# On the Still Unreasonable Effectiveness of Federated Averaging for Heterogeneous Distributed Learning

**Kumar Kshitij Patel** [*1] **Margalit Glasgow** [*2] **Lingxiao Wang** [1] **Nirmit Joshi** [1] **Nathan Srebro** [1]

## Abstract

Federated Averaging/local SGD is the most common optimization method for federated learning that has proven effective in many real-world applications, dominating simple baselines like mini-batch SGD for convex and non-convex objectives. However, theoretically showing the effectiveness of local SGD remains challenging, posing a huge gap between theory and practice. In this paper, we provide new lower bounds for local SGD for convex objectives, ruling out proposed heterogeneity assumptions that try to capture this *"unreasonable"* effectiveness of local SGD. We further show that accelerated mini-batch SGD is, in fact, min-max optimal under some of these heterogeneity notions. Our results indicate that strong convexity of a client's objective might be necessary to utilize several heterogeneity assumptions (Wang et al., 2022). This also highlights the need for new heterogeneity assumptions for federated optimization for the general convex setting, and we discuss some alternative assumptions.

## 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 $F_m := \mathbb{E}_{z_m\sim\mathcal{D}_m}[f(x;z_m)]$, $f(\cdot;\cdot)$ is a convex and differentiable function, and $\mathcal{D}_m$ is the data distribution on machine $m$. We assume that for all $m\in[M]$[1], the objective function $F_m$ is smooth, and the averaged objective $F$ has bounded optima.

**Assumption 1.1** (Smoothness)**.** For all $m\in[M]$, $F_m : \mathbb{R}^d\to\mathbb{R}$ is convex and $H$-smooth, i.e., for any $x,y\in\mathbb{R}^d$, $F_m(x)\le F_m(y)+\langle\nabla F_m(y), x-y\rangle+\frac{H}{2}\|x-y\|_2^2$.

**Assumption 1.2** (Bounded Optima)**.** For all $x^\star\in\arg\min_{x\in\mathbb{R}^d}F(x)$, $\|x^\star\|_2\le B$.

We want to solve problem (1) in the *intermittent communication* (IC) setting (Woodworth et al., 2018; 2021) where the machines work in parallel and are allowed to communicate $R$ times with $K$ time steps in between communication rounds. Thus optimization occurs over $T=KR$ time steps. We assume each machine can access stochastic gradients for its objective during these time steps.

**Assumption 1.3** (Unbiasedness and Bounded Variance)**.** All machines $m\in[M]$ sample $z_t^m\sim\mathcal{D}^m$ at time $t$ to obtain $\nabla f(.;z_t^m)$ s.t., (i) $\mathbb{E}_{z_t^m\sim\mathcal{D}^m}[\nabla f(.;z_t^m)]=\nabla F_m(.)$, and (ii) $\mathbb{E}_{z_t^m\sim\mathcal{D}^m}\left[\|\nabla f(.;z_t^m)-\nabla F_m(.)\|_2^2\right]\le\sigma^2$.

While several algorithms have been proposed for solving problem (1), the most popular algorithms in practice (Kairouz et al., 2019; Wang et al., 2021) are (variants of) local SGD/Federated Averaging (McMahan et al., 2016; Lin et al., 2018) and large mini-batch SGD/Federated SGD (Dekel et al., 2012; Woodworth et al., 2020a). With stochastic gradient access, we can write the local SGD updates[2] as follows for all $m\in[M]$, $t\in[0,T-1]$ (initialize $x_0^m=0$),

$$\begin{aligned} g_t^m &= \nabla f(x_t;z_t^m),\ z_t^m\sim\mathcal{D}^m,\\ x_{t+1}^m &= \begin{cases} x_t^m-\eta g_t^m, & \text{if } (t+1)\nmid K\\ \frac{1}{M}\sum\limits_{n\in[M]}(x_t^n-\eta g_t^n), & \text{if } (t+1)\mid K\end{cases}\end{aligned} \tag{2}$$

where we say $a\mid b$ if $b$ divides $a$, otherwise we say $a\nmid b$. Similarly, we can write the large mini-batch updates as follows (while noting that the iterate doesn't change between communication rounds) for all $m\in[M]$, $t\in[0,T-1]$ (initialize $x_0^m=0$),

$$g_t^m = \nabla f(x_t;z_t^m),\ z_t^m\sim\mathcal{D}^m,$$

[*]Equal contribution [1]Toyota Technological Institute, Chicago, USA [2]Department of Computer Science, Stanford University, USA. Correspondence to: Kumar Kshitij Patel <kkpatel@ttic.edu>.

[1]We denote the set $\{i,i+1,\dots,n\}$ by $[i,n]$ and when $i=1$ by $[n]$.

