a very wide variety of distributed optimization algorithms, including mini-batch SGD (Dekel et al., 2012), accelerated mini-batch SGD (Ghadimi and Lan, 2012), local SGD (McMahan et al., 2016a), as well as all the variance-reduced algorithms (Karimireddy et al., 2020a; Zhao et al., 2021; Khanduri et al., 2021). Non-distributed zero-respecting algorithms 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". We have also defined the smaller class of centralized algorithms $\mathcal{A}_{ZR}^{cent}$. These algorithms query the oracles at the same point within each communication round and use the combined $MK$ oracle queries each round to get a *"mini-batch"* estimate of the gradient. Thus, the class $\mathcal{A}_{ZR}^{cent}$ includes algorithms such as mini-batch SGD (Dekel et al., 2012; Woodworth et al., 2020b), but doesn't include local-update algorithms in $\mathcal{A}_{ZR}$ such as local-SGD.

### B.6. Proof of Theorem 11

For even $m$, let

$$F_m(x) := \frac{H}{2}\left((q^2+1)(q-x_1)^2 + \sum_{i=1}^{\lfloor (d-1)/2 \rfloor} (qx_{2i} - x_{2i+1})^2\right), \tag{18}$$

and for odd $m$, let

$$F_m(x) = \frac{H}{2}\left(\sum_{i=1}^{\lfloor d/2 \rfloor} (qx_{2i-1} - x_{2i})^2\right). \tag{19}$$

Thus we have

$$F(x) = \mathbb{E}_m[F_m(x)] = \frac{H}{2}\left((q^2+1)(q-x_1)^2 + \sum_{i=1}^{\boldsymbol{d}} (qx_i - x_{i+1})^2\right). \tag{20}$$

Observe that the optimum of $F$ is attained at $x^\star$, where $x_i^\star = q^i$. Theorem 11 improves on the previous best lower bounds by introducing the term $\frac{HB^2}{R^2}$. Combining the following lemma with standard arguments to achieve the $\frac{\sigma B}{\sqrt{MKR}}$ suffices to prove Theorem 11.

**Lemma 24** *For any $K \geq 2, R, M, H, B, \sigma$, there exist $f(x; xi)$ and distributions $\{\mathcal{D}_m\}$, each satisfying assumptions 2, 4, 3, and together satisfying $\frac{1}{M}\sum_{m=1}^{M} \|\nabla F_m(x^\star)\|_2^2 = 0$, such that for initialization $x^{(0,0)} = 0$, the final iterate $\hat{x}$ of any zero-respecting with $R$ rounds of communication and $KR$ gradient computations per machine satisfies*

$$\mathbb{E}\left[F(\hat{x})\right] - F(x^\star) \succeq \frac{HB^2}{R^2}. \tag{21}$$

**Proof** Consider the division of functions onto machines described above for some sufficiently large $d$.

Let $q = 1 - \frac{1}{R}$, and let $t = \frac{1}{2}\log_q\left(\frac{B^2}{R}\right)$. We begin at the iterate $x_0$, where the coordinate $(x_0)_i = q^i$ for all $i < t$, and $(x_0)_i = 0$ for $i \geq t$. Observe that $\|x_0 - x^\star\|^2 \leq \sum_{i=t}^{\infty} q^{2i} \leq \frac{q^{2t}}{1-q^2} \leq Rq^{2t} \leq B^2$.

Observe that for any zero-respecting algorithm, on odd machines, if for any $i$, we have $x_{2i}^m = x_{2i+1}^m = 0$, then after any number of local computations, we still have $x_{2i+1} = 0$. Similarly, on even machines, if for any $i$, we have $x_{2i-1}^m = x_{2i}^m = 0$, then after any number of local computations, we still have $x_{2i} = 0$.

Thus, after $R$ rounds of communication, on all machines, we have $x_i^m = 0$, for all $i > t + R$. Thus for $d$ sufficiently large, we have $\|\hat{x} - x^\star\|^2 \geq \sum_{i=t+R+1}^{d} q^{2i} \geq \frac{q^{2t+2R+2} - q^{2d}}{1-q^2} = \Omega\left(B^2 q^{2R+2}\right) = \Omega(B^2)$ since $q = 1 - \frac{1}{R}$.

Now observe that the Hessian of $F$ is a tridiagonal Toeplitz matrix with diagonal entries $H(q^2 + 1)$ and off-diagonal entries $-Hq$. It is well-known (see e.g., Golub (2005)) that the $d$ eigenvalues of $\tilde{M}$ are $(1+q^2)H + 2qH\cos\left(\frac{i\pi}{d+1}\right)$ for $i = 1, \ldots, d$. Thus since $\cos(x) \geq -1$, we know that $F$ has strong-convexity parameter at least $H(q^2 + 1 - 2q) = \Omega\left(\frac{H}{R^2}\right)$, so we have $F(\hat{x}) - F(x^*) \geq \Omega\left(B^2\right)\Omega\left(\frac{H}{R^2}\right)$, which gives the desired result. ∎

## Appendix C. On the Assumption of Wang et al. (2022)

We will now consider the first-order assumption introduced by Wang et al. (2022). As discussed in Section 2, Assumption 2 is very restrictive when $\mathcal{X} = \mathbb{R}^d$ . Wang et al. (2022) claim that even with $\mathcal{X} = S^\star$, the optima of $F(x)$, Assumption (2) (and by implication Assumption 5) can be restrictive in some settings. They propose an alternative assumption that instead tries to capture how much the local iterates move when initialized at an optimum of the averaged function $F$.

**Assumption 6 (Movement at Optima)** *Given an inner step-size $\eta$, local steps $K$, and $x^\star \in \arg\min_{x\in\mathbb{R}^d} F(x)$ assume,*

$$\frac{1}{M\eta K}\left\|\sum_{m\in[M]} x^\star - \hat{x}_K^m\right\|_2 \leq \rho,$$

*where $\hat{x}_K^m$ is the iterate on machine $m$ after making $K$ local steps (using exact gradients) initialized at $x^\star$.*

Unlike all other assumptions, note that $\rho$ in Assumption 6 can be a function of the hyper-parameters $\eta$ and $K$ despite normalizing with $\eta K$. Wang et al. (2022) argue that Assumption 6 is much less restrictive than the other first-order assumptions (Assumption 2 with $\mathcal{X} = \mathbb{R}^d$ or with $\mathcal{X} = S^\star$). And, when the client's objectives are strongly convex, they show a provable domination over large-mini-batch SGD in a regime of low heterogeneity (see the discussion in Section 3). However, it is unclear if the assumption is useful in the general convex setting, where we often empirically see that local SGD outperforms mini-batch SGD. We show in Section 3 that this assumption is not useful in the convex setting because of the following connection with Assumption 2 for $\mathcal{X} = S^\star$.

**Proposition 25** *If the functions of the machines $\{F_m\}_{m\in[M]}$ satisfy Assumption 2 for $\mathcal{X} = S^\star$ with $\zeta_\star$, then we have,*

$$\left\|\frac{1}{M\eta K}\sum_{m\in[M]} x^\star - \hat{x}_K^m\right\|_2 \leq \zeta_\star\left((1+\eta H)^{K-1} - 1\right).$$

**Proof** Define $\phi(k) := \left\|\frac{1}{M\eta K}\sum_{m\in[M]} x^\star - \hat{x}_k^m\right\|_2$ where $\hat{x}_k^m$ is the $k$th gradient descent iterate on machine $m$ initialized at $\hat{x}_0^m = x^\star$. Then note the following,

$$\begin{aligned}
\phi(K) &= \left\|\frac{1}{M\eta K}\sum_{m\in[M]} x^\star - \hat{x}_K^m\right\|_2, \\
&= \left\|\frac{1}{M\eta K}\sum_{m\in[M]} x^\star - \hat{x}_{K-1}^m + \eta\nabla F_m(\hat{x}_{K-1}^m)\right\|_2, \\
&\leq \left\|\frac{1}{M\eta K}\sum_{m\in[M]} x^\star - \hat{x}_{K-1}^m\right\|_2 + \left\|\frac{1}{MK}\sum_{m\in[M]} \nabla F_m(\hat{x}_{K-1}^m)\right\|_2, \\
&= \phi(K-1) + \left\|\frac{1}{MK}\sum_{m\in[M]} \nabla F_m(\hat{x}_{K-1}^m) - \nabla F_m(x^\star)\right\|_2,
\end{aligned}$$

$$
\begin{aligned}
&\leq \phi(K-1) + \frac{1}{MK}\sum_{m\in[M]} \left\|\nabla F_m(\hat{x}_{K-1}^m) - \nabla F_m(x^\star)\right\|_2, \\
&\leq \phi(K-1) + \frac{H}{MK}\sum_{m\in[M]} \left\|\hat{x}_{K-1}^m - x^\star\right\|_2, \\
&= \phi(K-1) + \frac{H}{K}\delta(K-1), \qquad (22)
\end{aligned}
$$

