The role of local steps in local SGD

Tiancheng Qin, S. Rasoul Etesami & César A. Uribe

Taylor & Francis
Taylor & Francis Group

Check for updates
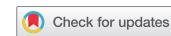
The role of local steps in local SGD

Tiancheng Qin [a], S. Rasoul Etesami [a] and César A. Uribe [b]

[a]Department of Industrial and Enterprise Systems Engineering, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL, USA; [b]Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA

ABSTRACT

We consider the distributed stochastic optimization problem where n agents want to minimize a global function given by the sum of agents' local functions and focus on the heterogeneous setting when agents' local functions are defined over non-i.i.d. datasets. We study the Local SGD method, where agents perform a number of local stochastic gradient steps and occasionally communicate with a central node to improve their local optimization tasks. We analyze the effect of local steps on the convergence rate and the communication complexity of Local SGD. In particular, instead of assuming a fixed number of local steps across all communication rounds, we allow the number of local steps during the jth communication round, H_j, to be different and arbitrary numbers. Our main contribution is to characterize the convergence rate of Local SGD as a function of $\{H_j\}_{j=1}^{R}$ under various settings of strongly convex, convex, and nonconvex local functions, where R is the total number of communication rounds. Based on this characterization, we provide sufficient conditions on the sequence $\{H_j\}_{j=1}^{R}$ such that Local SGD can achieve linear speedup with respect to the number of workers. Furthermore, we propose a new communication strategy with increasing local steps that is superior to constant local steps for strongly convex local functions. On the other hand, for convex and nonconvex local functions, we argue that fixed local steps are the best communication strategy for Local SGD and recover state-of-the-art convergence rate results. Finally, we justify our theoretical results through extensive numerical experiments.

1. Introduction

Stochastic Gradient Descent (SGD) is one of the most commonly used algorithms for parameter optimization of machine learning models. SGD tries to minimize a function f by iteratively updating parameters as: $\mathbf{x}^{t+1} = \mathbf{x}^t - \eta_t \hat{\mathbf{g}}^t$, where $\hat{\mathbf{g}}^t$ is a stochastic gradient of f at $\mathbf{x}^t$ and η_t is the learning rate. However, given the massive scale of many modern ML models and datasets, and taking into account data ownership, privacy, fault tolerance, and scalability, distributed training approaches have recently emerged as a suitable alternative

CONTACT Tiancheng Qin ✉ tq6@illinois.edu

over centralized ones, e.g. parameter server [4], federated learning [7,12,20,25], decentralized stochastic gradient descent [1,10,15,31], decentralized momentum SGD [36], decentralized ADAM [21], among others [3,17,32].

A naive distributed generalization of SGD consists of having multiple agents computing stochastic gradients distributedly, with a central node or fusion center, where local gradients are aggregated and sent back to the agents at every iteration. However, communicating at each iteration induces a large communication overhead, where at each iteration of the algorithm, all agents need to send their gradients to the central node. Then the central node needs to send the agents the aggregated information. Local SGD [18,29,33,38], which can be viewed as a special case of the well-known Federated Averaging algorithm [19] assuming full participation of the agents, presents a suitable solution to the problem. Specifically, in Local SGD, each agent independently runs SGD locally for a number of steps and then aggregates by a central node from time to time only. The main advantage of Local SGD is that multiple local updates would likely move the model parameters much faster to the optimal solution in each communication round, thus effectively reducing the communication overhead at the cost of more local computations.

On the other hand, it remains a delicate problem to choose the number of local steps during each communication round in Local SGD, as too few local steps would result in poor communication efficiency, while too many local steps would lead to slow convergence or even non-convergence of the algorithm. The problem is further complicated by the various scenarios the algorithm is facing, including different types of local objective functions, i.e. strongly convex, general convex or nonconvex functions, as well as whether all agents have the same objective function (the homogeneous case) [9,28,30] or different objective functions (the heterogeneous case) [5,8,9,24,35]. In this paper we focus on the more general heterogeneous case and study strongly convex, general convex and nonconvex local functions respectively.

1.1. Related work

For the case of homogeneous local functions, i.e. when all agents have the same objective function, it was shown in [9,30] that using $\mathcal{O}(n\text{polylog}(T))$ communication rounds, one can achieve convergence rate $\mathcal{O}(\frac{1}{nT})$ for Local SGD with strongly convex functions, where n is the number of agents and T is the number of iterations (or local gradient steps).

A number of recent works have focused on the convergence analysis of Local SGD in heterogeneous setting [5,8,9,24,35]. It is shown that $\mathcal{O}(\frac{1}{nT})$ is both a lower and upper bound for the convergence rate of Local SGD for strongly convex objective functions [8,24]. Moreover, it is known that $\mathcal{O}(\frac{1}{\sqrt{nT}})$ is both a lower and upper bound for the convergence rate of Local SGD for general convex and nonconvex objective functions [9,35]. These two convergence rates are often referred to as *linear speedup with respect to the number of agents N* for strongly convex and convex/nonconvex objective functions, respectively. The name *linear speedup* comes from the implication that with N agents, the algorithm converges N times faster than with just 1 agent [24]. Furthermore, for general convex and nonconvex local functions it is shown that Local SGD can achieve linear speedup with $\mathcal{O}(n^{\frac{3}{4}}T^{\frac{3}{4}})$ communication rounds [8,9]. For strongly convex local functions, the results in [11] implies that Local SGD can achieve $\mathcal{O}(\frac{\log(T)}{nT})$ convergence rate with $\mathcal{O}(\sqrt{nT})$ communication

rounds without the bounded gradient assumption; [24] showed that linear speedup can be achieved with $\mathcal{O}(\sqrt{nT})$ communication rounds, however, their analysis requires the bounded gradient assumption, which is unrealistic in certain cases (see, e.g. [9]).

On the other hand, while most of the works mentioned above assume a fixed number of local steps across all communication rounds, several recent works have proposed different communication strategies for Local SGD to reduce communication costs further. Specifically, in the homogeneous setting, [34] proposed an adaptive communication strategy that gradually increases communication frequency for training neural networks. [6] gave an error bound of Local SGD that explicitly relies on the sequence of local steps for nonconvex local functions that satisfy the Polyak–ojasiewicz condition and proposed decreasing communication frequency. Recently, [27] proposed a linearly increasing number of local steps for strongly convex objective functions and theoretically showed its better communication efficiency. This result has been further generalized in [22] to the network settings. In the heterogeneous setting, [16] proposed decreasing communication frequency such that a number of fully synchronized SGD steps are performed, followed by Local SGD with a fixed number of local steps. On the contrary, [14] proposed increasing communication frequency such that the number of local steps decreases exponentially until it reaches unit local steps.

However, with the exception of [6], the aforementioned works focused on certain specific communication strategies, thus failing to provide a comprehensive understanding of the role of local steps in the convergence rate of Local SGD. In this paper, we aim to systematically study the role of local steps in Local SGD in the more general heterogeneous setting and study strongly convex, general convex, and nonconvex local functions, respectively.

Finally, [26] proposed Local SGD with an exponentially increasing communication intervals, and showed upper bounds for the convergence rate of Local SGD that match with the bounds shown in this paper.[1] However, [26] only focused on a specific communication strategy that relies heavily on prior knowledge for the initialization of the algorithm (e.g. setting up initial stepsize and initial communication interval), while this paper provides a much broader analysis with less reliance on such prior knowledge.

1.2. Contributions and organization

In this paper, we allow the number of local steps during the jth communication round, H_j, to be different integer numbers, and characterize the convergence rate of Local SGD with respect to the sequence $\{H_j\}_{j=1}^R$, where R is the total number of communication rounds. Such a characterization enables us to study the convergence rate of Local SGD for any general communication pattern. We summarize our contributions as follows:

- We characterize the convergence rate of Local SGD explicitly as a function of $\{H_i\}_{i=1}^R$ under various settings of strongly convex, convex, and nonconvex local functions.
- We provide sufficient conditions on the sequence $\{H_j\}_{j=1}^R$ such that Local SGD can achieve linear speedup with respect to the number of agents, i.e. $\mathcal{O}(\frac{1}{nT})$ convergence rate for strongly convex local functions and $\mathcal{O}(\frac{1}{\sqrt{nT}})$ convergence rate for general convex or nonconvex local functions, that covers broad classes of communication strategies.

- For strongly convex local functions, we propose a new communication strategy for the Local SGD with an increasing number of local steps and show it can achieve linear speedup convergence rate with $\mathcal{O}(\sqrt{nT})$ communication rounds without any assumption on the boundedness of the gradients. To our knowledge, this is tighter than all existing bounds of Local SGD with constant number of local steps.[2] We also validate the superiority of the communication strategy through numerical experiments.
- Based on our convergence rate characterization, we argue that using fixed local steps is the best communication strategy for Local SGD in the case of convex and nonconvex local functions. Our results imply that Local SGD can achieve a linear speedup convergence rate with $\mathcal{O}(n^{\frac{3}{4}} T^{\frac{3}{4}})$ communication rounds, which matches the best-known results in this setting [8,9]. Moreover, we show through numerical experiments that this bound on the number of communication rounds to achieve linear speedup is almost tight.

The paper is organized as follows. Section 2 describes the problem statement. Section 3 states our main results for the case of strongly convex and convex objective functions. Section 3.3 extends our convergence rate analysis to the case of nonconvex functions. Simulation results are given in Section 4, followed by conclusions and future directions in Section 5. For ease of presentation, all the proof details are deferred to the supplementary materials.