[2]Another popular variant of local SGD uses an inner-outer stepsize. In this paper we want to highlight the effect of local steps alone, so we use the simpler variant of local SGD.

| Reference | Convergence Rate |
|---|---|
| Woodworth et al. (2020b) | $\frac{HB^2}{KR} + \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(H\zeta^2B^4)^{1/3}}{R^{2/3}}$ |
| Koloskova et al. (2020) | $\frac{HB^2}{R} + \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}}$ |
| Glasgow et al. (2022) | $\frac{HB^2}{KR} + \min\left\{\frac{\sigma B}{\sqrt{KR}}, \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}}\right\} + \frac{\sigma B}{\sqrt{MKR}} + \min\left\{\frac{\zeta_\star^2}{H}, \frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}}\right\}$ |
| Theorem 3.2 | $\frac{HB^2}{R} + \frac{(H\sigma^2B^4)^{1/3}}{K^{1/3}R^{2/3}} + \frac{\sigma B}{\sqrt{MKR}} + \frac{(H\zeta_\star^2B^4)^{1/3}}{R^{2/3}}$. |

*Table 1.* Summary of existing convergence analyses under assumptions 2.1 and 2.3.

$$x_{t+1}^m = \begin{cases} x_t^m, & \text{if } (t+1) \nmid K \\ x_t^m - \frac{\eta}{M} \sum_{\substack{n\in[M] \\ t'\in[t+1-K,t]}} g_{t'}^n, & \text{if } (t+1) \mid K \end{cases} \quad (3)$$

In the homogeneous setting, when $F_m = F$ for $m \in [M]$, and each machine has access to stochastic gradients of $F$, Woodworth et al. (2021) showed that local SGD doesn't dominate the best of large mini-batch and single machine SGD. This is disappointing because, in practice, local SGD often performs better than both these algorithms. This presents a big gap between the theory and practice of federated optimization. There is hope that in the homogeneous setting, the domination of local SGD might still be provable under a higher-order smoothness assumption (Yuan & Ma, 2020; Woodworth et al., 2021; Bullins et al., 2021; Glasgow et al., 2022), as local SGD is min-max optimal for quadratic objectives (Woodworth et al., 2020a). However, maybe the actual reason for this gap is that the homogeneous setting is too simplistic, and in real applications, the machines' objectives are *"similar"* but not the same (Wang et al., 2021). We begin by first revisting some heterogeneity assumptions in the next section.

**Notation.** We use $\cong$, $\preceq$, and $\succeq$ to refer to equality and inequality up to absolute numerical constants.

## 2. First-order Heterogeneity Assumptions

Several works (Khaled et al., 2020; Karimireddy et al., 2020; Koloskova et al., 2020; Woodworth et al., 2020b; Yuan & Ma, 2020; Glasgow et al., 2022; Wang et al., 2022) have tried to capture the similarity between machine's objectives by using *"heterogeneity"* assumptions. The goal is to show a clear theoretical advantage of using local SGD over mini-batch SGD, demonstrating the usefulness of local updates as a design primitive. Woodworth et al. (2020b) first showed such an advantage of local SGD under the following first-order assumption,

**Assumption 2.1** (Bounded First-Order Heterogeneity Everywhere)**.** A set of objectives $\{F_m\}_{m\in[M]}$ satisfy $\zeta$-first-order heterogeneity everywhere if for all $x \in \mathbb{R}^d$,

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

Unfortunately, this assumption can be too restrictive because of the supremum over all $x \in \mathbb{R}^d$. For instance, simple quadratic functions don't satisfy this assumption unless the objectives of the machines have the same hessian.

**Proposition 2.2.** *Let $F_m(x) = \frac{1}{2}x^TA_mx + b_m^Tx + c_m$ for all $m \in [M]$. If $\{F_m\}_{m\in[M]}$ satisfy assumption 2.1 for finite $\zeta$ then for all machines $m$, $A_m = A$, for $A = \frac{1}{M}\sum_{m\in[M]} A_m$.*

This restrictiveness of assumption 2.1 is precisely why several papers (Khaled et al., 2020; Koloskova et al., 2020) have instead considered the following more relaxed assumption[3], which only needs to be satisfied at the optima for $F$.

**Assumption 2.3** (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 $x^\star \in \arg\min_{x\in\mathbb{R}^d} F(x)$,

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

Unfortunately, no known analysis has shown that local SGD improves over large mini-batch SGD under assumption (2.3) (see table 1). And it is unclear, based on the best known lower bound by Glasgow et al. (2022) (see table 1) if these analyses are tight. Based on their lower bound when $\zeta_\star$ and $\sigma$ are small, and $K$ is large, this is better than the upper bound of large (accelerated) mini-batch SGD. 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 a 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

[3]We discuss another relaxation of this assumption used by Karimireddy et al. (Karimireddy et al., 2020) in section 4

such a regime. Thus, local SGD might be strictly better than large mini-batch SGD in some regimes under assumption (2.3). In the next section, we will show this is not the case using a new hard instance. In particular, our hard instance is such that there is a minimizer $x^\star$ of the averaged objective, which is also a minimizer of all machines ensuring $\zeta_\star = 0$.