where we define $\delta(k) := \frac{1}{M}\sum_{m\in[M]} \|\hat{x}_k^m - x^\star\|_2$. Now we consider another recursion on $\delta(k)$ to introduce the $\zeta_\star$ assumption:

$$
\begin{aligned}
\delta(k) &= \frac{1}{M}\sum_{m\in[M]} \|\hat{x}_k^m - x^\star\|_2, \\
&\leq \frac{1}{M}\sum_{m\in[M]} \left\|\hat{x}_{k-1}^m - x^\star\right\|_2 + \frac{\eta}{M}\sum_{m\in[M]} \left\|\nabla F_m(\hat{x}_m^{k-1})\right\|_2, \\
&\leq \frac{1}{M}\sum_{m\in[M]} \left\|\hat{x}_{k-1}^m - x^\star\right\|_2 + \frac{\eta}{M}\sum_{m\in[M]} \left(\left\|\nabla F_m(\hat{x}_m^{k-1}) - \nabla F_m(x^\star)\right\|_2 + \|\nabla F_m(x^\star)\|_2\right), \\
&\leq \frac{(1+\eta H)}{M}\sum_{m\in[M]} \left\|\hat{x}_{k-1}^m - x^\star\right\|_2 + \eta\zeta_\star, \\
&\leq (1+\eta H)\delta(k-1) + \eta\zeta_\star, \\
&\leq \frac{\zeta_\star}{H}\left((1+\eta H)^k - 1\right). \qquad (\delta(0)=0) \qquad (23)
\end{aligned}
$$

Plugging (23) back into 22 we get,

$$
\begin{aligned}
\phi(K) &\leq \phi(K-1) + \frac{\zeta_\star}{K}\left((1+\eta H)^{K-1} - 1\right), \\
&= \zeta_\star\left((1+\eta H)^{K-1} - 1\right),
\end{aligned}
$$

which proves the claim. ■

**Remark 26** *When Assumption 2 for $\mathcal{X} = S^\star$ can be satisfied with $\zeta_\star = 0$, we can satisfy Assumption 6 with $\rho = 0$.*

As might be evident already, Proposition 7 has implications for the heterogeneity assumption by Wang et al. (2022). This is because the construction used in the proposition satisfies Assumption 6 with $\rho = 0$. In particular, in the general convex setting under Assumption 6, local SGD can not hope to benefit in the optimization term from a large $K$. This implies that it is **not possible** to explain the effectiveness of local SGD by controlling the movement of the local iterates at the optima, as postulated by Wang et al. (2022). Note that Theorem 11 also applies to Assumption 6. Thus, large mini-batch SGD is also min-max optimal under Assumptions 2, 4, and 6 for convex objectives.

To reconcile this negative result with the upper bound of Wang et al. (2022), we note that their analyses crucially rely on the strong convexity of each machine's objective and not just the strong convexity of the average function, something which is lacking in our hard instance in Proposition 7. This highlights that controlling the client drift for local SGD between communication rounds in the general convex setting is much more challenging due to potential directions of no information on each machine (both $A_1, A_2$ are rank one in Proposition 7).

### C.1. Beating Mini-batch SGD

Finally, we note that for quadratic functions of the form discussed in Section 5, we can bound the average movement at optimum as follows,

$$\begin{aligned}
\rho &= \frac{1}{M\eta K}\left\|\sum_{m\in[M]} x^\star - x_K^m\right\|_2, \\
&= \frac{1}{M\eta K}\left\|\sum_{m\in[M]} x^\star - x_m^\star - (I-\eta A_m)^K(x_\star - x_m^\star)\right\|_2, \\
&= \frac{1}{M\eta K}\left\|\sum_{m\in[M]} \left(I-(I-\eta A_m)^K\right)(x_\star - x_m^\star)\right\|_2, \\
&\leq \frac{1}{M\eta K}\sum_{m\in[M]} \left\|\left(I-(I-\eta A_m)^K\right)\right\|_2 \left\|x_\star - x_m^\star\right\|_2, \\
&\leq \frac{1-(1-\eta L)^K}{\eta K}\zeta_\star.
\end{aligned}$$

If $\eta = \frac{1}{2HK^p}$ for $p<1$, and $K\to\infty$ we get that,

$$\begin{aligned}
\lim_{K\to\infty}\rho &\leq \lim_{K\to\infty}\frac{1-(1-\eta H)^K}{\eta K}\zeta_\star, \\
&\leq \lim_{K\to\infty}\frac{1-\left(1-\frac{1}{2K^p}\right)^K}{K^{1-p}}\cdot 2L\zeta_\star, \\
&= 0.
\end{aligned}$$

On the other hand, if $\eta = \frac{1}{2HK}$, then we get that,

$$\begin{aligned}
\lim_{K\to\infty}\rho &\leq \lim_{K\to\infty}\frac{1-(1-\eta H)^K}{\eta K}\zeta_\star, \\
&\leq \lim_{K\to\infty}\left(1-\left(1-\frac{1}{2K}\right)^K\right)\cdot 2L\zeta_\star, \\
&= \left(1-\frac{1}{\sqrt{e}}\right)\cdot 2L\zeta_\star.
\end{aligned}$$

This suggests that the Corollary 2 in Wang et al. (2022) must use $\eta = \Omega\left(\frac{1}{K}\right)$ to ensure that $\rho$ can reach arbitrarily small values. Unfortunately the corollary can only be obtained by setting

$\eta < \min\left\{\frac{1}{\mu K}, \frac{1}{L}\right\}$. Thus, in their existing result, it seems **impossible** to improve the convergence w.r.t. MB-SGD unconditionally. This also highlights that our result in Theorem 22 is much cleaner, depends on only problem-dependentth parameters, and can improve over mini-batch SGD.

## Appendix D. Missing Proofs from Section 4

### D.1. Proof of Poposition 13

**Proof** Denote the function $G_m := F_m - F$ for all $m \in [M]$. Then note that using Taylor expansion, we can write for any $x \in \mathbb{B}_2(D)$ and $x^\star \in S$,

$$\nabla G_m(x) - \nabla G_m(x^\star) = \left[\int_0^1 \nabla^2 G_m(x^\star + s(x - x^\star))ds\right](x - x^\star).$$

Re-arranging, taking norm, and applying triangle inequality implies,

$$\begin{aligned}
\|\nabla G_m(x)\|_2 &= \left\|\nabla G_m(x^\star) + \left[\int_0^1 \nabla^2 G_m(x^\star + s(x - x^\star))ds\right](x - x^\star)\right\|_2, \\
&\leq \|\nabla F_m(x^\star)\|_2 + \left\|\int_0^1 \nabla^2 G_m(x^\star + s(x - x^\star))ds\right\|_2 \|x - x^\star\|_2, \\
&\leq \|\nabla F_m(x^\star)\|_2 + \left[\int_0^1 \left\|\nabla^2 G_m(x^\star + s(x - x^\star))\right\|_2 ds\right] \|x - x^\star\|_2, \\
&\leq \|\nabla F_m(x^\star)\|_2 + \left[\int_0^1 \tau ds\right] D, \\
&\leq \|\nabla F_m(x^\star)\|_2 + \tau D.
\end{aligned}$$

Averaging over the machines and noting the definition of $\zeta_\star$ gives us the desired result. ■

### D.2. Convergence in the Strongly Convex Setting

Recall that the average objective $F$ is $\mu$-strongly convex, if for all $x, y \in \mathbb{R}^d$,

$$F(y) + \langle \nabla F(y), x - y\rangle + \frac{\mu}{2}\|x - y\|_2^2 \leq F(x).$$

We can give the following upper bound for local SGD for optimizing problems in the class $\mathcal{P}_\tau^{H,Q,B,\sigma}$ with $\mu$-strongly convex average objective.

**Theorem 27 (Informal, Strongly Convex)** *For any $K, R, M \geq 1$ and $H, B, Q, \sigma, \tau, \geq 0,\ \mu > 0$ consider a problem instance in the class $\mathcal{P}_\tau^{H,Q,B,\sigma}$, with $\mu$-strongly convex average objective $F$. Then there exists a sequence of weights $w_{r,k} \geq 0$, for $r \in [R], k \in [0, K-1]$ and a fixed inner stepsize of $\eta \leq \frac{1}{2H}$ such that with outer stepsize of $\beta = 1$ we have the following convergence rate (up to logarithmic factors) for local SGD where $W_{R,K} = \sum_{r\in[R],k\in[0,K-1]} w_{r,k}$:*

$$\mathbb{E}\left[F\left(\frac{1}{MW_{R,K}}\sum_{m=1}^{M}\sum_{r=1}^{R}\sum_{k=0}^{K-1} w_{r,k}x_{r,k}^m\right) - F(x^\star) + \mu\left\|\frac{1}{M}\sum_{m\in[M]} x_{R,K}^m - x^\star\right\|_2^2\right] \preceq$$

$$HB^2 \exp\left(-\frac{\mu KR}{H}\right) + \frac{\sigma^2}{\mu MKR} + \frac{\tau^2\sigma^2}{\mu^3 KR^2} + \frac{\tau^2\zeta^2}{\mu^3 R^2} + \frac{Q^2\sigma^4}{\mu^5 K^2 R^4} + \frac{Q^2\zeta^4}{\mu^5 R^4}\,.$$

*Proof of this Theorem can be found in D.4.*

### D.3. Useful Lemmas

In this section, we provide a list of useful lemmas which will be used in the proofs. We use the notation $\mathbb{E}_t$ for expectation conditioned on $x_t^1, ..., x_t^M$ and $\mathbb{E}$ for the unconditional expectation.