2. Problem formulation

We consider the distributed stochastic optimization problem with a set of $[n] = \{1, \ldots, n\}$ agents, where each agent $i \in [n]$ holds a local objective function $f_i \colon \mathbb{R}^d \to \mathbb{R}$ that can be expressed in a stochastic form

$$f_i(\mathbf{x}) = \mathbb{E}_{\xi_i \sim \mathcal{D}_i} F_i(\mathbf{x}, \xi_i). \tag{1}$$

Here, $\mathbf{x} \in \mathbb{R}^d$ is the optimization variable, and $\mathcal{D}_i$ denotes the distribution of random variable ξ_i over the parameter sample space Ω_i for agent i. The agents' goal is to minimize the global objective function $f \colon \mathbb{R}^d \to \mathbb{R}$ given by the average sum of all the local functions or, equivalently, solve the following unconstrained optimization problem

$$f^{\star} := \min_{\mathbf{x} \in \mathbb{R}^d} \left\{ f(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{x}) \right\}, \tag{2}$$

by performing local gradient steps and occasionally communicating with a central node to leverage the samples obtained by the other agents.

We assume throughout the paper that $f(\mathbf{x})$ is bounded below by $f^{\star}$ (i.e. a global minimum exists), $f_i(\mathbf{x})$ is L-smooth for every $i \in [n]$, and $\nabla F_i(\mathbf{x}, \xi_i)$ is an unbiased stochastic gradient of $f_i(\mathbf{x})$, which by now are standard assumptions in the context of federated learning [8,9].

Assumption 2.1: For every $i \in [n]$, $f_i(\mathbf{x})$ is L-smooth in terms of $\mathbf{x}$, that is

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \le L\|\mathbf{x} - \mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, i \in [n].$$

Moreover, for some of our results, we will require functions f_i to be μ-strongly convex with respect to the parameter $\mathbf{x}$ as defined next.

Assumption 2.2: We say $f_i \colon \mathbb{R}^d \to \mathbb{R}$ is μ-(strongly) convex for some $\mu \geq 0$ if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

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

If $\mu = 0$, then f_i is convex but not strongly-convex.

Next, as in [9], we consider the following definition, which allows us to measure the heterogeneity among local functions.

Definition 2.1: Assume (2) admits a unique optimal solution $\mathbf{x}^\star = \arg \min f(\mathbf{x})$. We define

$$\bar{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\xi_i} \left\| \nabla F_i(\mathbf{x}^*, \xi_i) \right\|_2^2.$$

It follows that for all non-degenerate sampling distribution $\mathcal{D}_i$, $\bar{\sigma}^2$ is well-defined and finite and serves as a natural measure of variance in local methods. However, for nonconvex objective functions where a unique $\mathbf{x}^*$ may not exist, as in [8], we consider the following assumption of bounded gradient dissimilarity.

Assumption 2.3 (bounded gradient dissimilarity): We say that the local functions f_i satisfy (G, B)-bounded gradient dissimilarity (or for short (G, B)-BGD) if there exist constants $G \geq 0$ and $B \geq 1$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(\mathbf{x})\|^2 \leq G^2 + B^2 \|\nabla f(\mathbf{x})\|^2, \quad \forall \mathbf{x}.$$

We also assume $\nabla F_i(\mathbf{x}, \xi_i)$ is an unbiased stochastic gradient of $f_i(\mathbf{x})$ with variance bounded by σ^2.

2.1. Local stochastic gradient descent

A popular method for solving (2) in a distributed manner is the local stochastic gradient descent (Local SGD) method. In Local SGD, each agent performs local gradient steps, and a central node will compute the average of all agents' iterates every once in a while to guide agents' iterates toward consensus. Let us denote the total number of iterations in Local SGD by T and the set of communication instances by $\mathcal{I} \subseteq [T]$. Then, in every iteration $t \in [T]$ of the Local SGD i) each agent $i \in [n]$ performs stochastic gradient descent update on its local objective function, and ii) if t is a communication time, i.e. $t \in \mathcal{I}$, each agent $i \in [n]$ sends its current local solution $\mathbf{x}_i^{(t)}$ to the central node and receives the average of all agents' local solutions. The pseudo-code for the Local SGD algorithm is summarized in Algorithm 1.

Finally, we consider the following definition of communication intervals in the Local SGD.

Algorithm 1 Local SGD

1: Input $\mathbf{x}_i^{(0)} = \mathbf{x}^{(0)}$ for $i \in [n]$, total number of iterations T, the step-size sequence $\{\eta_t\}_{t=0}^{T-1}$, the set of communication time instances $\mathcal{I} = \{\tau_j\}_{j=0}^{R}$.
2: **for** $t = 0, \ldots, T-1$ **do**
3:   **for** $i = 1, \ldots, n$ **do**
4:     Sample $\xi_i^{(t)}$, compute $\mathbf{g}_i^t = \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})$
5:     **if** $t + 1 \in \mathcal{I}$ **then**
6:       $\mathbf{x}_i^{(t+1)} = \frac{1}{n} \sum_{j=1}^{n} (\mathbf{x}_j^{(t)} - \eta_t \mathbf{g}_j^t)$
7:     **else**
8:       $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} - \eta_t \mathbf{g}_i^t$
9:     **end if**
10:   **end for**
11: **end for**

Definition 2.2: Given communication time instances $\mathcal{I} = \{\tau_j\}_{j=1}^{R}$, we let $H_j = \tau_j - \tau_{j-1}$ be the length of the jth communication interval, i.e. the number of local steps between the $(j-1)$th and jth communications. Moreover, for any time instance $t \in [\tau_j, \tau_{j+1})$, we define $k(t) = j$. In other words, $k(t)$ is the index such that $\tau_{k(t)} \leq t < \tau_{k(t)+1}$.

Our main objective in this work is to characterize the convergence rate of Algorithm 1 with respect to the sequence of the local steps $\{H_j\}_{j=1}^{R}$ as defined above, when applied to the optimization problem (2).

3. Convergence results for local SGD

In this section, we state our main result for the case of strongly convex and convex functions. To that end, let $\bar{\mathbf{x}}^{(t)}$ and $\bar{\mathbf{g}}^{(t)}$ be the average of agents' iterates and the average of their stochastic gradients at time t, respectively, i.e.

$$\bar{\mathbf{x}}^{(t)} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^{(t)}, \quad \bar{\mathbf{g}}^{(t)} = \frac{1}{n} \sum_{i=1}^{n} \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}).$$

Moreover, define the following parameters

$$r_t = \mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2, \quad V_t = \frac{1}{n} \mathbb{E} \sum_{i=1}^{n} \|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\|^2, \quad e_t = \mathbb{E}[f(\bar{\mathbf{x}}^{(t)})] - f(\mathbf{x}^*),$$

which represent the expected distance of the averaged iterates at time t to the optimum solution, the expected consensus error among agents at time t, and the expected optimality gap at time t.

3.1. Convergence result for strongly convex functions

Theorem 3.1: *Let Assumptions 2.1 and 2.2 hold with $\mu > 0$. Then, the sequence generated by Algorithm 1 with stepsize $\eta_t = \frac{2}{\mu(\beta+t)}$, and any sequence of communication intervals*

$\{H_j\}_{j=1}^R$ and parameter $\beta \geq \frac{20L}{\mu}$ such that

$$H_j \leq \frac{\mu(\beta + \sum_{k=1}^{j-1} H_k)}{12L} \quad \forall j,$$

has the following property:

$$r_T \leq \frac{(\beta - 1)^2}{T^2} r_0 + \frac{12\bar{\sigma}^2}{n\mu^2 T} + \frac{144L\bar{\sigma}^2}{\mu^3 T^2} \sum_{j=1}^R \frac{H_j^3}{\sum_{k=1}^{j-1} H_k + \beta}. \tag{3}$$

where L is the smoothness constant, R is the number of communication rounds, and β is a constant that can be tuned by the Local SGD algorithm to balance the first and third term in (3).

An immediate corollary of Theorem 3.1 is the set of sufficient conditions on the sequence $\{H_j\}_{j=1}^R$ that leads to linear speedup in the convergence of Algorithm 1.

Corollary 3.1: *Assume that $T \geq n$. Let the sequence of local steps $\{H_j\}_{j=1}^R$ have the following properties:*

$$(1)\ H_j \leq \frac{\mu(\beta + \sum_{k=1}^{j-1} H_k)}{12L}, \forall j \quad (2)\ \sum_{j=1}^R H_j = T \quad (3)\ \sum_{j=1}^R \frac{H_j^3}{\sum_{k=1}^{j-1} H_k + \beta} = \mathcal{O}\left(\frac{T}{n}\right).$$

Then, the sequence generated by in Algorithm 1 has the following property $r_T = \mathcal{O}(\frac{1}{nT})$.

Next, using Corollary 3.1, we can analyze two special communication strategies, one with a fixed number of local steps and the other with an increasing number of steps.

Proposition 3.1: *Let Assumptions 2.1 and 2.2 hold with $\mu > 0$. Then,*

(1) *For the constant number of local steps $H_j = \frac{T}{R}, \forall j$ and $\beta = \frac{12LT}{\mu R}$, the convergence rate of Local SGD is upper bounded by $r_T = \mathcal{O}(\frac{1}{nT})$, where $R = \mathcal{O}(\sqrt{nT}\log(nT))$ is the number of communication rounds.*

(2) *For the increasing number of local steps $H_j = \lfloor aj^s \rfloor, \forall j$ with $a = \mathcal{O}(n^{-\frac{s+1}{2}} T^{\frac{1-s}{2}})$, $0 < s < 1$ and $\beta = a\lceil \frac{24L}{\mu}\rceil^s \cdot \frac{12L}{\mu} + 1$,[3] the convergence rate of Local SGD is upper bounded by $r_T = \mathcal{O}(\frac{1}{nT})$ with $R = \mathcal{O}(\sqrt{nT})$ number of communication rounds.*

Remark 3.1: Based on the convergence rate upper bounds of Local SGD in Proposition 3.1, the communication strategy with an increasing number of local steps exhibits better communication efficiency than a fixed number of local steps. Moreover, it is shown in [35] that if adopting a fixed number of local steps, the convergence rate of Local SGD is lower bounded by $\mathcal{O}(\frac{1}{nT})$,[4] suggesting that using the proposed communication strategy of increasing number of local steps is at least aas good as fixed number of local steps.