## 3. Lower Bound Results

We begin with a simple motivating example in two dimensions and on two machines. Assume the objectives of the machines for all $x \in \mathbb{R}^2$ are given by,

$$F_1(x) := H\left(x(1) - x^\star(1)\right)^2, \\ F_2(x) := H\left(x(2) - x^\star(2)\right)^2, \tag{4}$$

where $x^\star \in \mathbb{R}^2$ is the unique optima of $F(x) = \frac{H}{2}\|x - x^\star\|_2^2$. Note that this instance satisfies assumption 1.1 with smoothness constant $2H$, and assumption 2.3 with $\zeta_\star = 0$. On the other hand, this instance is not homogeneous, and it doesn't satisfy (2.1) for any finite $\zeta$. Assume we run local SGD on both the machines initialized at $(0, 0)$ with step-size $\eta < \frac{1}{H}$, then the iterate after $R$ rounds of communication is given by,

$$\bar{x}_R = x^\star\left(1 - \left(\frac{1}{2} + \frac{(1-2\eta H)^K}{2}\right)^R\right). \tag{5}$$

This means that even if $K \to \infty$, $\bar{x}_R$ doesn't converge to $x^\star$ for small $R$[4]. In particular, this already highlights that the old lower bound by Glasgow et al. (2022) can not be tight, at least when $K$ is large. Note that this construction is possible because $\zeta_\star = 0$ doesn't imply we are in the homogeneous regime—all it says is that the machines should have some shared optima. This is in contrast to the stronger assumption (2.1), where $\zeta = 0$ indeed implies that the functions across the machines have to be the same except for constant terms. We formalize this idea further and present following the lower bound in the regime when $\zeta^\star = 0$ (proof in section B).

**Proposition 3.1.** *For any $K \geq 2, R, M, H, B, \sigma$, there exist $\{F_m\}_{m\in[M]}$ satisfying assumptions 1.1, 1.2 and 1.3 such that $\frac{1}{M}\sum_{m=1}^M \|\nabla F_m(\boldsymbol{x}^\star)\|_2^2 = 0$ for $x^\star \in \arg\min_{x\in\mathbb{R}^d} F(x)$, and the final iterate of local SGD initialized at zero with any step size satisfies:*

$$\mathbb{E}\left[F(\hat{x}_R)\right] - F(\boldsymbol{x}^\star) \succeq \frac{HB^2}{R} + \frac{\sigma B}{\sqrt{MKR}} \\ + \min\left\{\frac{\sigma B}{\sqrt{KR}}, \frac{H^{\frac{1}{3}}\sigma^{\frac{2}{3}}B^{\frac{4}{3}}}{K^{\frac{1}{3}}R^{\frac{2}{3}}}\right\}.$$

[4]Note the inner-outer step size variant of local SGD indeed benefits from local steps for the motivating example (5). However, our hard instance in proposition 3.1 works against the inner-outer step-size variant as well

In particular, combining with the previous lower bound by Glasgow et al. (2022), this gives us the following new lower bound for local-SGD.

**Theorem 3.2.** *For any $K \geq 2, R, M, H, B, \sigma, \zeta_\star$, under assumptions 1.1, 1.2, 1.3 and 2.3 the final iterate of local SGD initialized at zero with any step size satisfies:*

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

Combining this with the upper bound by Koloskova et al. (2020), we characterize the optimal convergence rate for local SGD under assumption 2.3. Thus, we can't hope to improve over large mini-batch SGD, which only has the first two terms in its convergence guarantee. But can we say which algorithm is min-max optimal in this setting? It turns out that it is accelerated large mini-batch SGD (Ghadimi & Lan, 2012). We show this by proving the following new algorithm independent lower bound.

**Theorem 3.3.** *For any $K \geq 2, R, M, H, B, \sigma$, there exist $\{F_m\}_{m\in[M]}$ satisfying assumptions 1.1, 1.2, 1.3 and 2.3, with $\frac{1}{M}\sum_{m=1}^M \|\nabla F_m(\boldsymbol{x}^\star)\|_2^2 = 0$, such that the final iterate $\hat{\boldsymbol{x}}$ of any zero-respecting initialized at zero with $R$ rounds of communication and $KR$ gradient computations per machine satisfies,*

$$\mathbb{E}\left[F(\hat{\boldsymbol{x}})\right] - F(\boldsymbol{x}^\star) \succeq \frac{HB^2}{R^2} + \frac{\sigma B}{\sqrt{MKR}}. \tag{6}$$

This lower bound fully characterizes the min-max complexity of distributed optimization under assumption 2.3), showing large mini-batch SGD is the min-max optimal algorithm. This conclusion is surprising, but it closes a recent line of work investigating the effectiveness of local-SGD under this assumption (Khaled et al., 2020; Karimireddy et al., 2020; Koloskova et al., 2020; Woodworth et al., 2020b; Glasgow et al., 2022; Wang et al., 2022). Comparing this with the min-max optimality of large mini-batch SGD in the homogenenous setting (Woodworth et al., 2021; Woodworth, 2021), we conjecture the following:

**Conjecture 3.1.** *There is a class of low heterogeneity problems (which are not homogeneous), under some notion of heterogeneity for which the best of single machine SGD and large mini-batch SGD is min-max optimal.*

We expect that this notion of heterogeneity will only look at the behaviour of different objectives around the optima. This also highlights that perhaps there is a need for assumptions such as assumption 2.1, which might be necessary to show a domination of local SGD over mini-batch SGD. In the next section we look at two alternative assumptions, and highlight how our lower bound implies why they can not be sufficient for showing such a domination.

## 4. Other Heterogeneity Assumptions

The hardness results in the previous section motivate considering other heterogeneity assumptions (and algorithms). In this paper, we will talk about three other heterogeneity assumptions considered in previous work. The first assumption is a second-order version of assumption 2.1, which has recently been successfully used in the non-convex setting by Patel et al. (2022) to show the dominance of local update algorithms.