**Lemma 28** *For a convex function $F$ we have:*

$$F\left(\frac{1}{M}\sum_{m=1}^{M} x_m\right) \leq \frac{1}{M}\sum_{m=1}^{M} F(x_m)\,. \tag{24}$$

**Lemma 29** *For a set of $M$ vectors $a_1, a_2, ..., a_M \in \mathbb{R}^d$ we have:*

$$\left\|\sum_{m=1}^{M} a_m\right\|_2 \leq \sum_{m=1}^{M} \|a_m\|_2\,. \tag{25}$$

**Lemma 30** *For a set of $M$ vectors $a_1, a_2, ..., a_M \in \mathbb{R}^d$ we have:*

$$\left\|\sum_{m=1}^{M} a_m\right\|_2^2 \leq M\sum_{m=1}^{M} \|a_m\|_2^2\,. \tag{26}$$

**Lemma 31** *For two arbitrary vectors $a, b \in \mathbb{R}^d$ and $\forall \gamma > 0$ we have:*

$$\|a+b\|_2^2 \leq (1+\gamma)\|a\|_2^2 + (1+\gamma^{-1})\|b\|_2^2\,. \tag{27}$$

**Lemma 32** *Let Assumption 3 hold. Then we have:*

$$\mathbb{E}_t\left\|\frac{1}{M}\sum_{m=1}^{M} g_t^m - \frac{1}{M}\sum_{m=1}^{M}\nabla F_m(x_t^m)\right\|_2^2 \leq \frac{\sigma^2}{M}\,. \tag{28}$$

**Lemma 33** *Let $F$ be a convex and $H$-smooth function. Then for any $x, y \in \mathbb{R}^d$ we have:*

$$\frac{1}{2H}\|\nabla F(x) - \nabla F(y)\|_2^2 \leq F(y) - F(x) + \langle \nabla F(x), x - y\rangle\,. \tag{29}$$

**Lemma 34** *Let Assumption 3 hold and $F(x) = \frac{1}{M}\sum_{m=1}^{M} F_m(x)$. Then for any $x, y \in \mathbb{R}^d$ we have the following inequality:*

$$\|\nabla F_m(x) - \nabla F(x) + \nabla F(y) - \nabla F_m(y)\|_2^2 \leq \tau^2\|x-y\|_2^2\,. \tag{30}$$

**Proof** We define the function $\Psi(z) = F_m(z) - F(z)$ so $\nabla^2\Psi(z) = \nabla^2 F_m(z) - \nabla^2 F(z)$. By Assumption 3 we know that $\forall z \in \mathbb{R}^d, \left\|\nabla^2\Psi(z)\right\|_2 \leq \tau$ which means $\Psi(z)$ is $\tau$-smooth. From the smoothness property we know:

$$\|\nabla\Psi(x) - \nabla\Psi(y)\|_2 \leq \tau \|x - y\|_2 .$$

By replacing the definition of $\Psi$ we get:

$$\|\nabla F_m(x) - \nabla F(x) - \nabla F_m(y) + \nabla F(y)\|_2 \leq \tau \|x - y\|_2 .$$

■

**Lemma 35 (Kovalev et al. 2019, Definition 2)** *Let function $F$ satisfy Assumption 4, then for any $x, y \in \mathbb{R}^d$, we have the following inequality:*

$$\left\|\nabla F(x) - \nabla F(y) - \nabla^2 F(y)(x-y)\right\|_2 \leq \frac{Q}{2}\|x - y\|_2^2 . \tag{31}$$

**Lemma 36** *Let Assumptions 3 and 2 hold. We can upper bound the consensus error $\Xi_t$ using the Lemma from Woodworth et al. (2020b). With a fixed learning rate of $\eta_t = \eta$ we have:*

$$\Xi_t := \frac{1}{M}\sum_{m=1}^{M}\|x_t^m - \bar{x}_t\|_2^2 \leq 3K\sigma^2\eta^2 + 6K^2\eta^2\zeta^2 , \tag{32}$$

*where $x_t^m$ is the parameters of client $m$ at time step $t$ and $\bar{x}_t = \frac{1}{M}\sum_{m=1}^{M} x_t^m$.*

**Lemma 37 (Stepsize Tuning)** *Our upper bound has the following recursive form:*

$$r_{t+1} \leq (1 - A\eta_t)r_t - \eta_t e_t + C\eta_t^2 + E\eta_t^3 + G\eta_t^5 , \tag{33}$$

*where $\{r_t\}_{t\geq 0}$ and $\{e_t\}_{t\geq 0}$ are two non-negative sequences and $\eta_t \leq \frac{1}{D}$ for a $D > 0$. With the following choice of parameters we recover our recursive upper bound used in the proof of Theorem 27 (see below):*

$$r_t = \|\bar{x}_t - x^\star\|_2^2 , \quad e_t = \Big[F(\bar{x}_t) - F(x^\star)\Big], \quad C = \frac{\sigma^2}{M}, \quad D = 2H, \quad A = \frac{\mu}{2} ,$$
$$E = \frac{24K\tau^2\sigma^2}{\mu} + \frac{48K^2\tau^2\zeta^2}{\mu}, \quad G = \frac{36K^2Q^2\sigma^4}{\mu} + \frac{144Q^2K^4\zeta^4}{\mu} .$$

The recursion (33) can be solved as follows:

$$\frac{1}{W_T}\sum_{t=0}^{T} w_t e_t + A r_{T+1} \leq D r_0 \exp\left(\frac{-AT}{D}\right) + \tilde{\mathcal{O}}\left(\frac{C}{AT}\right) + \tilde{\mathcal{O}}\left(\frac{E}{A^2T^2}\right) + \tilde{\mathcal{O}}\left(\frac{G}{A^4T^4}\right) ,$$

where we define the coefficient $w_t = (1 - A\eta)^{-(t+1)}$ and $W_T = \sum_{t=0}^{T} w_t$.

**Proof** The proof is similar to the works (Stich, 2019; Koloskova et al., 2020). We divide both sides of the recursive inequality by $\eta_t$ and rearrange the terms:

$$e_t \leq \frac{(1 - A\eta_t)}{\eta_t} r_t - \frac{1}{\eta_t} r_{t+1} + C\eta_t + E\eta_t^2 + G\eta_t^4$$

Note that learning rate is fixed $\eta_t = \eta$. Then we multiply both sides of the inequality by $w_t$ and sum over $t$ which gives us:

$$\begin{aligned}
\frac{1}{W_T}\sum_{t=0}^{T} w_t e_t &\le \frac{1}{W_T}\sum_{t=0}^{T}\left[\frac{(1-A\eta)^{-(t+1)}(1-A\eta)}{\eta}r_t - \frac{(1-A\eta)^{-(t+1)}}{\eta}r_{t+1}\right] + C\eta + E\eta^2 + G\eta^4 \\
&= \frac{1}{W_T}\sum_{t=0}^{T}\left[\frac{(1-A\eta)^{-t}}{\eta}r_t - \frac{(1-A\eta)^{-(t+1)}}{\eta}r_{t+1}\right] + C\eta + E\eta^2 + G\eta^4 \\
&= \frac{1}{W_T}\sum_{t=0}^{T}\left[\frac{w_{t-1}}{\eta}r_t - \frac{w_t}{\eta}r_{t+1}\right] + C\eta + E\eta^2 + G\eta^4 \\
&= \left[\frac{r_0}{\eta W_T} - \frac{w_T}{\eta W_T}r_{T+1}\right] + C\eta + E\eta^2 + G\eta^4\,.
\end{aligned}$$

So we have:

$$\frac{1}{W_T}\sum_{t=0}^{T} w_t e_t + \frac{w_T}{\eta W_T}r_{T+1} \le \frac{r_0}{\eta W_T} + C\eta + E\eta^2 + G\eta^4\,.$$