3.2. Convergence result for convex functions

In this part, we relax the assumption of strong convexity on the local function to merely convex functions and analyze the convergence rate of Algorithm 1 in terms of the number of local steps.

Theorem 3.2: *Let Assumptions 2.1 and 2.2 be satisfied with $\mu = 0$ and set a stepsize as $\eta_t = c\sqrt{\frac{n}{T}}, \forall c \leq \frac{1}{10L}\sqrt{\frac{T}{n}}$. Moreover, set the communication intervals to satisfy $H_j \leq \frac{1}{7L\eta} = \frac{\sqrt{T}}{7Lc\sqrt{n}}, \forall j$. Thus, the iterates generated by Algorithm 1 have the following property:*

$$\frac{1}{T}\sum_{t=0}^{T-1} e_t \leq \frac{2r_0 + 6c^2\bar{\sigma}^2}{c\sqrt{nT}} + \frac{24L\bar{\sigma}^2 c^2 n}{T^2}\sum_{j=1}^{R} H_j^3. \tag{4}$$

An immediate corollary of Theorem 3.2 is a sufficient condition on the sequence $\{H_j\}_{j=1}^{R}$ that leads to linear speedup in the convergence of Algorithm 1.

Corollary 3.2: *Assume that $T \geq n^3$. In order to achieve Linear Speedup, $\frac{1}{T}\sum_{t=0}^{T-1} e_t = \mathcal{O}(\frac{1}{\sqrt{nT}})$, it is enough to select the local steps sequence $\{H_j\}_{j=1}^{R}$ such that*

$$(1)\ H_j \leq \frac{1}{7L\eta} = \frac{\sqrt{T}}{7Lc\sqrt{n}}, \forall j, \quad (2)\ \sum_{j=1}^{R} H_j = T, \quad (3)\ \sum_{j=1}^{R} H_j^3 = \mathcal{O}(\frac{T^{\frac{3}{2}}}{n^{\frac{3}{2}}}).$$

Remark 3.2: A closer look at the bound (4) reveals that in order to minimize the error bound of Local SGD, the sequence $\{H_j\}_{j=1}^{R}$ should minimize $\sum_{j=1}^{R} H_j^3$ subject to $\sum_{j=1}^{R} H_j = T$. This leads to the communication strategy of a fixed number of local steps, i.e. $H_j = \frac{T}{R}$. Therefore, based on the convergence rate of Local SGD derived in Theorem 3.2, for convex local functions, the fixed number of local steps is the best communication strategy for Local SGD. Moreover, from Corollary 3.2, we immediately get that in order to achieve *linear speedup*, the number of communication rounds should be $R = \mathcal{O}((nT)^{3/4})$,[5] which correspond to the number of local steps $H_j = \frac{T^{1/4}}{n^{3/4}}\ \forall j$. This choice of the number of local steps also results in the first and second term in the left-hand side of inequality (4) to be of the same order, which is 'efficient' in a sense.

3.3. Convergence result for non-convex functions

In this section, we focus on the class of nonconvex local functions. However, we need to impose the additional (G, B)-BGD assumption to analyze the convergence rate versus communication complexity trade-off. To state our main result, let us define

$$h_t = \|\nabla f(\bar{\mathbf{x}}^{(t)})\|^2,$$

which is the gradient norm of the average iterates in the Local SGD. Then, we have the following theorem.

Theorem 3.3: *Let Assumptions 2.1 and 2.3 hold, fix a stepsize $\eta_t = c\sqrt{\frac{n}{T}}$, $\forall c \leq \frac{1}{4L}\sqrt{\frac{T}{n}}$,, and set a sequence of communication intervals that satisfy $H_j \leq \frac{1}{7LB\eta} = \frac{\sqrt{T}}{7LBc\sqrt{n}}$, $\forall j$. Then, the sequence generated by Algorithm 1 has the following property:*

$$\frac{1}{T}\sum_{t=0}^{T-1} h_t \leq \frac{8e_0 + 4c^2\sigma^2}{c\sqrt{nT}} + \frac{48L^2(\sigma^2 + G^2)c^2 n}{T^2}\sum_{j=1}^{R} H_j^3. \tag{5}$$

As a corollary of Theorem 3.2, we obtain the following set of sufficient conditions on the sequence $\{H_j\}_{j=1}^{R}$ that leads to linear speedup in the convergence of Algorithm 1.

Corollary 3.3: *Assume that $T \geq n^3$. In order to achieve Linear Speedup, $\frac{1}{T}\sum_{t=0}^{T-1} h_t = \mathcal{O}(\frac{1}{\sqrt{nT}})$, it is enough to select the local steps sequence $\{H_j\}_{j=1}^{R}$ such that*

$$(1)\ H_j \leq \frac{1}{7LB\eta} = \frac{\sqrt{T}}{7LBc\sqrt{n}}, \forall j, \quad (2)\ \sum_{j=1}^{R} H_j = T, \quad (3)\ \sum_{j=1}^{R} H_j^3 = \mathcal{O}\left(\frac{T^{\frac{3}{2}}}{n^{\frac{3}{2}}}\right).$$

Remark 3.3: By taking a closer look at the bound (5), it is easy to see that in order to minimize the error bound of Local SGD, the local steps sequence $\{H_j\}_{j=1}^{R}$ should minimize $\sum_{j=1}^{R} H_j^3$ subject to $\sum_{j=1}^{R} H_j = T$. This leads to the communication strategy of a fixed number of local steps, i.e. $H_j = \frac{T}{R}$. Therefore, using the convergence rate of Local SGD obtained in Theorem 3.3, for nonconvex local functions, we conclude that a fixed number of local steps is the best communication strategy for Local SGD. Moreover, from Corollary 3.3, we immediately get that in order to achieve linear speedup, the number of communication rounds should be $R = \mathcal{O}((nT)^{3/4})$,[6] which correspond to the number of local steps $H_j = \frac{T^{1/4}}{n^{3/4}}$ $\forall j$. This choice of the number of local steps also results in the first and second term in the LHS of inequality (5) to be of the same order, which is 'efficient' in a sense.

4. Numerical results

This section shows the results for two sets of experiments on the MNIST dataset [13] to validate our theoretical findings. We focus on strongly-convex loss functions for the first set of experiments, where we train a logistic regression model with l_2 regularization. We focus on nonconvex loss functions for the second set of experiments, where we train a small, fully connected neural network.

4.1. Logistic regression model for MNIST

In this set of experiments, we distribute the MNIST dataset to $n = 20$ agents and apply Local SGD to train a multinomial logistic regression model with l_2 regularization. We first sort the data by digit label, then divide the dataset into 100 shards and assign each of 20 agents 5 shards. Each agent will have examples of approximately five digits, reflecting moderately heterogeneous data sets.
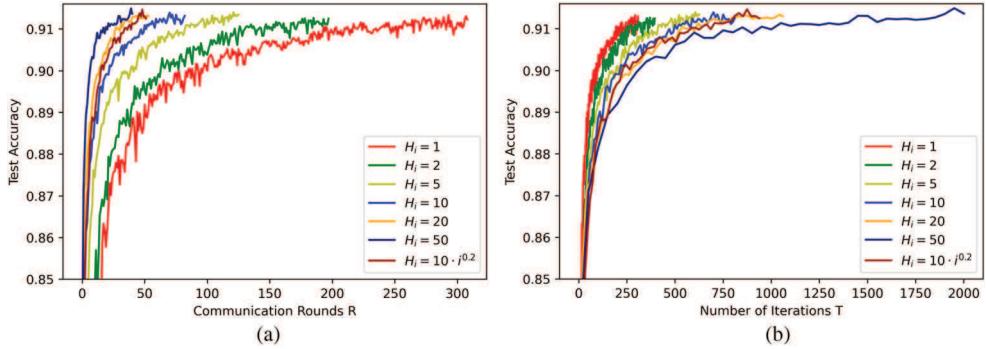
Figure 1. Logistic regression for MNIST. (a) Test accuracy vs. communication rounds for different communication strategies. (b) test accuracy vs. the number of iterations for different communication strategies. Simulation results averaged over 5 runs of the experiment.
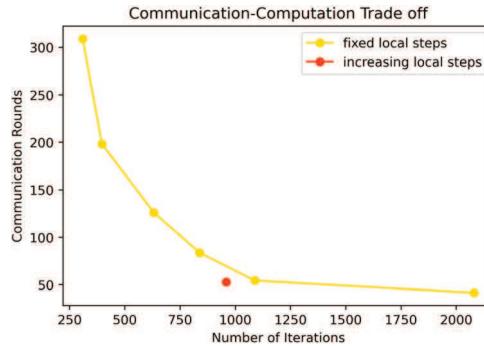

Figure 2. Logistic regression for MNIST. Summary of the number of communication rounds and iterations needed for the model to reach a 91.5% accuracy on the MNIST test dataset for different communication strategies. Yellow dots: fixed number of local steps. Red dot: increasing number of local steps as in Proposition 3.1 with $a = 10$, $s = 0.2$. Simulation results averaged over 5 runs of the experiment.

We evaluate different communication strategies (i.e. various numbers of local steps when following communication strategy with a fixed number of local steps and $a = 10$, $s = 0.2$ when following communication strategy with an increasing number of local steps as in Proposition 3.1) the corresponding communication rounds and iterations needed for the model to reach a 91.5% accuracy on the MNIST test dataset. The simulation results are averaged over 5 independent runs of the experiments and are shown in Figures 1 and 2.