**Assumption 4.1** (Second-order Heterogeneity)**.** A set of doubly-differentiable objectives $\{F_m\}_{m\in[M]}$ satisfy $\tau$-second-order heterogeneity if for all $x \in \mathbb{R}^d$,

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

Unfortunately, local SGD can not improve upon large mini-batch SGD under assumption 4.1. To see this recall the motivating example in (4) and note that we can easily modify it to satisfy assumption 4.1. In particular for all $x \in \mathbb{R}^3$ define,

$$\begin{aligned} F_1(x) &:= \tau\left(x(1) - x^\star(1)\right)^2 + \frac{H}{2}\left(x(3) - x^\star(3)\right)^2, \\ F_2(x) &:= \tau\left(x(2) - x^\star(2)\right)^2 + \frac{H}{2}\left(x(3) - x^\star(3)\right)^2, \quad (7) \\ F(x) &= \frac{\tau}{2}\left(x(1) - x^\star(1)\right)^2 + \frac{\tau}{2}\left(x(1) - x^\star(1)\right)^2 \\ &\quad + \frac{H}{2}\left(x(3) - x^\star(3)\right)^2. \end{aligned}$$

Note that this construction satisfies assumption 1.1 with smoothness $H$, and assumption 4.1 with $\tau$. The only difference compared to the example in (4) is an additional dimension, where the machines share the objective. Based on a similar calculation as for the example in (4), we can show that the sub-optimality in the first two dimensions does not improve with large $K$. As a result, we can't show an upper bound for local SGD that gets arbitrarily small with $K$, thus precluding any domination over mini-batch SGD.

Next, we consider the first-order assumption introduced by Wang et al. (2022). As discussed, assumption (2.1) is very restrictive. Wang et al. (2022) claim that even assumption (2.3) 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 4.2** (Movement at Optima)**.** Given a 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 4.2 can be a function of $\eta$ and $K$ despite normalizing with $\eta K$. Wang et al. (2022) **conjecture** that assumption 4.2 is much less restrictive than the other first-order assumptions ((2.1) and (2.3)). 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[5]. 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 next that it is not useful in the convex setting. We first note the following proposition, which shows assumption 4.2 can always be satisfied if assumption 2.3 is true.

**Proposition 4.3.** *If the functions of the machines $\{F_m\}_{m\in[M]}$ satisfy assumption 2.3 then we have,*

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

In particular, when $\zeta_\star = 0$ in assumption 2.3, we can take $\rho = 0$ in assumption 4.2. Thus our hard instance in proposition 3.1 satisfies assumption 4.2 with $\rho = 0$. This precludes any improvement in the analysis of local SGD. This disproves the conjecture by (Wang et al., 2022) that controlling the movement of the local iterates at the optima is sufficient to explain the *"unreasonable"* effectiveness of local SGD as even the simplest problem under their notion of heterogeneity doesn't benefit from making local steps! We conjecture the following for non-zero $\rho$.

**Conjecture 4.1.** *Large mini-batch SGD is also min-max optimal under assumptions 1.1, 1.2, and 4.1 for convex objectives.*

This seems true based on our proof of theorem 3.3. Overall, our new lower bounds raise several more questions about the working of local SGD and highlight that none of the existing assumptions are satisfactory in explaining the workings of local SGD in the convex setting. This presents a big gap between the theory and practice of local SGD. It is also an open question to understand if strong convexity of the client objectives is necessary to avoid the ill-effects of data-heterogeneity. Finally, we end this discussion by looking at the relaxed first-order heterogeneity assumption first used by Karimireddy et al. (2020).

**Assumption 4.4** (Relaxed First-Order Heterogeneity Everywhere)**.** A set of objectives $\{F_m\}_{m\in[M]}$ satisfy $(G, D)$-first-order heterogeneity everywhere if for all $x \in \mathbb{R}^d$,

$$\frac{1}{M}\sum_{m\in[M]} \left\|\nabla F_m(x)\right\|_2^2 \leq G^2 + D^2\left\|\nabla F(x)\right\|_2^2.$$

[5]Note that if the client objectives are convex, but the average objective is strongly convex, it is still not sufficient to show an advantage of local updates (c.f., (5)).

Note that this assumption "interpolates" between assumptions 2.3 and 2.1 for different values of $(G, D)$. Furthermore, the restrictiveness of assumption 2.1 pointed out in proposition 2.2 doesn't extend to this assumption. This suggests that this assumption[6] **might be key to proving meaningful results about local SGD**. None of the existing analyses achieves this, including the one by Karimireddy et al. (2020). We leave this direction and the conjectures in this paper to future work.

## Acknowledgements

We thank the anonymous reviewers who helped us improve the writing of the paper. We would also like to thank Anastasia Kolsokova, Sebastian Stich, Satyen Kale, and Zheng Xu for several useful discussions. This research was partly supported by the NSF-BSF award 1718970, the NSF TRIPOD IDEAL award, and the NSF-Simons funded Collaboration on the Theoretical Foundations of Deep Learning.

## References

Bullins, B., Patel, K., Shamir, O., Srebro, N., and Woodworth, B. E. A stochastic newton algorithm for distributed convex optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

Dekel, O., Gilad-Bachrach, R., Shamir, O., and Xiao, L. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research*, 13(1), 2012.