Also using the fact that $W_T \le \frac{w_T}{a\eta}$ and $W_T \ge (1-A\eta)^{-(T+1)}$ we have:

$$\begin{aligned}
\frac{1}{W_T}\sum_{t=0}^{T} w_t e_t + a r_{T+1} &\le \frac{(1-A\eta)^{-(T+1)}r_0}{\eta} + C\eta + E\eta^2 + G\eta^4 \\
&\le \frac{r_0\exp\left(-a\eta(T+1)\right)}{\eta} + C\eta + E\eta^2 + G\eta^4\,.
\end{aligned}$$

Now we consider two cases:

1) $\frac{1}{D} \ge \frac{\ln\left(\max\left\{2, \frac{A^2T^2r_0}{C}\right\}\right)}{AT}$: We choose $\eta = \frac{\ln\left(\max\left\{2, \frac{A^2T^2r_0}{C}\right\}\right)}{AT}$. With this choice we get a rate of:

$$\begin{aligned}
\frac{1}{W_T}\sum_{t=0}^{T} w_t e_t + A r_{T+1} &\le \frac{r_0 AT}{\ln\left(\max\left\{2, \frac{A^2T^2r_0}{C}\right\}\right)}\exp\left(-\ln\left(\max\left\{2, \frac{A^2T^2r_0}{C}\right\}\right)\right) + \tilde{\mathcal{O}}\left(\frac{C}{AT}\right) \\
&\quad + \tilde{\mathcal{O}}\left(\frac{E}{A^2T^2}\right) + \tilde{\mathcal{O}}\left(\frac{G}{A^4T^4}\right) \\
&\le \tilde{\mathcal{O}}\left(\frac{r_0 A}{T^2}\right) + \tilde{\mathcal{O}}\left(\frac{C}{AT}\right) + \tilde{\mathcal{O}}\left(\frac{E}{A^2T^2}\right) + \tilde{\mathcal{O}}\left(\frac{G}{A^4T^4}\right)\,.
\end{aligned}$$

2) $\frac{1}{D} \le \frac{\ln\left(\max\left\{2, \frac{A^2T^2r_0}{C}\right\}\right)}{AT}$: We choose $\eta = \frac{1}{D}$. With this choice we get a rate of:

$$\begin{aligned}
\frac{1}{W_T}\sum_{t=0}^{T} w_t e_t + A r_{T+1} &\le D r_0 \exp\left(\frac{-AT}{D}\right) + \frac{C}{D} + \frac{E}{D^2} + \frac{G}{D^4} \\
&\le D r_0 \exp\left(\frac{-AT}{D}\right) + \tilde{\mathcal{O}}\left(\frac{C}{AT}\right) + \tilde{\mathcal{O}}\left(\frac{E}{A^2T^2}\right) + \tilde{\mathcal{O}}\left(\frac{G}{A^4T^4}\right)\,.
\end{aligned}$$

■

### D.4. Proof of Theorem 27

**Proof** The proof follows a similar approach to Yuan and Ma (2020) while we assume that we have heterogeneity. In the following proof, we use the same notation as the work Woodworth et al. (2020b). We drop the subscript $r$ for simplicity and use $t$ instead which is in the range $[0, KR-1]$. We use the notation $x_t^m$ as the parameters of client $m$ at time step $t$ and $g_t^m$ as the stochastic gradient on client $m$ at time step $t$. Also because $\beta = 1$, the update rule $x_{r+1} = x_r + \frac{\beta}{M}\sum_m (x_{r,K}^m - x_r)$ reduces to $x_{r+1} = \frac{1}{M}\sum_m x_{r,K}^m$. We also define $\bar{x}_t = \frac{1}{M}\sum_{m=1}^M x_t^m$ as the average of parameters over all clients at time step $t$. Starting with the distance from the optimal point and taking the conditional expectation on the previous iterate $x_t^m, \forall m \in [M]$ we have:

$$
\begin{aligned}
&\mathbb{E}_t \left\|\bar{x}_{t+1} - x^\star\right\|_2^2 \\
&= \mathbb{E}_t \left\| \bar{x}_t - x^\star - \frac{\eta_t}{M}\sum_{m=1}^M \nabla F_m(x_t^m) + \frac{\eta_t}{M}\sum_{m=1}^M \nabla F_m(x_t^m) - \frac{\eta_t}{M}\sum_{m=1}^M g_t^m \right\|_2^2 \\
&\overset{\text{(Lemma 32)}}{\leq} \left\| \bar{x}_t - x^\star - \frac{\eta_t}{M}\sum_{m=1}^M \nabla F_m(x_t^m)\right\|_2^2 + \frac{\eta_t^2\sigma^2}{M} \\
&= \left\| \bar{x}_t - x^\star - \eta_t \nabla F(\bar{x}_t) + \eta_t \nabla F(\bar{x}_t) - \frac{\eta_t}{M}\sum_{m=1}^M \nabla F_m(x_t^m)\right\|_2^2 + \frac{\eta_t^2\sigma^2}{M} \\
&\overset{\text{(Lemma 31)}}{\leq} \left(1 + \frac{\eta_t\mu}{2}\right)\left\|\bar{x}_t - x^\star - \eta_t\nabla F(\bar{x}_t)\right\|_2^2 + \eta_t^2\left(1 + \frac{2}{\eta_t\mu}\right)\left\|\nabla F(\bar{x}_t) - \frac{1}{M}\sum_{m=1}^M \nabla F_m(x_t^m)\right\|_2^2 + \frac{\eta_t^2\sigma^2}{M} .
\end{aligned} \tag{34}
$$

For the first term in (34) we have:

$$
\left\|\bar{x}_t - x^\star - \eta_t\nabla F(\bar{x}_t)\right\|_2^2 = \left\|\bar{x}_t - x^\star\right\|_2^2 + \eta_t^2\left\|\nabla F(\bar{x}_t)\right\|_2^2 - 2\eta_t\left\langle \bar{x}_t - x^\star, \nabla F(\bar{x}_t)\right\rangle .
$$

For the second term in the above equation we have:

$$
\begin{aligned}
\eta_t^2\left\|\nabla F(\bar{x}_t)\right\|_2^2 &= \eta_t^2\left\|\nabla F(\bar{x}_t) - \nabla F(x^\star)\right\|_2^2 \\
&\overset{\text{(Lemma 33)}}{\leq} 2H\eta_t^2\Big[F(\bar{x}_t) - F(x^\star)\Big] .
\end{aligned}
$$

For the third term in the equality we have:

$$
-2\eta_t\left\langle \bar{x}_t - x^\star, \nabla F(\bar{x}_t)\right\rangle \leq -2\eta_t\Big[F(\bar{x}_t) - F(x^\star)\Big] - \eta_t\mu\left\|\bar{x}_t - x^\star\right\|_2^2
$$

Now by putting everything together we have:

$$
\begin{aligned}
&\left\|\bar{x}_t - x^\star - \eta_t\nabla F(\bar{x}_t)\right\|_2^2 \\
&= \left\|\bar{x}_t - x^\star\right\|_2^2 + \eta_t^2\left\|\nabla F(\bar{x}_t)\right\|_2^2 - 2\eta_t\left\langle \bar{x}_t - x^\star, \nabla F(\bar{x}_t)\right\rangle \\
&\leq \left\|\bar{x}_t - x^\star\right\|_2^2 + 2H\eta_t^2\Big[F(\bar{x}_t) - F(x^\star)\Big] - 2\eta_t\Big[F(\bar{x}_t) - F(x^\star)\Big] - \eta_t\mu\left\|\bar{x}_t - x^\star\right\|_2^2 .
\end{aligned}
$$

With the choice of $\eta_t \leq \frac{1}{2H}$ we have:

$$\|\bar{x}_t - x^\star - \eta_t \nabla F(\bar{x}_t)\|_2^2 \leq (1 - \eta_t \mu) \|\bar{x}_t - x^\star\|_2^2 - \eta_t \Big[F(\bar{x}_t) - F(x^\star)\Big].$$

Then we multiply both sides by the coefficient $(1 + \frac{\eta_t \mu}{2})$ and we have:

$$\begin{aligned}\left(1 + \frac{\eta_t \mu}{2}\right) \|\bar{x}_t - x^\star - \eta_t \nabla F(\bar{x}_t)\|_2^2 &\leq \left(1 + \frac{\eta_t \mu}{2}\right) (1 - \eta_t \mu) \|\bar{x}_t - x^\star\|_2^2 - \eta_t \left(1 + \frac{\eta_t \mu}{2}\right) \Big[F(\bar{x}_t) - F(x^\star)\Big] \\ &\leq \left(1 - \frac{\eta_t \mu}{2}\right) \|\bar{x}_t - x^\star\|_2^2 - \eta_t \Big[F(\bar{x}_t) - F(x^\star)\Big]\end{aligned}$$