For the set of hyperparameters, we use a training batch size of 8, l_2 regularization parameter $\mu = 0.001$, $\beta = 1000$ and set stepsize at iteration t to be $\eta_t = \frac{\beta}{t+\beta}\eta_0$, where the initial stepsize η_0 is chosen based on a grid search of resolution 10^{-3}.

Figure 1 shows the details of the runs of the experiment. Figure 2 shows the summary of the runs. For example, the upper left yellow dot in Figure 2 corresponds to the average of 5 runs of Local SGD with constant $H_i = 1$, showing that with constant $H_i = 1$ it took the algorithm an average of ~ 305 communication rounds as well as total iterations to reach 91.5% accuracy.
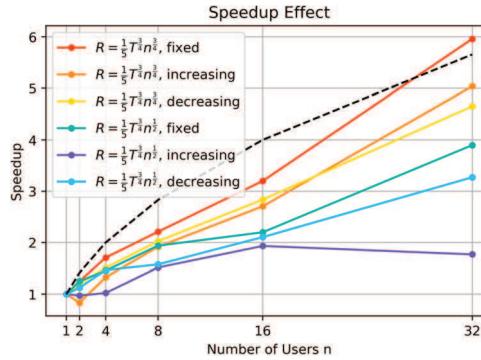
Figure 3. Speedup curve for Local SGD with different communication strategies. Fixed: fixed number of local steps subject to a total of $R = \frac{1}{5}T^{\frac{3}{4}}n^{\frac{3}{4}}$ or $R = \frac{1}{5}T^{\frac{3}{4}}n^{\frac{1}{2}}$ communication rounds. Increasing: $H_i \propto i^2$ subject to a total of $R = \frac{1}{5}T^{\frac{3}{4}}n^{\frac{3}{4}}$ or $R = \frac{1}{5}T^{\frac{3}{4}}n^{\frac{1}{2}}$ communication rounds. Decreasing: $H_i \propto (R-i)^2$ subject to a total of $R = \frac{1}{5}T^{\frac{3}{4}}n^{\frac{3}{4}}$ or $R = \frac{1}{5}T^{\frac{3}{4}}n^{\frac{1}{2}}$ communication rounds. The dashed black line corresponds to speedup $= \sqrt{n}$.

Communication-Computation Trade-Off: In general, we can observe a communication-computation trade-off such that with more local computation (corresponding to a larger number of iterations T), less communication is needed (corresponding to a smaller number of communication rounds R) for the model to reach a certain accuracy.

Better Communication Efficiency with Increasing Number of Local Steps: As we can see from Figure 2, the red dot lies to the bottom left of the yellow line, which shows that the communication strategy of an increasing number of local steps is indeed more communication efficient than a fixed number of local steps, thus validating our claim in Remark 3.1.

4.2. Neural network for MNIST

In this set of experiments, we distribute the MNIST dataset to n agents and apply Local SGD to train a fully-connected neural network (2NN) with 2-hidden layers with 50 units each using ReLu activations (42310 total parameters).[7] We first sort the data by digit label, then divide the dataset into n shards and assign each of n agents 1 shards. Each agent will have examples of approximately one digit, reflecting the most heterogeneous data sets.

We evaluate the speedup effect of the number of agents n for different communication strategies. In particular, we set a fixed number of $T = 20,000$ iterations and run Local SGD for T iterations with different communication strategies, a different number of agents n, and a different number of communication rounds R. After that, a speedup factor is derived by dividing the expected error of a single worker SGD at the final iterate T by the expected error of Local SGD with different communication strategies and a different number of agents n at the final iterate T. We plot the speedup curve in Figure 3. In the case of linear speedup, we should expect the dashed black line on the graph, corresponding to speedup $= \sqrt{n}$.

We use a training batch size of 1 and choose stepsize η based on a grid search of resolution 10^{-3}. The simulation results are averaged over 5 independent runs of the experiments.

Better Performance with Fixed Number of Local Steps: We can observe from Figure 3 that Local SGD with fixed number of local steps significantly outperforms its increasing or decreasing number of local steps counterparts in both settings of $R = \frac{1}{5}T^{3/4}n^{3/4}$ (corresponding to sufficient communication) and $R = \frac{1}{5}T^{3/4}n^{1/2}$ (corresponding to insufficient communication). This validates our claim in Remark 3.3 that a fixed number of local steps is the best communication strategy for Local SGD for nonconvex local functions.

Almost Tight Bound for $R = \mathcal{O}((nT)^{3/4})$ to Achieve Linear Speedup: Another observation from Figure 3 is that while setting $R = \mathcal{O}((nT)^{3/4})$ and following a communication strategy of a fixed number of local steps, Local SGD successfully achieved linear speedup, as expected, decreasing R by a factor of $n^{\frac{1}{4}}$ fails for Local SGD to achieve linear speedup, even with the best communication strategy of a fixed number of local steps. This suggests that the bound of $R = \mathcal{O}((nT)^{3/4})$ to achieve linear speedup is close to tight.

5. Conclusions

In this paper, we analyzed the role of local steps in Local SGD in the heterogeneous data setting. We characterized the convergence rate of Local SGD as a function of the sequence of the local steps $\{H_i\}_{i=1}^{R}$ under various settings of strongly convex, convex, and nonconvex local functions. Based on this characterization, we gave sufficient conditions on the sequence $\{H_i\}_{i=1}^{R}$ that covers broad classes of communication strategies such that Local SGD can achieve linear speedup. Furthermore, for strongly convex local functions, we proposed a new communication strategy with increasing local steps that enjoy better performance than the vanilla fixed local steps communication strategy theoretically and in numerical experiments. We argued that fixed local steps are the best communication strategy for Local SGD and recover state-of-the-art convergence rate results for convex and nonconvex local functions. Such an argument is validated by numerical experiments, which showed that the results are almost tight.

As a future research direction, one can consider analyzing the role of local steps in other federated optimization methods, e.g.SCAFFOLD [8], FedAC [37]. Moreover, generalizing our work to directed networks in which agents communicate with their neighbours rather than a central node is another interesting research problem, *e.g.* for Stochastic Gradient Push algorithm [2]. Also, we only considered the role of local steps in Local SGD with full agent participation; generalizing it to the partial participation setting is yet another interesting problem.

Notes

1. Our work was done independently of [26], which was brought to our attention during the review process.
2. For constant number of local steps, the results in [11] implies that Local SGD can achieve $\mathcal{O}(\frac{\log(T)}{nT})$ convergence rate with $\mathcal{O}(\sqrt{nT})$ communication rounds without the bounded gradient assumption; [24] showed that linear speedup can be achieved with $\mathcal{O}(\sqrt{nT})$ communication

rounds, however, their analysis requires the bounded gradient assumption. Recently, [26] proposed a communication strategy with exponentially increasing number of local steps (different from ours) and also showed $\mathcal{O}(\frac{1}{nT})$ convergence rate with $\mathcal{O}(\sqrt{nT})$ communication rounds.

3. It is easy to see that for such choice of parameters conditions (2) and (3) in Corollary 3.1 are satisfied. For the proof of condition (1), we refer to Appendix A.5.

4. In fact, Theorem 2 in [35] stated a lower bound of $\mathcal{O}(\frac{1}{nT}) + \mathcal{O}(\frac{T}{n^2 R^3}) + \min\{\mathcal{O}(\exp(-\frac{R^2}{T})),$ $\mathcal{O}(\frac{T}{R^3})\}$. The higher order terms also suggest that in order to achieve $\mathcal{O}(\frac{1}{nT})$ convergence rate, R needs to be at least $\max(\mathcal{O}(n^{-1/3}T^{2/3}), \mathcal{O}((T\log(nT))^{1/2}))$.

5. This matches the best-known results in the setting [9].

6. This matches the best-known results in the setting [8].

7. We have deliberately chosen to train a small neural network to avoid getting an overparameterized model, in which case the convergence rate of Local SGD would be different [23].

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

Notes on contributors

Tiancheng Qin is a Ph.D student in the department of industrial and systems engineering at the University of Illinois Urbana-Champaign (UIUC) under the supervision of Dr. S. Rasoul Etesami, where his research interests include distributed optimization, federated learning and reinforcement learning. He received a bachelor's degree from the School of Mathematical Sciences, Peking University, China in 2018.

S. Rasoul Etesami is an sssistant professor in the department of industrial and systems engineering at the University of Illinois Urbana-Champaign (UIUC), where he is also affiliated with the department of electrical and computer engineering and the Coordinated Science Laboratory. From 2016 to 2017, he was a postdoctoral research fellow in the department of electrical engineering at Princeton University and WINLAB. He received his Ph.D. in electrical and eomputer Engineering from the University of Illinois Urbana-Champaign in 2015. His research interests include the analysis of complex social, communication and decision-making systems using tools from control and game theory, optimization and learning theory. He was the recipient of the NSF CAREER Award in 2020 and the Air Force Young Investigator Award in 2023.

César A. Uribe is the Louis Owen Jr. Assistant Professor at the department of electrical and computer engineering at Rice University. He received his M.Sc. degrees in systems and control from Delft University of Technology, in The Netherlands, and applied mathematics from the University of Illinois at Urbana-Champaign, in 2013 and 2016, respectively. He also received a PhD degree in electrical and computer engineering at the University of Illinois at Urbana-Champaign in 2018. He was a postdoctoral associate in the Laboratory for Information and Decision Systems-LIDS at the Massachusetts Institute of Technology-MIT until 2020 and is a visiting professor at the Moscow Institute of Physics and Technology. His research interests include distributed learning and optimization, decentralized control, algorithm analysis and computational optimal transport.