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

Glasgow, M. R., Yuan, H., and Ma, T. Sharp bounds for federated averaging (local sgd) and continuous perspective. In *International Conference on Artificial Intelligence and Statistics*, pp. 9050–9090. PMLR, 2022.

GOLUB, G. H. Cme 302: Numerical linear algebra fall 2005/06 lecture 10. 2005.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. corr. *arXiv preprint arXiv:1912.04977*, 2019.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

[6] Or perhaps a variant such that $\frac{1}{M}\sum_{m\in[M]}\|\nabla F_m(x) - \nabla F(x)\|_2^2 \leq G^2 + D^2\|\nabla F(x)\|_2^2$.

Khaled, A., Mishchenko, K., and Richtárik, P. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4519–4529. PMLR, 2020.

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pp. 5381–5393. PMLR, 2020.

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

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data (2016). *arXiv preprint arXiv:1602.05629*, 2016.

Patel, K. K., Wang, L., Woodworth, B., Bullins, B., and Srebro, N. Towards optimal communication complexity in distributed non-convex optimization. In *Advances in Neural Information Processing Systems*, 2022.

Wang, J., Charles, Z., Xu, Z., Joshi, G., McMahan, H. B., Al-Shedivat, M., Andrew, G., Avestimehr, S., Daly, K., Data, D., et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

Wang, J., Das, R., Joshi, G., Kale, S., Xu, Z., and Zhang, T. On the unreasonable effectiveness of federated averaging with heterogeneous data. *arXiv preprint arXiv:2206.04723*, 2022.

Woodworth, B. The minimax complexity of distributed optimization. *arXiv preprint arXiv:2109.00534*, 2021.

Woodworth, B., Patel, K. K., Stich, S., Dai, Z., Bullins, B., Mcmahan, B., Shamir, O., and Srebro, N. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pp. 10334–10343. PMLR, 2020a.

Woodworth, B. E., Wang, J., Smith, A., McMahan, B., and Srebro, N. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. *Advances in neural information processing systems*, 31, 2018.

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

Woodworth, B. E., Bullins, B., Shamir, O., and Srebro, N. The min-max complexity of distributed stochastic convex optimization with intermittent communication. In *Conference on Learning Theory*, pp. 4386–4437. PMLR, 2021.

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

# A. Missing Proofs from Section 2

## A.1. Proof of Proposition 2.2

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

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

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,

$$\begin{aligned} sup_{x\in\mathbb{R}^d} \left\|\nabla F_m(x) - \nabla F(x)\right\|_2 &\geq \delta \left\|(A_m - A)e_i\right\|_2 - \left\|b_m - b\right\|_2, \\ &\geq \delta \left\|c_{m,i}\right\|_2 - \left\|b_m\right\|_2 - \left\|b\right\|_2. \end{aligned}$$

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. □

# B. Missing Proofs from Section 3

We will use almost the same instance in the proof of Theorem 3.3 and Proposition 3.1.

For Theorem 3.3, for even $m$, let

$$F_m(\boldsymbol{x}) := \frac{H}{2}\left((q^2+1)(q-x_1)^2 + \sum_{i=1}^{\lfloor (d-1)/2 \rfloor} (qx_{2i} - x_{2i+1})^2\right), \tag{8}$$

and for odd $m$, let

$$F_m(\boldsymbol{x}) = \frac{H}{2}\left(\sum_{i=1}^{\lfloor d/2 \rfloor} (qx_{2i-1} - x_{2i})^2\right). \tag{9}$$

Thus we have

$$F(\boldsymbol{x}) = \mathbb{E}_m[F_m(\boldsymbol{x})] = \frac{H}{2}\left((q^2+1)(q-x_1)^2 + \sum_{i=1}^{\boldsymbol{d}} (qx_i - x_{i+1})^2\right). \tag{10}$$

For technical reasons, we make a slight modification for Propositions 3.1. For even $m$, we let

$$F_m(\boldsymbol{x}) := \frac{H}{2}\left((q-x_1)^2 + \sum_{i=1}^{\lfloor (d-1)/2 \rfloor} (qx_{2i} - x_{2i+1})^2\right) + q^2 x_d^2, \tag{11}$$

and for odd $m$, let

$$F_m(\boldsymbol{x}) = \frac{H}{2}\left(\sum_{i=1}^{\lfloor d/2 \rfloor} (qx_{2i-1} - x_{2i})^2\right). \tag{12}$$

Thus we have

$$F(\boldsymbol{x}) = \mathbb{E}_m[F_m(\boldsymbol{x})] = \frac{H}{2}\left((q^2+1)(q-x_1)^2 + \sum_{i=1}^{\boldsymbol{d}} (qx_i - x_{i+1})^2\right). \tag{13}$$

Observe that in both cases, the optimum of $F$ is attained at $\boldsymbol{x}^*$, where $\boldsymbol{x}_i^* = q^i$.

### B.1. Proof of Theorem 3.3

Theorem 3.3 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 3.3.