For the second term in (34) we have:

$$\begin{aligned}&\eta_t^2 \left(1 + \frac{2}{\eta_t \mu}\right) \left\|\frac{1}{M} \sum_{m=1}^M \nabla F_m(x_t^m) - \nabla F(\bar{x}_t)\right\|_2^2 \\ &\leq \frac{4\eta_t}{\mu} \left\|\frac{1}{M} \sum_{m=1}^M \nabla F_m(x_t^m) - \nabla F(\bar{x}_t)\right\|_2^2 \\ &= \frac{4\eta_t}{\mu} \left\|\frac{1}{M} \sum_{m=1}^M \Big(\nabla F_m(x_t^m) - \nabla F(x_t^m) + \nabla F(\bar{x}_t) - \nabla F_m(\bar{x}_t)\Big) + \frac{1}{M} \sum_{m=1}^M \nabla F(x_t^m) - \nabla F(\bar{x}_t)\right\|_2^2 \\ &\overset{\text{(Lemma 28,30)}}{\leq} \frac{8\eta_t}{\mu M} \sum_{m=1}^M \|\nabla F_m(x_t^m) - \nabla F(x_t^m) + \nabla F(\bar{x}_t) - \nabla F_m(\bar{x}_t)\|_2^2 + \frac{8\eta_t}{\mu} \left\|\frac{1}{M} \sum_{m=1}^M \nabla F(x_t^m) - \nabla F(\bar{x}_t)\right\|_2^2 \\ &\overset{\text{(Lemma 34)}}{\leq} \frac{8\eta_t}{\mu} \tau^2 \Xi_t + \frac{8\eta_t}{\mu} \left\|\frac{1}{M} \sum_{m=1}^M \nabla F(x_t^m) - \nabla F(\bar{x}_t)\right\|_2^2.\end{aligned}$$

For the second term in the above inequality we have:

$$\begin{aligned}&\frac{8\eta_t}{\mu} \left\|\frac{1}{M} \sum_{m=1}^M \nabla F(x_t^m) - \nabla F(\bar{x}_t)\right\|_2^2 \\ &= \frac{8\eta_t}{\mu} \left\|\frac{1}{M} \sum_{m=1}^M \Big(\nabla F(x_t^m) - \nabla F(\bar{x}_t) - \nabla^2 F(\bar{x}_t)^\top (x_t^m - \bar{x}_t)\Big) + \underbrace{\frac{1}{M} \sum_{m=1}^M \nabla^2 F(\bar{x}_t)^\top (x_t^m - \bar{x}_t)}_{=0}\right\|_2^2 \\ &= \frac{8\eta_t}{\mu} \left(\left\|\frac{1}{M} \sum_{m=1}^M \Big(\nabla F(x_t^m) - \nabla F(\bar{x}_t) - \nabla^2 F(\bar{x}_t)^\top (x_t^m - \bar{x}_t)\Big)\right\|_2\right)^2 \\ &\overset{\text{(Lemma 29)}}{\leq} \frac{8\eta_t}{\mu M^2} \left(\sum_{m=1}^M \left\|\nabla F(x_t^m) - \nabla F(\bar{x}_t) - \nabla^2 F(\bar{x}_t)^\top (x_t^m - \bar{x}_t)\right\|_2\right)^2 \\ &\overset{\text{(Lemma 35)}}{\leq} \frac{8\eta_t}{\mu M^2} \left(\sum_{m=1}^M \frac{Q}{2} \|x_t^m - \bar{x}_t\|_2^2\right)^2\end{aligned}$$

$$
\begin{aligned}
&= \frac{2Q^2\eta_t}{\mu}\left(\frac{1}{M}\sum_{m=1}^{M}\|x_t^m - \bar{x}_t\|_2^2\right)^2 \\
&\le \frac{2Q^2\eta_t}{\mu}\Xi_t^2 .
\end{aligned}
$$

Now by plugging everything back into (34) we have:

$$
\mathbb{E}_t\|\bar{x}_{t+1} - x^\star\|_2^2 \le \left(1 - \frac{\eta_t\mu}{2}\right)\|\bar{x}_t - x^\star\|_2^2 - \eta_t\Big[F(\bar{x}_t) - F(x^\star)\Big] + \frac{8\tau^2\eta_t}{\mu}\Xi_t + \frac{2Q^2\eta_t}{\mu}\Xi_t^2 + \frac{\eta_t^2\sigma^2}{M} .
$$

Then we divide both sides by $\eta_t = \eta$, rearrange the terms and take the unconditional expectation, and we have:

$$
\begin{aligned}
&\mathbb{E}\Big[F(\bar{x}_t) - F(x^\star)\Big] \\
&\le \left(\frac{1}{\eta_t} - \frac{\mu}{2}\right)\mathbb{E}\|\bar{x}_t - x^\star\|_2^2 - \frac{1}{\eta_t}\mathbb{E}\|\bar{x}_{t+1} - x^\star\|_2^2 + \frac{8\tau^2}{\mu}\Xi_t + \frac{2Q^2}{\mu}\Xi_t^2 + \frac{\eta\sigma^2}{M} \\
&\overset{\text{(Lemma 36)}}{\le} \left(\frac{1}{\eta} - \frac{\mu}{2}\right)\mathbb{E}\|\bar{x}_t - x^\star\|_2^2 - \frac{1}{\eta}\mathbb{E}\|\bar{x}_{t+1} - x^\star\|_2^2 + \frac{\eta\sigma^2}{M} + \\
&\qquad\qquad \frac{24K\tau^2\sigma^2\eta^2}{\mu} + \frac{48K^2\tau^2\zeta^2\eta^2}{\mu} + \frac{36K^2Q^2\sigma^4\eta^4}{\mu} + \frac{144Q^2K^4\zeta^4\eta^4}{\mu} .
\end{aligned}
$$

After tuning the stepsize using the Lemma 37 with the weights $w_t = (1 - \frac{\mu\eta}{2})^{-(t+1)}$ and $W_T = \sum_{t=0}^{KR-1} w_t$ and by summing over $t = 0, .., KR-1$ we have:

$$
\begin{aligned}
&\mathbb{E}\left[F\left(\frac{1}{W_T}\sum_{t=0}^{KR-1} w_t\bar{x}_t\right) - F(x^\star)\right] + \frac{\mu}{2}\|\bar{x}_{T+1} - x^\star\|_2^2 \\
&\le 2HB^2\exp\left(-\frac{\mu T}{4H}\right) + \tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu MKR}\right) + \tilde{\mathcal{O}}\left(\frac{\tau^2\sigma^2}{\mu^3KR^2}\right) + \tilde{\mathcal{O}}\left(\frac{\tau^2\zeta^2}{\mu^3R^2}\right) + \tilde{\mathcal{O}}\left(\frac{Q^2\sigma^4}{\mu^5K^2R^4}\right) + \tilde{\mathcal{O}}\left(\frac{Q^2\zeta^4}{\mu^5R^4}\right) .
\end{aligned}
$$

It's worth mentioning that the notation $\tilde{\mathcal{O}}(.)$ means the bound holds up to some logarithmic factors appearing in the nominator for the last five terms. The logarithmic factor is in the form of $\frac{\ln\left(\max\{2,T^2\}\right)}{AT}$ (some absolute constants are ignored). It is also common in the literature to ignore these factors Yuan and Ma (2020); Koloskova et al. (2020). ■

### D.5. Proof of Corollary 15

We use the regularization technique to derive the convergence rate for the convex case from the strongly-convex result. This technique is standard in the literature, see e.g. (Hazan et al., 2016). For the sake of completeness, we repeat the argument here.

Let $F(x)$ be a convex function. We construct a regularized version of this function $F_\mu(x)$ as:

$$
F_\mu(x) = F(x) + \frac{\mu}{2}\|x - x_0\|_2^2 .
$$

Next we define:

$$x_\mu^\star = \arg\min_x F_\mu(x)\,,$$
$$x^\star = \arg\min_x F(x)\,.$$

We have that $F_\mu(x_\mu^\star) \leq F_\mu(x^\star)$. Then we upper bound the function sub-optimality for the convex function $F(x)$:

$$\begin{aligned} F(\bar{x}_t) - F(x^\star) &\leq F_\mu(\bar{x}_t) - \frac{\mu}{2}\left\|\bar{x}_t - x_0\right\|_2^2 - F_\mu(x^\star) + \frac{\mu}{2}\left\|x^\star - x_0\right\|_2^2 \\ &\leq F_\mu(\bar{x}_t) - F_\mu(x^\star) + \frac{\mu}{2}\left\|x^\star - x_0\right\|_2^2 \\ &\leq F_\mu(\bar{x}_t) - F_\mu(x_\mu^\star) + \frac{\mu}{2}\left\|x^\star - x_0\right\|_2^2 \\ &\leq F_\mu(\bar{x}_t) - F_\mu(x_\mu^\star) + \frac{\mu}{2}B^2\,. \end{aligned}$$

The last step is to tune the $\mu$. We assume that we want to achieve $\epsilon$-accuracy when running local SGD on the convex function $F$ so we have:

$$F(\bar{x}_t) - F(x^\star) \leq \big(F_\mu(\bar{x}_t) - F_\mu(x_\mu^\star)\big) + \frac{\mu}{2}B^2 \leq \epsilon\,.$$