ORCID

Tiancheng Qin 🔟 http://orcid.org/0000-0002-0652-7263
S. Rasoul Etesami 🔟 http://orcid.org/0000-0002-2087-6136
César A. Uribe 🔟 http://orcid.org/0000-0002-7080-9724

References

[1] M. Assran and M. Rabbat, *Asynchronous subgradient-push* (2018). Available at arXiv preprint arXiv:1803.08950.

[2] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, *Stochastic gradient push for distributed deep learning*, in *International Conference on Machine Learning*, PMLR, 2019, pp. 344–353.

[3] D. Bertsekas and J. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, 2015.

[4] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and Q. Le, *Large scale distributed deep networks*, in *Advances in Neural Information Processing Systems*, 2012, pp. 1223–1231.

[5] E. Gorbunov, F. Hanzely, and P. Richtárik, *Local SGD: Unified theory and new efficient methods*, in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2021, pp. 3556–3564.

[6] F. Haddadpour, M.M. Kamani, M. Mahdavi, and V. Cadambe, *Local SGD with periodic averaging: Tighter analysis and adaptive synchronization*, Adv. Neural. Inf. Process. Syst. 32 (2019).

[7] P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, and R.G. D'Oliveira, *Advances and open problems in federated learning*, preprint (2019). Available at arXiv:1912.04977.

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

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

[10] A. Koloskova, S.U. Stich, and M. Jaggi, *Decentralized stochastic optimization and gossip algorithms with compressed communication*, preprint (2019), Available at arXiv:1902.00340.

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

[12] J. Konečny, H.B. McMahan, D. Ramage, and P. Richtárik, *Federated optimization: Distributed machine learning for on-device intelligence*, preprint (2016), Available at arXiv:1610.02527.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, Proc. IEEE 86 (1998), pp. 2278–2324.

[14] X. Li, W. Yang, S. Wang, and Z. Zhang, *Communication-efficient local decentralized SGD methods*, preprint (2019), Available at arXiv:1910.09126.

[15] X. Lian, C. Zhang, H. Zhang, C.J. Hsieh, W. Zhang, and J. Liu, *Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent*, in *Advances in Neural Information Processing Systems*, 2017, pp. 5330–5340.

[16] T. Lin, S.U. Stich, K.K. Patel, and M. Jaggi, *Don't use large mini-batches, use local SGD*, preprint (2018), Available at arXiv:1808.07217.

[17] Y. Lu and C. De Sa, *Moniqua: Modulo quantized communication in decentralized SGD*, preprint (2020), Available at arXiv:2002.11787.

[18] Y. Lu, J. Nash, and C. De Sa, *Mixml: A unified analysis of weakly consistent parallel learning*, preprint (2020), Available at arXiv:2005.06706.

[19] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B.A. y Arcas, *Communication-efficient learning of deep networks from decentralized data*, in *Artificial Intelligence and Statistics*, PMLR, 2017, pp. 1273–1282.

[20] H.B. McMahan, E. Moore, D. Ramage, and B.A. y Arcas, *Federated learning of deep networks using model averaging. corr abs/1602.05629 (2016)*, preprint (2016), arXiv:1602.05629.

[21] P. Nazari, D.A. Tarzanagh, and G. Michailidis, *Dadam: A consensus-based distributed adaptive gradient method for online optimization*, preprint (2019), Available at arXiv:1901.09109.

[22] T. Qin, S.R. Etesami, and C.A. Uribe, *Communication-efficient decentralized local SGD over undirected networks*, in *2021 60th IEEE Conference on Decision and Control (CDC)*, IEEE, 2021, pp. 3361–3366.

[23] T. Qin, S.R. Etesami, and C.A. Uribe, *Faster convergence of local SGD for over-parameterized models*, preprint (2022), Available at arXiv:2201.12719.

[24] Z. Qu, K. Lin, J. Kalagnanam, Z. Li, J. Zhou, and Z. Zhou, *Federated learning's blessing: Fedavg has linear speedup*, preprint (2020), Available at arXiv:2007.05690.

[25] N. Rieke, J. Hancox, W. Li, F. Milletari, H.R. Roth, S. Albarqouni, S. Bakas, M.N. Galtier, B.A. Landman, K. Maier-Hein, and S. Ourselin, *The future of digital health with federated learning*, NPJ Digit. Med. 3 (2020), pp. 1–7.

[26] S. Shen, Y. Cheng, J. Liu, and L. Xu, *STL-SGD: Speeding up local SGD with stagewise communication period*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2021, pp. 9576–9584.

[27] A. Spiridonoff, A. Olshevsky, and I. Paschalidis, *Communication-efficient SGD: From local SGD to one-shot averaging*, Adv. Neural. Inf. Process. Syst. 34 (2021), pp. 24313–24326.

[28] A. Spiridonoff, A. Olshevsky, and I.C. Paschalidis, *Local SGD with a communication overhead depending only on the number of workers*, preprint (2020), Available at arXiv:2006.02582.

[29] S.U. Stich, *Local SGD converges fast and communicates little*, preprint (2018), Available at arXiv:1805.09767.

[30] S.U. Stich and S.P. Karimireddy, *The error-feedback framework: Better rates for SGD with delayed gradients and compressed communication*, preprint (2019), Available at arXiv:1909.05350.

[31] H. Tang, X. Lian, M. Yan, C. Zhang, and J. Liu, *Decentralized training over decentralized data*, preprint (2018), Available at arXiv:1803.07068.

[32] H. Tang, C. Yu, X. Lian, T. Zhang, and J. Liu, *Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression*, in *International Conference on Machine Learning*, PMLR, 2019, pp. 6155–6165.

[33] J. Wang and G. Joshi, *Cooperative SGD: A unified framework for the design and analysis of communication-efficient SGD algorithms*, preprint (2018), Available at arXiv:1808.07576.

[34] J. Wang and G. Joshi, *Adaptive communication strategies to achieve the best error-runtime trade-off in local-update SGD*, Proc. Mach. Learn. Syst. 1 (2019), pp. 212–229.

[35] B. Woodworth, K.K. Patel, and N. Srebro, *Minibatch vs local SGD for heterogeneous distributed learning*, preprint (2020), Available at arXiv:2006.04735.

[36] H. Yu, R. Jin, and S. Yang, *On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization* preprint (2019), Available at arXiv:1905.03817.

[37] H. Yuan and T. Ma, *Federated accelerated stochastic gradient descent*, Adv. Neural. Inf. Process. Syst. 33 (2020), pp. 5332–5344.

[38] M. Zinkevich, M. Weimer, L. Li, and A.J. Smola, *Parallelized stochastic gradient descent*, in *Advances in Neural Information Processing Systems*, 2010, pp. 2595–2603.

Appendix. Omitted proofs

A.1   Proof of Theorem 3.1

In order to prove Theorem 3.1, we first establish the following two lemmas. The first lemma allows us to establish a descent property for the distance of iterates from the optimal point, while the second lemma bounds the consensus error among the agents. The proofs of these lemmas are given in Appendix A.4.

Lemma A.1 (Decent lemma): *Let Assumption 2.2 hold. Then,*

$$r_{t+1} \le (1 - \mu\eta_t)r_t - \eta_t e_t + \frac{3\bar{\sigma}^2}{n}\eta_t^2 + 2L\eta_t V_t.$$

Lemma A.2 (Consensus error lemma): *Let Assumption 2.2 hold. Then,*

$$V_t \le H_{k(t)+1} \sum_{j=\tau_{k(t)}}^{t-1} \eta_j^2(12Le_j + 6\bar{\sigma}^2).$$

where $k(t)$ is the index such that $\tau_{k(t)} \le t < \tau_{k(t)+1}$.

Using Lemmas A.1 and A.2, we can now prove Theorem 3.1.

Proof of Theorem 3.1: For $\eta_t = \frac{2}{\mu(\beta+t)}$, it is easy to see that $(t + \beta)^2(1 - \mu\eta_t) = (t + \beta)(t + \beta - 2) \le (t + \beta - 1)^2$. Thus, if we multiply both sides of the expression in Lemma A.1 by $(t + \beta)^2$, we can write

$$(t + \beta)^2 r_{t+1} \le (t + \beta - 1)^2 r_t - (t + \beta)^2 \eta_t e_t + (t + \beta)^2 \frac{3\bar{\sigma}^2}{n}\eta_t^2 + (t + \beta)^2 2L\eta_t V_t$$

$$= (t + \beta - 1)^2 r_t - \frac{2(t + \beta)}{\mu}e_t + \frac{12\bar{\sigma}^2}{n\mu^2} + \frac{4L(t + \beta)}{\mu}V_t.$$

Summing this relation over $t = 0, \ldots, T - 1$, we get

$$(T + \beta - 1)^2 r_T \le (\beta - 1)^2 r_0 - \sum_{t=0}^{T-1}\frac{2(t + \beta)}{\mu}e_t + \frac{12\bar{\sigma}^2 T}{n\mu^2} + \frac{4L}{\mu}\sum_{t=0}^{T-1}(t + \beta)V_t. \qquad \text{(A1)}$$