**Lemma B.1.** *For any $K \geq 2, R, M, H, B, \sigma$, there exist $f(\boldsymbol{x}; \xi)$ and distributions $\{\mathcal{D}_m\}$, each satisfying asssumptions 1.1, 1.2, 1.3, and together satisfying $\frac{1}{M}\sum_{m=1}^{M} \|\nabla F_m(\boldsymbol{x}^\star)\|_2^2 = 0$, such that for initialization $\boldsymbol{x}^{(0,0)} = 0$, the final iterate $\hat{\boldsymbol{x}}$ of any zero-respecting with $R$ rounds of communication and $KR$ gradient computations per machine satisfies*

$$\mathbb{E}\left[F(\hat{\boldsymbol{x}})\right] - F(\boldsymbol{x}^\star) \succeq \frac{HB^2}{R^2}. \tag{14}$$

*Proof.* Consider the division of functions onto machines as 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 $\boldsymbol{x}_0$, where the coordinate $(\boldsymbol{x}_0)_i = q^i$ for all $i < t$, and $(\boldsymbol{x}_0)_i = 0$ for $i \geq t$. Observe that $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 \leq \sum_{i=t}^{\infty} q^{2i} \leq \frac{q^{2t}}{1-q^2} \leq Rq^{2t} \leq B^2$.

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 $\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\|^2 \geq \sum_{i=t+R+1}^{d} q^{2i} \geq \frac{q^{2t+2R+2} - q^{2d}}{1-q^2} = \Omega\left(B^2 q^{2R+2}\right) = \Omega(B^2)$ 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 eg. (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 $F(\hat{\boldsymbol{x}}) - F(\boldsymbol{x}^*) \geq \Omega\left(B^2\right)\Omega\left(\frac{H}{R^2}\right)$, which gives the desired result. □

### B.2. Proof of Proposition 3.1

Proposition 3.1 improves on the previous best lower bounds by introducing the term $\frac{HB^2}{R}$. Combining the following lemma with the instances used in (Glasgow et al., 2022) to achieve the first two terms in the lower bound suffices to prove Proposition 3.1.

**Lemma B.2.** *For any $K \geq 2, R, M, H, B, \sigma$, there exist $f(\boldsymbol{x}; \xi)$ and distributions $\{\mathcal{D}_m\}$, satisfying assumptions 1.1, 1.2, 1.3 and together satisfying $\frac{1}{M}\sum_{m=1}^{M} \|\nabla F_m(\boldsymbol{x}^\star)\|_2^2 = 0$, such that for initialization $\boldsymbol{x}^{(0,0)} = 0$ the final iterate of local SGD with any step size satisfies:*

$$\mathbb{E}\left[F(\boldsymbol{x}^{(R,0)})\right] - F(\boldsymbol{x}^\star) \geq \Omega\left(\frac{HB^2}{R}\right). \tag{15}$$

*Proof of Lemma B.2.* In this proof, we will use the notation $\boldsymbol{x}^{m,r,k}$ to denote the $k$th iterate of local SGD in the $r$th round on machine $m$.

Consider the division of functions onto machines as described above for some sufficiently large $d$. First, we claim that we can reduce understanding local SGD on this instance to understanding GD on $F(\boldsymbol{x})$.

*Claim* 1. Fix any $\boldsymbol{x}^{(r,0)}$, such that we have $\boldsymbol{x}^{(r+1,0)} = \mathbb{E}_m \boldsymbol{x}^{(m,r,K)} = \frac{1}{2}\left(\boldsymbol{x}^{(1,r,K)} + \boldsymbol{x}^{(2,r+1,0)}\right)$. Then for some step size $\bar{\eta} := \frac{\left(1-\left(1-2\eta(q^2+1)\right)^k\right)}{2(q^2+1)}$, we have

$$\boldsymbol{x}^{(r+1,0)} = \boldsymbol{x}^{(r,0)} - \bar{\eta}\nabla F(\boldsymbol{x}^{(r,0)}). \tag{16}$$

*Proof.* To abbreviate, will define $\tilde{\boldsymbol{x}}^{1,k} := \boldsymbol{x}^{(1,r,k)} - \boldsymbol{x}^*$, and $\tilde{\boldsymbol{x}}^{2,k} := \boldsymbol{x}^{(2,r,k)} - \boldsymbol{x}^*$.

Let $M_1$ be the hessian of $F_1$, such that for $k \geq 1$, we have

$$\tilde{\boldsymbol{x}}^{1,k} = (I - \eta M_1)\,\tilde{\boldsymbol{x}}^{1,k-1}. \tag{17}$$

Similarly, letting $M_2$ be the hessian of $F_2$, we have

$$\tilde{\boldsymbol{x}}^{2,k} = (I - \eta M_2)\, \tilde{\boldsymbol{x}}^{2,k-1}. \tag{18}$$

Now we claim that all the eigenvalues of $M_1$ and $M_2$ are either 0 or $H(q^2+1)$. Indeed, we have the following block decomposition of $M_1$ and $M_2$, where $A := H \begin{pmatrix} q^2 & -q \\ -q & 1 \end{pmatrix}$.

$$M_1 = \begin{pmatrix} H(q^2+1) & & & & & & & \\ & A & & & & & & \\ & & A & & & & & \\ & & & A & & & & \\ & & & & A & & & \\ & & & & & A & & \\ & & & & & & \cdots & \end{pmatrix} \tag{19}$$

$$M_2 = \begin{pmatrix} A & & & & & & & \\ & A & & & & & & \\ & & A & & & & & \\ & & & A & & & & \\ & & & & A & & & \\ & & & & & \cdots & & \\ & & & & & & A & \\ & & & & & & & 0 \end{pmatrix} \tag{20}$$

Since the eigenvalues of $A$ are 0 and $H(q^2+1)$, all eigenvalues of $M_1$ and $M_2$ are 0 or $H(q^2+1)$.