To satisfy this condition, it is enough to ensure that both terms in the rate are less than $\frac{\epsilon}{2}$. We solve this for the regularization term and we get:

$$\frac{\mu}{2}B^2 \leq \frac{\epsilon}{2}\,,$$

which results in:

$$\mu \leq \frac{\epsilon}{B^2}\,.$$

Now we proceed by replacing $\mu$ in all terms and deriving the rate.
**Proof** By assuming $\mu = \frac{\epsilon}{B^2}$, we will find the condition on $K$ and $R$ to reach $\epsilon$-accuracy using Theorem 27. For simplicity, we drop all constant numbers which do not affect the rate. To recall, in the strongly convex case we had proven a rate of:

$$HB^2 \exp\left(-\frac{\mu KR}{H}\right) + \tilde{\mathcal{O}}\left(\frac{\sigma^2}{\mu MKR}\right) + \tilde{\mathcal{O}}\left(\frac{\tau^2\sigma^2}{\mu^3 KR^2}\right) + \tilde{\mathcal{O}}\left(\frac{\tau^2\zeta^2}{\mu^3 R^2}\right) + \tilde{\mathcal{O}}\left(\frac{Q^2\sigma^4}{\mu^5 K^2 R^4}\right) + \tilde{\mathcal{O}}\left(\frac{Q^2\zeta^4}{\mu^5 R^4}\right)\,.$$

For the first term we solve the inequality:

$$HB^2 \exp\left(-\frac{\epsilon KR}{HB^2}\right) \leq \epsilon \quad\Longrightarrow\quad KR \geq \frac{HB^2}{\epsilon}\ln\left(\frac{\epsilon}{HB^2}\right) = \tilde{\mathcal{O}}\left(\frac{HB^2}{\epsilon}\right)\,.$$

For the second term we solve the inequality:

$$\frac{\sigma^2 B^2}{\epsilon MKR} \leq \epsilon \quad\Longrightarrow\quad KR \geq \frac{\sigma^2 B^2}{M\epsilon^2}\,.$$

For the third term we solve the inequality:

$$\frac{\tau^2\sigma^2 B^6}{\epsilon^3 K R^2} \le \epsilon \quad \Longrightarrow \quad K R^2 \ge \frac{\tau^2\sigma^2 B^6}{\epsilon^4}\,.$$

For the fourth term we solve the inequality:

$$\frac{\tau^2\zeta^2 B^6}{\epsilon^3 R^2} \le \epsilon \quad \Longrightarrow \quad R^2 \ge \frac{\tau^2\zeta^2 B^6}{\epsilon^4}\,.$$

For the fifth term we solve the inequality:

$$\frac{Q^2\sigma^4}{\mu^5 K^2 R^4} \le \epsilon \quad \Longrightarrow \quad K^2 R^4 \ge \frac{Q^2\sigma^4 B^{10}}{\epsilon^6}\,.$$

And for the last term:

$$\frac{Q^2\zeta^4 B^{10}}{\epsilon^5 R^4} \le \epsilon \quad \Longrightarrow \quad R^4 \ge \frac{Q^2\zeta^4 B^{10}}{\epsilon^6}\,.$$

Finally, the rate would be the maximum of above terms which we can write:

$$\tilde{\mathcal{O}}\left(\max\left\{\frac{HB^2}{\epsilon}, \frac{\sigma^2 B^2}{M\epsilon^2}, \frac{\tau^2\sigma^2 B^6}{\epsilon^4}, \frac{\tau^2\zeta^2 B^6}{\epsilon^4}, \frac{Q^2\sigma^4 B^{10}}{\epsilon^6}, \frac{Q^2\zeta^4 B^{10}}{\epsilon^6}\right\}\right) \le$$
$$\tilde{\mathcal{O}}\left(\frac{HB^2}{\epsilon} + \frac{\sigma^2 B^2}{M\epsilon^2} + \frac{\tau^2\sigma^2 B^6}{\epsilon^4} + \frac{\tau^2\zeta^2 B^6}{\epsilon^4} + \frac{Q^2\sigma^4 B^{10}}{\epsilon^6} + \frac{Q^2\zeta^4 B^{10}}{\epsilon^6}\right)\,.$$

Or in terms of $K$ and $R$:

$$\tilde{\mathcal{O}}\left(\frac{HB^2}{KR} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(\tau\sigma B^3)^{1/2}}{K^{1/4}R^{1/2}} + \frac{(\tau\zeta B^3)^{1/2}}{R^{1/2}} + \frac{(Q\sigma^2 B^5)^{1/3}}{K^{1/3}R^{2/3}} + \frac{(Q\zeta^2 B^5)^{1/3}}{R^{2/3}}\right)\,.$$

It's worth mentioning that during the process of deriving rate for convex regime from strongly convex regime, we also ignore some logarithmic factors which is why we used the notation $\tilde{\mathcal{O}}(.)$. In fact the term $\frac{\ln\left(\max\left\{2, \frac{M\mu^2 T^2 B^2}{\sigma^2}\right\}\right)}{\mu T}$ (some constants are dropped for simplicity) that we use for tuning the stepsize depends on $\mu$. By the choice of $\mu = \frac{\epsilon}{B^2}$ this term becomes $\frac{B^2\ln\left(\max\left\{2, \frac{M\epsilon^2 T^2}{B^2\sigma^2}\right\}\right)}{\epsilon T}$ which will be used to upper bound the terms in our rate which are in the form of $\frac{C}{D} + \frac{E}{D^2} + \frac{G}{D^4}$ (refer to D.3). We use the fact that $\frac{1}{D} \le \frac{B^2\ln\left(\max\left\{2, \frac{M\epsilon^2 T^2}{B^2\sigma^2}\right\}\right)}{\epsilon T}$. If $2 \ge \frac{M\epsilon^2 T^2}{B^2\sigma^2}$, then there is no extra logarithmic factor appearing in our rate. Only in the case of $2 \le \frac{M\epsilon^2 T^2}{B^2\sigma^2}$ we get an extra factor of $\ln(T^2)$. In this case, for the term $\frac{C}{D}$ as an example we have:

$$\frac{C}{D} \le \frac{CB^2\ln\left(\frac{M\epsilon^2 T^2}{B^2\sigma^2}\right)}{\epsilon T}$$

And to achieve $\epsilon$-accuracy we want this term to be less that $\epsilon$ so we have:

$$\frac{CB^2\ln\left(\frac{M\epsilon^2 T^2}{B^2\sigma^2}\right)}{\epsilon T} \le \epsilon$$

So we need to set $T$ in the following way:

$$
\begin{aligned}
T &\geq \frac{CB^2 \ln\left(\frac{M\epsilon^2T^2}{B^2\sigma^2}\right)}{\epsilon^2} \\
&\geq \frac{CB^2 \ln\left(\frac{M\epsilon^2}{B^2\sigma^2}\right)}{\epsilon^2} = \tilde{\mathcal{O}}(\frac{1}{\epsilon^2})
\end{aligned}
$$

And for the other terms we do the same.

■

## Appendix E. Missing Details from Section 5

### E.1. A stronger variant of Assumption 5 on $S^\star$

We will use the following stronger version of Assumption 2 with $\mathcal{X} = S^\star$, which makes more sense for strongly convex objectives.

**Assumption 7 (Bounded Optima Heterogeneity)** *A set of objectives $H$-smooth and strongly convex objectives $\{F_m\}_{m\in[M]}$, satisfy $\zeta_\star$-optima-heterogeneity if,*

$$
\frac{1}{M}\sum_{m\in[M]} \|x_m^\star - x^\star\|_2^2 \leq \frac{\zeta_\star^2}{H^2},
$$

*where $x_m^\star$ is the optimum of $F_m$ for all $m \in [M]$ and $x^\star$ is the optimum of the average objective $F$.*

Note that using smoothness along with Assumption 7 implies Assumption 2 for $\mathcal{X} = S^\star$. Intuitively, when the machines have unique optima, the most natural measure of heterogeneity is the distance between these optima and the optima of the average objective, which the above assumption measures. Note that a problem instance with objectives of the form (11) satisfying Assumption 7 lies in the problem class $\mathcal{P}_{\zeta_\star}^{H,B,\sigma}$.