Next, we use Lemma A.2 to bound the last term $\sum_{t=0}^{T-1}(t + \beta)V_t$ in the above expression (A1). We have

$$\sum_{t=0}^{T-1}(t + \beta)V_t \le \sum_{t=0}^{T-1}(t + \beta)H_{k(t)+1}\sum_{j=\tau_{k(t)}}^{t-1}\eta_j^2(12Le_j + 6\bar{\sigma}^2)$$

$$= \sum_{j=0}^{T-2}\eta_j^2(12Le_j + 6\bar{\sigma}^2)\sum_{t=j+1}^{\tau_{k(j)+1}-1}(t + \beta)H_{k(t)+1}$$

$$= \sum_{j=0}^{T-2}\eta_j^2(12Le_j + 6\bar{\sigma}^2)H_{k(j)+1}\sum_{t=j+1}^{\tau_{k(j)+1}-1}(t + \beta), \qquad \text{(A2)}$$

where the last equality holds because $k(t) = k(j)$ for any $t \in [j + 1, \tau_{k(j)+1} - 1]$. Moreover, using the assumption on the communication intervals, we have

$$\tau_{k(j)+1} - \tau_{k(j)} = H_{k(j)+1} \le \beta + \sum_{\ell=1}^{k(j)}H_\ell = \beta + \tau_{k(j)},$$

which implies $\tau_{k(j)+1} \le \beta + 2\tau_{k(j)}$. Using this relation together with $\tau_{k(j)} \le j < \tau_{k(j)+1}$, we can write

$$\sum_{t=j+1}^{\tau_{k(j)+1}-1}(t + \beta) \le \sum_{t=\tau_{k(j)}}^{\tau_{k(j)+1}-1}(t + \beta)$$

$$\le (\tau_{k(j)+1} - \tau_{k(j)})\left(\beta + \frac{\tau_{k(j)+1} + \tau_{k(j)}}{2}\right)$$

$$\leq H_{k(j)+1}\left(\beta + \frac{\beta + 3\tau_{k(j)}}{2}\right) \leq \frac{3}{2}H_{k(j)+1}(j + \beta).$$

Substituting this relation into (A2), we get

$$\sum_{t=0}^{T-1}(t + \beta)V_t \leq \sum_{j=0}^{T-2}\eta_j^2(18Le_j + 9\bar{\sigma}^2)(j + \beta)H_{k(j)+1}^2$$

$$= \sum_{t=0}^{T-2}\left(\frac{72L}{\mu^2}e_t + \frac{36}{\mu^2}\bar{\sigma}^2\right)\frac{H_{k(t)+1}^2}{t + \beta}.$$

where in the second equality we have used $\eta_t = \frac{2}{\mu(\beta+t)}$ and relabeled the index j by t. Finally, if we substitute the above relation into (A1), we obtain

$$(T + \beta - 1)^2 r_T - (\beta - 1)^2 r_0 \leq \frac{12\bar{\sigma}^2 T}{n\mu^2} + \sum_{t=0}^{T-1}\left(\frac{288L^2}{\mu^3}\frac{H_{k(t)+1}^2}{t + \beta} - \frac{2(t + \beta)}{\mu}\right)e_t$$

$$+ \frac{144L\bar{\sigma}^2}{\mu^3}\sum_{t=0}^{T-2}\frac{H_{k(t)+1}^2}{t + \beta}. \tag{A3}$$

Now, using the condition on the length of communication intervals in the theorem statement, we know that

$$H_{k(t)+1} \leq \frac{\mu(\beta + \sum_{j=1}^{k(t)}H_j)}{12L} = \frac{\mu(\beta + \tau_{k(t)})}{12L} \leq \frac{\mu(\beta + t)}{12L}.$$

Substituting this bound in (A3) we obtain

$$(T + \beta - 1)^2 r_T - (\beta - 1)^2 r_0 \leq \frac{12\bar{\sigma}^2 T}{n\mu^2} + \frac{144L\bar{\sigma}^2}{\mu^3}\sum_{t=0}^{T-2}\frac{H_{k(t)+1}^2}{t + \beta}$$

$$= \frac{12\bar{\sigma}^2 T}{n\mu^2} + \frac{144L\bar{\sigma}^2}{\mu^3}\sum_{i=1}^{R}\sum_{t=\tau_{i-1}}^{\tau_i-1}\frac{H_{k(t)+1}^2}{t + \beta}$$

$$\leq \frac{12\bar{\sigma}^2 T}{n\mu^2} + \frac{144L\bar{\sigma}^2}{\mu^3}\sum_{i=1}^{R}\frac{H_i^3}{\tau_{i-1} + \beta}$$

$$= \frac{12\bar{\sigma}^2 T}{n\mu^2} + \frac{144L\bar{\sigma}^2}{\mu^3}\sum_{i=1}^{R}\frac{H_i^3}{\sum_{j=1}^{i-1}H_j + \beta},$$

where the second equality holds because for any $t \in [\tau_{i-1}, \tau_i)$, we have $k(t) + 1 = i$. Dividing both sides by T^2, we obtain the desired bound. ■

A.2 Proof of Theorem 3.2

Proof: Let us set $\eta_t = \eta, \forall t$, for some parameter η to be determined later. Substituting $\mu = 0$ in Lemma A.1 and summing over $t = 0, \ldots, T - 1$, we get

$$\eta\sum_{t=0}^{T-1}e_t \leq r_0 - r_T + \frac{3\bar{\sigma}^2\eta^2 T}{n} + 2L\eta\sum_{t=0}^{T-1}V_t. \tag{A4}$$

Next, we use Lemma A.2 to bound $\sum_{t=0}^{T-1} V_t$. We have,

$$\sum_{t=0}^{T-1} V_t \leq \sum_{t=0}^{T-1} H_{k(t)+1} \sum_{j=\tau_{k(t)}}^{t-1} \eta^2 (12Le_j + 6\bar{\sigma}^2)$$

$$\leq 12L\eta^2 \sum_{j=0}^{T-2} e_j \sum_{t=j+1}^{\tau_{k(j)+1}-1} H_{k(t)+1} + 6\bar{\sigma}^2\eta^2 \sum_{t=0}^{T-1} H_{k(t)+1}(t - \tau_{k(t)})$$

$$\leq 12L\eta^2 \sum_{j=0}^{T-2} e_j H_{k(j)+1}^2 + 6\bar{\sigma}^2\eta^2 \sum_{t=0}^{T-1} H_{k(t)+1}^2$$

$$\leq \frac{1}{4L} \sum_{t=0}^{T-2} e_t + 6\bar{\sigma}^2\eta^2 \sum_{i=0}^{R} H_i^3, \tag{A5}$$

where in the third inequality we have used the fact that $k(t) = k(j)$ for any $t \in [j + 1, \tau_{k(j)+1} - 1]$, and $t - \tau_{k(t)} \leq H_{k(t)+1}$, and in the last inequality we have used $H_i \leq \frac{1}{7L\eta}, \forall i$. Now, we can write

$$\frac{1}{2} \sum_{t=0}^{T-1} e_t \leq \frac{r_0 - r_T}{\eta} + \frac{3\bar{\sigma}^2\eta T}{n} + 12L\bar{\sigma}^2\eta^2 \sum_{i=0}^{R} H_i^3.$$

Dividing both sides of the above inequality by T and using the choice of $\eta = c\sqrt{\frac{n}{T}}$, we obtain

$$\frac{1}{T} \sum_{t=0}^{T-1} e_t \leq \frac{2r_0 + 6c^2\bar{\sigma}^2}{c\sqrt{nT}} + \frac{24L\bar{\sigma}^2 c^2 n}{T^2} \sum_{i=1}^{R} H_i^3.$$

∎

A.3   Proof of Theorem 3.3

To prove Theorem 3.3, we first establish an analogous descent lemma and consensus error lemma for the case of nonconvex local functions. The proofs of these lemmas are given in Appendix A.4.

Lemma A.3 (Decent lemma, non-convex): *Assume that $\nabla F_i(\mathbf{x}, \xi_i)$ is an unbiased stochastic gradient of $f_i(\mathbf{x})$ with variance bounded by σ^2. We have*

$$e_{t+1} \leq e_t - \frac{\eta_t}{4} h_t + \frac{L\sigma^2}{2n} \eta_t^2 + L^2 \eta_t V_t.$$

Lemma A.4 (Consensus error lemma, non-convex): *Let Assumption 2.3 hold. Moreover, assume that $\nabla F_i(\mathbf{x}, \xi_i)$ is an unbiased stochastic gradient of $f_i(\mathbf{x})$ with variance bounded by σ^2. For any t, define $k(t)$ be the index such that $\tau_{k(t)} \leq t < \tau_{k(t)+1}$. We have*

$$V_t \leq H_{k(t)+1} \sum_{j=\tau_{k(t)}}^{t-1} 6\eta_j^2 (B^2 h_j + \sigma^2 + G^2).$$

Proof of Theorem 3.3: Let us choose $\eta_t = \eta$, for some η to be specified later. By summing Lemma A.3 over $t = 0, \ldots, T - 1$, we get

$$\frac{\eta}{4} \sum_{t=0}^{T-1} h_t \leq e_0 - e_T + \frac{L\sigma^2\eta^2 T}{2n} + L^2\eta \sum_{t=0}^{T-1} V_t. \tag{A6}$$

Next, we use Lemma A.4 to bound $\sum_{t=0}^{T-1} V_t$. Using the same idea as in deriving expression (A5) in the proof of Theorem 3.2, we can get

$$\sum_{t=0}^{T-1} V_t \leq \sum_{t=0}^{T-1} H_{k(t)+1} \sum_{j=\tau_{k(t)}}^{t-1} 6\eta^2(B^2 h_j + \sigma^2 + G^2)$$

$$\leq 6B^2\eta^2 \sum_{j=0}^{T-2} e_j \sum_{t=j+1}^{\tau_{k(j)+1}-1} H_{k(t)+1} + 6(\sigma^2 + G^2)\eta^2 \sum_{t=0}^{T-1} H_{k(t)+1}(t - \tau_{k(t)})$$

$$\leq 6B^2\eta^2 \sum_{j=0}^{T-2} e_j H_{k(j)+1}^2 + 6(\sigma^2 + G^2)\eta^2 \sum_{t=0}^{T-1} H_{k(t)+1}^2$$

$$\leq \frac{1}{8L^2} \sum_{t=0}^{T-2} h_t + 6(\sigma^2 + G^2)\eta^2 \sum_{i=0}^{R} H_i^3,$$

where in the last inequality we have used $H_i \leq \frac{1}{7BLn}, \forall i$. Now, we can write

$$\frac{1}{8} \sum_{t=0}^{T-1} h_t \leq \frac{e_0 - e_T}{\eta} + \frac{L\sigma^2 \eta T}{2n} + 6L^2(\sigma^2 + G^2)\eta^2 \sum_{i=0}^{R} H_i^3.$$