It follows that

$$\tilde{\boldsymbol{x}}^{1,K} = (1 - \eta M_1)^K\, \tilde{\boldsymbol{x}}^{1,0} = \left( I - \frac{\left(I - \left(1 - \eta(q^2+1)\right)^K\right)}{q^2+1} M_1 \right) \tilde{\boldsymbol{x}}^{1,0} = (I - \bar{\eta} M_1)\, \tilde{\boldsymbol{x}}^{1,0}, \tag{21}$$

and similarly

$$\tilde{\boldsymbol{x}}^{2,K} = (I - \bar{\eta} M_2)\, \tilde{\boldsymbol{x}}^{2,0}. \tag{22}$$

Thus

$$\boldsymbol{x}^{(r+1,0)} - \boldsymbol{x}^* = \frac{1}{2}\left(\tilde{\boldsymbol{x}}^{(1,K)} + \tilde{\boldsymbol{x}}^{(2,K)}\right) = \left(I - \frac{\bar{\eta}}{2}(M_1 + M_2)\right)\left(\boldsymbol{x}^{(r,0)} - \boldsymbol{x}^*\right). \tag{23}$$

Now since $\nabla F(\boldsymbol{x}) = \frac{1}{2}\left(M_1 + M_2\right)(\boldsymbol{x} - \boldsymbol{x}^*)$, we have

$$\boldsymbol{x}^{(r+1,0)} = \boldsymbol{x}^{(r,0)} - \bar{\eta} \nabla F(\boldsymbol{x}^{(r,0)}). \tag{24}$$

which proves the claim. □

We now need to show the following claim: for any step size, GD on $F$ will converge slowly.

*Claim* 2. For any $R \leq d$, there exists some choice of $q$ and $\boldsymbol{x}_0$ with $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \leq B$, such that for any step size $\eta$, $F(\boldsymbol{x}_R) - F(\boldsymbol{x}^*) \geq \Omega\left(\frac{HB^2}{R}\right)$.

*Proof.* To prove this claim, we need a lower bound for the condition number of the Hessian of $F$. We can then apply Lemma B.3 to yield the result directly.

We will compare the Hessian of $F$ to the nearby matrix with a well-understood spectrum. Let $M := M_1 + M_2$ be the Hessian of $F$, and let $\tilde{M} := M - Hq^2 e_1 e_1^T + Hq^2 e_d e_d^T$, where $e_i$ denotes the $i$th standard basis vector. Then $\tilde{M}$ is a tridiagonal Toeplitz matrix with diagonal entries $H(q^2+1)$, and off-diagonal entries $-Hq$.

It is well-known (see eg. (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 in particular, letting $q = 1 - \frac{1}{R}$, we see that $\tilde{M}$ has at least 2 eigenvalues which are at most $\frac{H}{10R}$ (consider $i = d-1, d$). Let $v_d$ and $v_{d-1}$ be their associated eigenvectors.

We now consider the Rayleigh quotient of $M = \tilde{M} + Hq^2 e_1 e_1^T - Hq^2 e_d e_d^T$. Let $w$ be any unit vector in the span of $v_{d-1}$ and $v_d$ which is orthogonal to $e_1$. Then

$$w^T M w = w^T \left(\tilde{M} + Hq^2 e_1 e_1^T - Hq^2 e_d e_d^T\right) w < w^T \tilde{M} w \leq \frac{H}{10R}. \tag{25}$$

If we can show that $M$ is full rank, then since we already know $M$ is PSD since it is a Hessian, if follows that all eigenvalues of $M$ are strictly positive. Thus this Rayleigh quotient calculation will suffice to show that $M$ has some eigenvector $v$ with positive eigenvalue at most $\frac{H}{10R}$.

To show the full rank of $M$, consider for the sake of contradiction any vector $\boldsymbol{y}$ such that $M\boldsymbol{y} = 0$. Observe (working from the bottom right corner of $M$), that we must have $y_d = qy_{d-1}$, and inductively, $y_{i+1} = qy_i$ for all $i \geq 1$. However, when we get to the constraint given by the first row of $M$, we see that $y_2 = qy_1$ is not possible. Thus we have a contradiction. It follows that $M$ has a positive eigenvalue of at most $\frac{H}{10R}$.

To show that the condition number of $M$ is at least $\Omega\left(\frac{1}{R}\right)$, observe that the top eigenvalue is at least $\frac{1}{d}$ times the trace, which is $H(1+q^2)$. This yields the result.