### E.2. Proof of Proposition 19

Assuming the local SGD algorithm converges, i.e., the hyper-parameters are set to achieve that and $R \to \infty$, we would like to calculate its fixed point. Let's denote the fixed point by $x_\infty$. Then $x_\infty$ must satisfy the following requirements,

$$
x_\infty = x_\infty + \frac{\beta}{M}\sum_{m\in[M]} \Delta^m(x_\infty) \equiv \sum_{m\in[M]} \Delta^m(x_\infty) = 0,
$$

where $\Delta^m(x_\infty)$ is the update on machine $m$ for a communication round starting at the fixed point $x_\infty$. Note that the above equation does not depend on $\beta$. Also, note that this is very similar to the average drift assumption of Wang et al. (2022). Unwinding the update, we get the following,

$$
\sum_{m\in[M]}\sum_{k\in[K]} A_m(I - \eta A_m)^{k-1}(x_\infty - x_m^\star) = 0,
$$

$$\Leftrightarrow^{\text{if } \eta = 0} x_\infty = \frac{1}{M} \sum_{m \in [M]} A^{-1} A_m x_m^\star = x^\star,$$

$$\Leftrightarrow^{\text{if } \eta > 0} x_\infty = \frac{1}{M} \sum_{m \in [M]} C^{-1} C_m x_m^\star,$$

where $C_m := I - (I - \eta C_m)^K$, and $C := \frac{1}{M} \sum_{m \in [M]} C_m$. Note that $x_\infty(\eta, K)$ is a function of $\eta, K$ and is unaffected by the choice of $\beta$. The above proof also highlights a key limitation of the assumption 6; they assume that local SGD is going to converge to $x^\star$ as opposed to any other fixed point, defining the average drift at the optimum. This makes their analysis more complicated than it should have been.

### E.3. Proof of Proposition 19

Assuming the local SGD algorithm converges, i.e., the hyper-parameters are set to achieve that and $R \to \infty$, we would like to calculate its fixed point. Let's denote the fixed point by $x_\infty$. Then $x_\infty$ must satisfy the following requirements,

$$x_\infty = x_\infty + \frac{\beta}{M} \sum_{m \in [M]} \Delta^m(x_\infty) \equiv \sum_{m \in [M]} \Delta^m(x_\infty) = 0,$$

where $\Delta^m(x_\infty)$ is the update on machine $m$ for a communication round starting at the fixed point $x_\infty$. Note that the above equation does not depend on $\beta$. Also, note that this is very similar to the average drift assumption of Wang et al. (2022). Unwinding the update, we get the following,

$$\sum_{m \in [M]} \sum_{k \in [K]} A_m (I - \eta A_m)^{k-1} (x_\infty - x_m^\star) = 0,$$

$$\Leftrightarrow^{\text{if } \eta = 0} x_\infty = \frac{1}{M} \sum_{m \in [M]} A^{-1} A_m x_m^\star = x^\star,$$

$$\Leftrightarrow^{\text{if } \eta > 0} x_\infty = \frac{1}{M} \sum_{m \in [M]} C^{-1} C_m x_m^\star,$$

where $C_m := I - (I - \eta C_m)^K$, and $C := \frac{1}{M} \sum_{m \in [M]} C_m$. Note that $x_\infty(\eta, K)$ is a function of $\eta, K$ and is unaffected by the choice of $\beta$. The above proof also highlights a key limitation of the assumption 6; they assume that local SGD is going to converge to $x^\star$ as opposed to any other fixed point, defining the average drift at the optimum. This makes their analysis more complicated than it should have been.

### E.4. Proof of Proposition 21

**Proof** We first bound the distance between $x^\star$ and $\bar{x}^\star$,

$$\|x^\star - \bar{x}^\star\|_2 = \left\| \frac{1}{M} \sum_{m \in [M]} \left( I - A^{-1} A_m \right) x_m^\star \right\|_2,$$

$$
\begin{aligned}
&= \left\| \frac{1}{M} \sum_{m\in[M]} A^{-1}\left(A - A_m\right)\left(x_m^\star - x^\star\right) \right\|_2, \\
&\leq \frac{1}{M} \sum_{m\in[M]} \left\| A^{-1}\left(A - A_m\right)\left(x_m^\star - x^\star\right) \right\|_2, \\
&\leq \frac{1}{M} \sum_{m\in[M]} \left\| A^{-1}\right\|_2 \left\|A - A_m\right\|_2 \left\|x_m^\star - x^\star\right\|_2, \\
&\leq \frac{1}{M} \sum_{m\in[M]} \frac{1}{\mu}\cdot\tau\cdot \left\|x_m^\star - x^\star\right\|_2, \\
&= \frac{\tau\zeta_\star}{H\mu}.
\end{aligned}
$$

We now bound the distance between $x_\infty(K > 1, \eta)$ and $\bar{x}^\star$ similarly,

$$
\begin{aligned}
\left\|x_\infty - \bar{x}^\star\right\|_2 &= \left\| \frac{1}{M} \sum_{m\in[M]} \left(I - C^{-1}C_m\right) x_m^\star \right\|_2, \\
&= \left\| \frac{1}{M} \sum_{m\in[M]} C^{-1}\left(C - C_m\right)\left(x_m^\star - x^\star\right) \right\|_2, \\
&\leq \frac{1}{M} \sum_{m\in[M]} \left\| C^{-1}\left(C - C_m\right)\left(x_m^\star - x^\star\right) \right\|_2, \\
&\leq \frac{1}{M} \sum_{m\in[M]} \left\| C^{-1}\right\|_2 \left\|C - C_m\right\|_2 \left\|x_m^\star - x^\star\right\|_2, \\
&\leq \frac{\zeta_\star}{H} \left\| C^{-1}\right\|_2 \frac{1}{M} \sum_{m\in[M]} \left\|C - C_m\right\|_2, \\
&= \frac{\zeta_\star}{H} \cdot \frac{1}{\lambda_{min}(C)} \frac{1}{M} \sum_{m\in[M]} \left\|C - C_m\right\|_2, \\
&\leq \frac{\zeta_\star}{H} \cdot \frac{1}{1 - \left(1 - \eta \inf_{m\in[M]} \lambda_{min}(A_m)\right)^K} \frac{1}{M} \sum_{m\in[M]} \left\|C - C_m\right\|_2, \\
&\leq \frac{\zeta_\star}{H} \cdot \frac{1}{1 - (1-\eta\mu)^K} \frac{1}{M^2} \sum_{m,n\in[M]} \left\|C_n - C_m\right\|_2, \\
&\leq \frac{\zeta_\star}{H} \cdot \frac{1}{1 - (1-\eta\mu)^K} \sup_{m,n\in[M]} \left\|(I - \eta A_m)^K - (I - \eta A_n)^K\right\|_2, \\
&\leq \frac{\zeta_\star}{H} \cdot \frac{1}{1 - (1-\eta\mu)^K} \sup_{m,n\in[M]} \left\|\eta(A_m - A_n)\right\|_2 K(1-\eta\mu)^{K-1}, \\
&\leq \frac{\zeta_\star}{H} \cdot \frac{\tau\eta K(1-\eta\mu)^{K-1}}{1 - (1-\eta\mu)^K},
\end{aligned}
$$

where to get the second last inequality, we note the following sequence of inequalities with a clever adding and subtracting of several terms,

$$
\begin{aligned}
(I-\eta A_m)^K-(I-\eta A_n)^K &= (I-\eta A_m)^K - \sum_{k\in[K]}(I-\eta A_n)^k(I-\eta A_m)^{K-k} \\
&\quad + \sum_{k\in[K]}(I-\eta A_n)^k(I-\eta A_m)^{K-k} - (I-\eta A_n)^K \ , \\
&= (I-\eta A_m)^K - \sum_{k\in[K]}(I-\eta A_n)^k(I-\eta A_m)^{K-k} \\
&\quad + \sum_{k\in[K-1]}(I-\eta A_n)^k(I-\eta A_m)^{K-k} + (I-\eta A_n)^K - (I-\eta A_n)^K \ , \\
&= (I-\eta A_m)^K + \sum_{k\in[K-1]}(I-\eta A_n)^k(I-\eta A_m)^{K-k} \\
&\quad - \sum_{k\in[K]}(I-\eta A_n)^k(I-\eta A_m)^{K-k} \ , \\
&= \sum_{k\in[0,K-1]}(I-\eta A_n)^k(I-\eta A_m)^{K-k} - \sum_{k\in[K]}(I-\eta A_n)^k(I-\eta A_m)^{K-k} \ , \\
&= \sum_{k\in[K]}(I-\eta A_n)^{k-1}(I-\eta A_m)^{K-k+1} - \sum_{k\in[K]}(I-\eta A_n)^k(I-\eta A_m)^{K-k} \ , \\
&= \sum_{k\in[K]}\left[(I-\eta A_n)^{k-1}(I-\eta A_m)^{K-k+1} - (I-\eta A_n)^k(I-\eta A_m)^{K-k}\right] \ , \\
&= \sum_{k\in[K]}\left[\eta(I-\eta A_n)^{k-1}(A_n-A_m)(I-\eta A_m)^{K-k}\right] \ ,
\end{aligned}
$$

which upon taking the norm implies,

$$
\begin{aligned}
\left\|(I-\eta A_m)^K-(I-\eta A_n)^K\right\|_2 &\leq \sum_{k\in[K]}\eta\left\|I-\eta A_n\right\|_2^{k-1}\left\|A-n-A_m\right\|_2\left\|I-\eta A_m\right\|_2^{K-k} \ , \\
&\leq \eta\sum_{k\in[K]}(1-\eta\mu)^{k-1}\tau(1-\eta\mu)^{K-k} \ , \\
&\leq \eta\tau K(1-\eta\mu)^{K-1} \ ,
\end{aligned}
$$

which finishes the proof of the inequality and hence the proof as well. ■

### E.5. Proof of Theorem 22

Having identified the convergence point of local SGD in the previous section as $x_\infty(\eta, K)$, in this section, we will characterize the rate of convergence to this point.