Substituting $\eta_t = c\sqrt{\frac{n}{T}}$ into the above inequality and dividing both sides by T we get the desired bound. ∎

A.4 Proof of lemmas

Proof of Lemma A.1: Consider the filtration $\{\mathcal{F}^t\}_{t=1}^{\infty}$ adapted to the history of random variables $\{\xi_i^{(t)}\}$, i.e.

$$\mathcal{F}^t = \{\xi_i^{(k)} | i \in [n], 0 \leq k \leq t - 1\}, \tag{A7}$$

and note that $\mathbf{g}_i^{t-1} = \nabla F_i(\mathbf{x}_i^{t-1}, \xi_i^{(t-1)})$ and $\mathbf{x}_i^t$ are $\mathcal{F}^t$-measurable, but $\mathbf{g}_i^t$ is not. Using the definition of $r_t = \mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2$ and $\bar{\mathbf{g}}^{(t)} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i^t$, we have

$$r_{t+1} = \mathbb{E}[\|\bar{\mathbf{x}}^{(t+1)} - \mathbf{x}^*\|^2]$$

$$= \mathbb{E}[\|\bar{\mathbf{x}}^{(t)} - \eta_t \bar{\mathbf{g}}^{(t)} - \mathbf{x}^*\|^2]$$

$$= \mathbb{E}\left[\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^* - \frac{\eta_t}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t) - \frac{\eta_t}{n}\sum_{i=1}^{n}(\bar{\mathbf{g}}^{(t)} - \nabla f_i(\mathbf{x}_i^t))\right\|^2\right]$$

$$= \mathbb{E}\left[\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^* - \frac{\eta_t}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t)\right\|^2\right] \eta_t^2 \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}(\bar{\mathbf{g}}^{(t)} - \nabla f_i(\mathbf{x}_i^t))\right\|^2\right]$$

$$- \frac{2\eta_t}{n} \mathbb{E}_{\mathcal{F}^t}\left[\left\langle \bar{\mathbf{x}}^{(t)} - \mathbf{x}^* - \frac{\eta_t}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^t), \mathbb{E}\left[\sum_{i=1}^{n}(\bar{\mathbf{g}}^{(t)} - \nabla f_i(\mathbf{x}_i^t)) \mid \mathcal{F}^t\right]\right\rangle\right],$$

where the last term is obtained by first conditioning on $\mathcal{F}^t$ and then taking expectation with respect to $\mathcal{F}^t$. However, the last inner product in the above expression is zero, we have $\mathbb{E}[\mathbf{g}_i^t | \mathcal{F}^t] =$

$\nabla f_i(\mathbf{x}_i^t), \forall i$, and thus $\mathbb{E}[\sum_{i=1}^n (\bar{\mathbf{g}}^{(t)} - \nabla f_i(\mathbf{x}_i^t))|\mathcal{F}^t] = 0$. Therefore, we have

$$r_{t+1} = \mathbb{E}\left[\left\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^* - \frac{\eta_t}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t)\right\|^2\right] + \eta_t^2 \mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n (\bar{\mathbf{g}}^{(t)} - \nabla f_i(\mathbf{x}_i^t))\right\|^2\right]. \tag{A8}$$

We can bound the second term in Equation (A8) using [11, Proposition 5] as the following:

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n (\bar{\mathbf{g}}^{(t)} - \nabla f_i(\mathbf{x}_i^t))\right\|^2\right] \leq \frac{3L^2}{n^2}\sum_{i=1}^n \mathbb{E}\|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 + \frac{6L}{n}\left(\mathbb{E}[f(\bar{\mathbf{x}})] - f(\mathbf{x}^\star)\right) + \frac{3\bar{\sigma}^2}{n}$$

$$= \frac{3L^2}{n^2}V_t + \frac{6L}{n}e_t + \frac{3\bar{\sigma}^2}{n}.$$

In order to bound the first term in (A8), we can write

$$\mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^* - \frac{\eta_t}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t)\|^2$$

$$= \mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 + \frac{\eta_t^2}{n^2}\mathbb{E}\left\|\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t)\right\|^2 - \frac{2\eta_t}{n}\mathbb{E}\langle \bar{\mathbf{x}}^{(t)} - \mathbf{x}^*, \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t)\rangle. \tag{A9}$$

To bound $\mathbb{E}\|\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t)\|^2$ in (A9), we can write

$$\mathbb{E}\|\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t)\|^2 \leq 2\mathbb{E}\left\|\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t) - \sum_{i=1}^n \nabla f_i(\bar{\mathbf{x}}^t)\right\|^2 + 2\mathbb{E}\left\|\sum_{i=1}^n \nabla f_i(\bar{\mathbf{x}}^t) - \sum_{i=1}^n \nabla f_i(\mathbf{x}_*^t)\right\|^2$$

$$\leq 2n\mathbb{E}\sum_{i=1}^n \|\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\bar{\mathbf{x}}^t)\|^2 + 2n\mathbb{E}\sum_{i=1}^n \|\nabla f_i(\bar{\mathbf{x}}^t) - \nabla f_i(\mathbf{x}_*^t)\|^2$$

$$\leq 2nL^2\mathbb{E}\sum_{i=1}^n \|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\|^2 + 4n^2L(\mathbb{E}f(\bar{\mathbf{x}}^{(t)}) - f(\mathbf{x}^*))$$

$$= 2n^2L^2V_t + 4n^2Le_t,$$

where in the last inequality, we have used Lemma A.8. To bound $\mathbb{E}\langle \bar{\mathbf{x}}^{(t)} - \mathbf{x}^*, \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t)\rangle$ in (A9), we have

$$\mathbb{E}\langle \bar{\mathbf{x}}^{(t)} - \mathbf{x}^*, \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t)\rangle$$

$$= \sum_{i=1}^n \mathbb{E}\langle \bar{\mathbf{x}}^{(t)} - \mathbf{x}^*, \nabla f_i(\mathbf{x}_i^t)\rangle$$

$$= \sum_{i=1}^n \mathbb{E}\langle \bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}, \nabla f_i(\mathbf{x}_i^t)\rangle + \sum_{i=1}^n \mathbb{E}\langle \mathbf{x}_i^{(t)} - \mathbf{x}^*, \nabla f_i(\mathbf{x}_i^t)\rangle$$

$$\geq \sum_{i=1}^n \mathbb{E}\left[f_i(\bar{\mathbf{x}}^{(t)}) - f_i(\mathbf{x}_i^{(t)}) - \frac{L}{2}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\|^2\right] + \sum_{i=1}^n \mathbb{E}\left[f_i(\mathbf{x}_i^{(t)}) - f_i(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x}_i^{(t)} - \mathbf{x}^*\|^2\right]$$

$$= \sum_{i=1}^n \mathbb{E}[f_i(\bar{\mathbf{x}}^{(t)}) - f_i(\mathbf{x}^*)] + \frac{\mu n}{2}\mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\|^2 - \left(\frac{L-\mu}{2}\right)\sum_{i=1}^n \mathbb{E}\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^{(t)}\|^2$$

$$= ne_t + \frac{\mu}{2}nr_t - \left(\frac{L-\mu}{2}\right)nV_t,$$

where the first inequality follows from the strong convexity assumption and Lemma A.8. Moreover, in the last equality, we used Lemma A.7. Finally, if we put the above bounds into (A9) and substitute the result into (A8), we obtain

$$r_{t+1} \leq r_t + \frac{\eta_t^2}{n^2}(2n^2 L^2 V_t + 4n^2 L e_t) - \frac{2\eta_t}{n}\left(ne_t + \frac{\mu}{2}nr_t - \frac{L-\mu}{2}nV_t\right)$$

$$+ \eta_t^2\left(\frac{3L^2}{n}V_t + \frac{6L}{n}e_t + \frac{3\bar{\sigma}^2}{n}\right)$$

$$= (1 - \mu\eta_t)r_t - \left(2\eta_t - 4\eta_t^2 L - \frac{6L\eta_t^2}{n}\right)e_t + \left(2L^2\eta_t^2 + (L-\mu)\eta_t + \frac{3L^2\eta_t^2}{n}\right)V_t + \frac{3\bar{\sigma}^2\eta_t^2}{n}$$

$$\leq (1 - \mu\eta_t)r_t - \eta_t e_t + 2L\eta_t V_t + \frac{3\bar{\sigma}^2\eta_t^2}{n},$$

where the last inequality holds because $\eta_t \leq \frac{1}{10L}$. ∎

In order to prove the consensus error lemma (Lemma A.2), we first state and prove the following auxiliary lemma, which bounds the expected sum of the gradient norms across all agents.