□

**Lemma B.3.** *Let $F(\boldsymbol{x})$ be a convex quadratic function whose Hessian has top eigenvalue $H$ and condition number $\kappa \geq 6$ and unique minima $\boldsymbol{x}^*$. Let $\hat{\boldsymbol{x}}_R$ be any linear combination of the iterates $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_R$. Then for any $B$, there exists some $\boldsymbol{x}_0$ with $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2 \leq B$ such that for any step size $\eta$, $F(\hat{\boldsymbol{x}}_R) - F(\boldsymbol{x}^*) \geq HB^2 \frac{1}{\kappa} e^{-R/\kappa}$. In particular, if $\kappa = \Omega(R)$, then $F(\hat{\boldsymbol{x}}_R) - F(\boldsymbol{x}^*) \geq \Omega\left(\frac{HB^2}{R}\right)$.*

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

Let $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ be the eigenvectors of $M$ with the greatest and least eigenvalues. Consider the initialization $\boldsymbol{x}_0 := B\left(\frac{\boldsymbol{v}_1 + \boldsymbol{v}_2}{\sqrt{2}}\right) + \boldsymbol{x}^*$, which satisfies $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|_2 = B$.

Then solving for the GD iterates in closed form, we have

$$\boldsymbol{x}_R - \boldsymbol{x}^* = \frac{B}{\sqrt{2}}(1 - \eta H)^R \boldsymbol{v}_1 + \frac{B}{\sqrt{2}}\left(1 - \eta\frac{H}{\kappa}\right)^R \boldsymbol{v}_2. \tag{26}$$

Observe that if $\eta \geq \frac{3}{H}$, then the iterates explode and we have $F(\boldsymbol{x}_R) \geq F(\boldsymbol{x}_0) \geq \Omega\left(HB^2\right)$.

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

$$F(\boldsymbol{x}_R) - F(\boldsymbol{x}^*) \geq \frac{1}{2}\left(\frac{B}{\sqrt{2}}\left(1 - \frac{3}{\kappa}\right)^R \boldsymbol{v}_2\right)^T M \left(\frac{B}{\sqrt{2}}\left(1 - \frac{3}{\kappa}\right)^R \boldsymbol{v}_2\right) \tag{27}$$

$$= \frac{B^2}{4}\left(1 - \frac{3}{\kappa}\right)^{2R} \boldsymbol{v}_2^T M \boldsymbol{v}_2 \tag{28}$$

$$= \frac{B^2}{4}\left(1 - \frac{3}{\kappa}\right)^{2R} \frac{H}{\kappa} \tag{29}$$

$$\geq \frac{HB^2}{4\kappa} e^{-12R/\kappa}. \tag{30}$$

The result follows. □

□

## C. Missing Proofs from Section 4

### C.1. Proof of Proposition 4.3

*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,

$$\begin{aligned}
\phi(K) &= \left\| \frac{1}{M\eta K} \sum_{m\in[M]} x^\star - \hat{x}_K^m \right\|_2, \\
&= \left\| \frac{1}{M\eta K} \sum_{m\in[M]} x^\star - \hat{x}_{K-1}^m + \eta \nabla F_m(\hat{x}_{K-1}^m) \right\|_2, \\
&\le \left\| \frac{1}{M\eta K} \sum_{m\in[M]} x^\star - \hat{x}_{K-1}^m \right\|_2 + \left\| \frac{1}{MK} \sum_{m\in[M]} \nabla F_m(\hat{x}_{K-1}^m) \right\|_2, \\
&= \phi(K-1) + \left\| \frac{1}{MK} \sum_{m\in[M]} \nabla F_m(\hat{x}_{K-1}^m) - \nabla F_m(x^\star) \right\|_2, \\
&\le \phi(K-1) + \frac{1}{MK} \sum_{m\in[M]} \left\| \nabla F_m(\hat{x}_{K-1}^m) - \nabla F_m(x^\star) \right\|_2, \\
&\le \phi(K-1) + \frac{H}{MK} \sum_{m\in[M]} \left\| \hat{x}_{K-1}^m - x^\star \right\|_2, \\
&= \phi(K-1) + \frac{H}{K} \delta(K-1), \qquad (31)
\end{aligned}$$

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:

$$\begin{aligned}
\delta(k) &= \frac{1}{M} \sum_{m\in[M]} \|\hat{x}_k^m - x^\star\|_2, \\
&\le \frac{1}{M} \sum_{m\in[M]} \left\| \hat{x}_{k-1}^m - x^\star \right\|_2 + \frac{\eta}{M} \sum_{m\in[M]} \left\| \nabla F_m(\hat{x}_m^{k-1}) \right\|_2, \\
&\le \frac{1}{M} \sum_{m\in[M]} \left\| \hat{x}_{k-1}^m - x^\star \right\|_2 + \frac{\eta}{M} \sum_{m\in[M]} \left( \left\| \nabla F_m(\hat{x}_m^{k-1}) - \nabla F_m(x^\star) \right\|_2 + \|\nabla F_m(x^\star)\|_2 \right), \\
&\le \frac{(1+\eta H)}{M} \sum_{m\in[M]} \left\| \hat{x}_{k-1}^m - x^\star \right\|_2 + \eta\zeta_\star, \\
&\le (1+\eta H)\delta(k-1) + \eta\zeta_\star, \\
&\le \frac{\zeta_\star}{H} \left( (1+\eta H)^k - 1 \right). \qquad (\delta(0) = 0) \qquad (32)
\end{aligned}$$

Plugging (32) back into 31 we get,

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

which proves the claim. □