**Proof** We first write the inner loop iterated on machine $m$ as follows,

$$
\begin{aligned}
x_{r,K}^m &= x_{r,K-1}^m - \eta\left(A_m(x_{r,K-1}^m - x_m^\star)\right), \\
&= x_m^\star + (I-\eta A_m)^K\left(x_{r,0}^m - x_m^\star\right),
\end{aligned}
$$

$$= x_m^\star + (I - \eta A_m)^K (x_{r-1} - x_m^\star).$$

Subtracting $x_{r-1}$ from both sides,

$$\begin{aligned} x_{r,K}^m - x_{r-1} &= x_m^\star - x_{r-1} + (I - \eta A_m)^K (x_{r-1} - x_m^\star), \\ &= \left(I - (I - \eta A_m)^K\right)(x_m^\star - x_{r-1}). \end{aligned}$$

With this in hand, we can write the update across communication rounds as follows,

$$\begin{aligned} x_r &= x_{r-1} + \frac{\beta}{M} \sum_{m \in [M]} \left(x_{r,K}^m - x_{r-1}\right), \\ &= x_{r-1} + \frac{\beta}{M} \sum_{m \in [M]} \left(I - (I - \eta A_m)^K\right)(x_m^\star - x_{r-1}). \end{aligned}$$

Subtracting the fixed point $\hat{x}$ on both sides, we get,

$$\begin{aligned} x_r - \hat{x} &= x_{r-1} - \hat{x} + \frac{\beta}{M} \sum_{m \in [M]} \left(I - (I - \eta A_m)^K\right)(x_m^\star - x_{r-1}), \\ &= x_{r-1} - \hat{x} + \frac{\beta}{M} \sum_{m \in [M]} C_m (x_m^\star - x_{r-1}), \\ &= (I - \beta C) x_{r-1} - x_\infty + \frac{\beta B}{M} \sum_{m \in [M]} C^{-1} C_m x_m^\star, \\ &= (I - \beta C) x_{r-1} - x_\infty + \beta C x_\infty, \\ &= (I - \beta C)(x_{r-1} - x_\infty). \end{aligned}$$

Unrolling this recursion for $R$ communication rounds,

$$x_R - x_\infty = (I - \beta B)^R (x_0 - x_\infty).$$

Noting the initialization $x_0 = 0$ we get,

$$\begin{aligned} x_R &= \left(I - (I - \beta C)^R\right) x_\infty, \\ &= \left(I - \left(I - \frac{\beta}{M} \sum_{m \in [M]} \left(I - (I - \eta A_m)^K\right)\right)^R\right) x_\infty, \\ &= \left(I - \left((1 - \beta) \cdot I + \frac{\beta}{M} \sum_{m \in [M]} (I - \eta A_m)^K\right)^R\right) x_\infty. \end{aligned}$$

In particular, we can bound the distance between $x_R$ and $x_\infty$ as follows,

$$\|x_R - x_\infty\|_2 = \left\|(I - \beta C)^R (x_0 - x_\infty)\right\|_2,$$

$$
\begin{aligned}
&\leq \|I - \beta C\|_2^R \|x_\infty\|_2 \,, \\
&= \left\| I - \beta \cdot \frac{1}{M} \sum_{m \in [M]} \left( I - (I - \eta A_m)^K \right) \right\|_2^R \|x_\infty\|_2 \,,
\end{aligned}
$$

If we ensure that $\beta < 1/\|C\|_2$, then the distance contracts with more communication rounds. Note that,

$$
\begin{aligned}
\|C\|_2 &= \left\| \frac{1}{M} \sum_{m \in [M]} \left( I - (I - \eta A_m)^K \right) \right\|_2 \,, \\
&= \left\| I - \frac{1}{M} \sum_{m \in [M]} (I - \eta A_m)^K \right\|_2 \,, \\
&\leq \max_{m \in [M]} \left\| I - (I - \eta A_m)^K \right\|_2 \,, \\
&\leq 1 - (1 - \eta H)^K.
\end{aligned}
$$

This implies that we will have contraction with communication if $\beta < \frac{1}{1-(1-\eta H)^K}$. In this regime, we can calculate the overall convergence as follows,

$$
\begin{aligned}
\|x_R - x_\infty\|_2 &\leq \|I - \beta C\|_2^R \|x_\infty\|_2 \,, \\
&\leq (1 - \beta \left\|C^{-1}\right\|_2^{-1})^R \|x_\infty\|_2 \,, \\
&\leq \left(1 - \beta \left(1 - (1 - \eta\mu)^K\right)\right)^R \|x_\infty\|_2 \,, \\
&\leq \left(1 - \beta \left(1 - (1 - \eta\mu)^K\right)\right)^R \left\|C^{-1}\right\|_2 \left\| \frac{1}{M} \sum_{m \in [M]} C_m x_m^\star \right\|_2 \,, \\
&\leq \left(1 - \beta \left(1 - (1 - \eta\mu)^K\right)\right)^R \cdot \frac{1 - (1 - \eta H)^K}{1 - (1 - \eta\mu)^K} B \ .
\end{aligned}
$$

We can note that when $\beta = 1$ this simplifies to,

$$
\|x_R - x_\infty\|_2 \leq (1 - \eta\mu)^{KR} \cdot \frac{1 - (1 - \eta H)^K}{1 - (1 - \eta\mu)^K} B \ .
$$

When $\beta = \frac{1}{c} \cdot \frac{1}{1-(1-\eta H)^K}$ for some $c > 1$ this reduces to,

$$
\begin{aligned}
\|x_R - x_\infty\|_2 &\leq \left(1 - \frac{1}{c} \cdot \frac{1 - (1 - \eta\mu)^K}{1 - (1 - \eta H)^K}\right)^R \cdot \frac{1 - (1 - \eta H)^K}{1 - (1 - \eta\mu)^K} B, \\
&= \left(1 - \frac{1}{c\kappa'}\right)^R \kappa' B,
\end{aligned}
$$

where we define,

$$
\kappa' := \frac{1 - (1 - \eta H)^K}{1 - (1 - \eta\mu)^K}.
$$

This finishes the convergence proof for local GD. To finish the proof of the theorem, we combine this result with the fixed point discrepancy implied by Proposition 21. ■

### E.6. Two Stage Algorithms

Based on the discussion in this section, a natural question is how to reduce the fixed-point discrepancy of local SGD. One approach based on the observation in Figure 1 is to reduce the inner step-size of local SGD as iterations proceed so that we get the benefit of more aggressive updates at the beginning of training but can also avoid the fixed-point discrepancy at the end of training. Note that reducing the inner step size actually reduces local SGD to mini-batch SGD in the limit. Thus, an even easier approach to fixing the fixed point issue is sharply transitioning from running local SGD to mini-batch SGD (which has no fixed point issues) at the end of training.

Such a two-stage algorithm has been considered by Hou et al. (2021), motivated by the results of Woodworth et al. (2020b). Theorem 22 provides a more fine-grained perspective on this two-stage algorithm by capturing the effect of $\zeta_\star, \tau$ as opposed to $\zeta(\mathbb{R}^d)$ in Hou et al. (2021). In particular, we run local GD for the first phase until the optimization (first) term in Theorem 22 dominates the convergence, i.e., $R_1 = \theta\left(\frac{\kappa}{K}\ln\left(\frac{H\mu}{\zeta_\star\tau B}\right)\right)$ rounds. At this point, we switch to running GD, i.e., the noiseless variant of mini-batch SGD, for $R_2 = \theta\left(\kappa\ln\left(\frac{\zeta_\star\tau}{H\mu\epsilon}\right)\right)$, for some target accuracy $\epsilon$. Thus, in total, the communication complexity of this algorithm is given by,

$$R(\epsilon) = R_1 + R_2 = \theta\left(\kappa\ln\left(\left(\frac{H\mu}{\zeta_\star\tau}\cdot\frac{1}{B}\right)^{1/K}\left(\frac{\zeta_\star\tau}{H\mu}\cdot\frac{1}{\epsilon}\right)\right)\right).$$

This communication complexity exemplifies the effect of local update steps, where we notice that in the limit $K\to\infty$ the communication complexity of the two-stage algorithm local GD-¿GD is much better than running GD alone as long as $\frac{\zeta_\star\tau}{\mu} << HB$.