Lemma A.5: *For strongly convex L-smooth local functions, we have*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\|^2 \leq 3L^2 V_t + 6L e_t + 3\bar{\sigma}^2.$$

Proof: Starting from the left-hand side, we can write

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left\|\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) - \nabla F_i(\bar{\mathbf{x}}^{(t)}, \xi_i^{(t)}) + \nabla F_i(\bar{\mathbf{x}}^{(t)}, \xi_i^{(t)}) - \nabla F_i(\mathbf{x}^*, \xi_i^{(t)}) + \nabla F_i(\mathbf{x}^*, \xi_i^{(t)})\right\|^2$$

$$\leq \frac{3}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla F_i(\bar{\mathbf{x}}^{(t)}, \xi_i^{(t)}) - \nabla F_i(\mathbf{x}^*, \xi_i^{(t)})\|^2$$

$$+ \frac{3}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) - \nabla F_i(\bar{\mathbf{x}}^{(t)}, \xi_i^{(t)})\|^2 + 3\bar{\sigma}^2$$

$$\leq \frac{3}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla F_i(\bar{\mathbf{x}}^{(t)}, \xi_i^{(t)}) - \nabla F_i(\mathbf{x}^*, \xi_i^{(t)})\|^2 + 3L^2 V_t + 3\bar{\sigma}^2, \tag{A10}$$

where the first inequality uses Definition 2.1, and the second inequality uses L-smooth assumption. We have

$$\sum_{i=1}^{n}\mathbb{E}\|\nabla F_i(\bar{\mathbf{x}}^{(t)}, \xi_i^{(t)}) - \nabla F_i(\mathbf{x}^*, \xi_i^{(t)})\|^2$$

$$\leq 2L\sum_{i=1}^{n}\mathbb{E}\left[F_i(\bar{\mathbf{x}}^{(t)}, \xi_i^{(t)}) - F_i(\mathbf{x}^*, \xi_i^{(t)})\right] - 2L\sum_{i=1}^{n}\mathbb{E}\langle\nabla F_i(\mathbf{x}^*, \xi_i^{(t)}), \bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\rangle$$

$$= 2L\sum_{i=1}^{n}\mathbb{E}\left[f_i(\bar{\mathbf{x}}^{(t)}) - f_i(\mathbf{x}^*)\right] - 2L\sum_{i=1}^{n}\mathbb{E}\langle\nabla f_i(\mathbf{x}^*), \bar{\mathbf{x}}^{(t)} - \mathbf{x}^*\rangle$$

$$= 2nL\mathbb{E}\big[f(\bar{\mathbf{x}}^{(t)}) - f(\mathbf{x}^*) - \langle \nabla f(\mathbf{x}^*), \bar{\mathbf{x}}^{(t)} - \mathbf{x}^* \rangle\big]$$

$$= 2nL\mathbb{E}\big[f(\bar{\mathbf{x}}^{(t)}) - f(\mathbf{x}^*)\big] = 2nLe_t,$$

where the first inequality uses Lemma A.8, and the last equality holds because $\mathbf{x}^*$ is the global minimum of f, and hence $\nabla f(\mathbf{x}^*) = 0$. Substituting the above relation into (A10) completes the proof. ∎

Proof of Lemma A.2: As $\tau_{k(t)} \le t < \tau_{k(t)+1}$, agents do not communicate during the time interval $(\tau_{k(t)}, t]$ and only perform local gradient steps. Thus, we have

$$\mathbf{x}_i^{(t)} = \mathbf{x}_i^{(\tau_{k(t)})} - \sum_{j=\tau_{k(t)}}^{t-1} \eta_j \nabla F_i(\mathbf{x}_i^{(j)}, \xi_i^{(j)}).$$

Moreover, at the communication time $\tau_{k(t)}$, all the agents update their local vectors to the same average vector received from the center node. Therefore, $\bar{\mathbf{x}}^{\tau_{k(t)}} = \mathbf{x}_i^{(\tau_{k(t)})} \forall i$, and we have

$$\bar{\mathbf{x}}^t = \bar{\mathbf{x}}^{\tau_{k(t)}} - \sum_{j=\tau_{k(t)}}^{t-1} \eta_j \bar{\mathbf{g}}^{(j)} = \mathbf{x}_i^{(\tau_{k(t)})} - \sum_{j=\tau_{k(t)}}^{t-1} \eta_j \bar{\mathbf{g}}^{(j)}.$$

If we substitute the above relations into V_t, we get

$$nV_t = \mathbb{E} \sum_{i=1}^{n} \|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^t\|^2$$

$$= \sum_{i=1}^{n} \mathbb{E} \left\| \sum_{j=\tau_{k(t)}}^{t-1} \eta_j \nabla F_i(\mathbf{x}_i^{(j)}, \xi_i^{(j)}) - \sum_{j=\tau_{k(t)}}^{t-1} \eta_j \bar{\mathbf{g}}^{(j)} \right\|^2$$

$$\le \sum_{i=1}^{n} \mathbb{E} \left\| \sum_{j=\tau_{k(t)}}^{t-1} \eta_j \nabla F_i(\mathbf{x}_i^{(j)}, \xi_i^{(j)}) \right\|^2$$

$$\le (t - \tau_{k(t)}) \sum_{i=1}^{n} \sum_{j=\tau_{k(t)}}^{t-1} \eta_j^2 \mathbb{E} \|\nabla F_i(\mathbf{x}_i^{(j)}, \xi_i^{(j)})\|^2$$

$$\le n(t - \tau_{k(t)}) \sum_{j=\tau_{k(t)}}^{t-1} \eta_j^2 (3L^2 V_j + 6Le_j + 3\bar{\sigma}^2)$$

$$\le nH_{k(t)+1} \sum_{j=\tau_{k(t)}}^{t-1} \eta_j^2 (3L^2 V_j + 6Le_j + 3\bar{\sigma}^2).$$

where the first inequality uses Lemma A.7, and the third inequality follows from Lemma A.5. Moreover, by our choice of step-size $\eta_j \le \frac{1}{3LH_{k(t)+1}}, \forall j \ge \tau_{k(t)}$. Thus, for any time instance in the interval $[\tau_{k(t)}, \tau_{k(t)+1})$, we have shown that

$$V_t \le \frac{1}{3H_{k(t)+1}} \sum_{j=\tau_{k(t)}}^{t-1} V_j + H_{k(t)+1} \sum_{j=\tau_{k(t)}}^{t-1} \eta_j^2 (6Le_j + 3\bar{\sigma}^2).$$

By recursively unrolling $V_j, j = t - 1, \dots, \tau_{k(t)} + 1$, and noting that $V_{\tau_{k(t)}} = 0$, we obtain

$$V_t \le \frac{1}{3H_{k(t)+1}} V_{t-1} + \frac{1}{3H_{k(t)+1}} \sum_{j=\tau_{k(t)}}^{t-2} V_j + H_{k(t)+1} \sum_{j=\tau_{k(t)}}^{t-1} \eta_j^2 (6Le_j + 3\bar{\sigma}^2)$$

$$\overset{\text{(Unroll } V_{t-1})}{\leq} \left(\left(\frac{1}{3H_{k(t)+1}} \right)^2 + \frac{1}{3H_{k(t)+1}} \right) \sum_{j=\tau_{k(t)}}^{t-2} V_j$$

$$+ \left(1 + \frac{1}{3H_{k(t)+1}} \right) H_{k(t)+1} \sum_{j=\tau_{k(t)}}^{t-1} \eta_j^2 (6Le_j + 3\bar{\sigma}^2)$$

$$\overset{\text{(Unroll } V_{t-2})}{\leq} \left(1 + \frac{1}{3H_{k(t)+1}} \right) \cdot \frac{1}{3H_{k(t)+1}} \sum_{j=\tau_{k(t)}}^{t-3} V_j$$

$$+ \left(1 + \frac{1}{3H_{k(t)+1}} \right) H_{k(t)+1} \sum_{j=\tau_{k(t)}}^{t-1} \eta_j^2 (6Le_j + 3\bar{\sigma}^2)$$

$$+ \left(1 + \frac{1}{3H_{k(t)+1}} \right) \cdot \left(\frac{1}{3H_{k(t)+1}} \right)^2 \sum_{j=\tau_{k(t)}}^{t-3} V_j$$

$$+ \left(1 + \frac{1}{3H_{k(t)+1}} \right) \cdot \left(\frac{1}{3H_{k(t)+1}} \right) H_{k(t)+1} \sum_{j=\tau_{k(t)}}^{t-1} \eta_j^2 (6Le_j + 3\bar{\sigma}^2)$$

$$= \left(1 + \frac{1}{3H_{k(t)+1}} \right)^2 \cdot \frac{1}{3H_{k(t)+1}} \sum_{j=\tau_{k(t)}}^{t-3} V_j + \left(1 + \frac{1}{3H_{k(t)+1}} \right)^2 H_{k(t)+1} \sum_{j=\tau_{k(t)}}^{t-1} \eta_j^2 (6Le_j + 3\bar{\sigma}^2)$$

$$\overset{\text{(Unroll } V_{t-3}, \dots, V_{\tau_{k(t)}+1})}{\leq} \left(1 + \frac{1}{3H_{k(t)+1}} \right)^{t-\tau_{k(t)}} \cdot \frac{1}{3H_{k(t)+1}} V_{\tau_{k(t)}}$$

$$+ \left(1 + \frac{1}{3H_{k(t)+1}} \right)^{t-\tau_{k(t)}} H_{k(t)+1} \sum_{j=\tau_{k(t)}}^{t-1} \eta_j^2 (6Le_j + 3\bar{\sigma}^2)$$

$$\overset{(V_{\tau_{k(t)}}=0)}{\leq} \left(1 + \frac{1}{3H_{k(t)+1}} \right)^{H_{k(t)+1}} H_{k(t)+1} \sum_{j=\tau_{k(t)}}^{t-1} \eta_j^2 (6Le_j + 3\bar{\sigma}^2).$$

Finally, by replacing $(1 + \frac{1}{3H_{k(t)+1}})^{H_{k(t)+1}} \leq 2$ into the above relation we obtain the desired bound. ∎

Proof of Lemma A.3: Using Taylor expansion and the L-smoothness assumption, we can write

$$e_{t+1} = \mathbb{E}f(\bar{\mathbf{x}}^{(t+1)}) - f(\mathbf{x}^*) \tag{A11}$$

$$= \mathbb{E}f\big(\bar{\mathbf{x}}^{(t)} - \eta_t \bar{\mathbf{g}}^{(t)}\big) - f(\mathbf{x}^*) \tag{A12}$$

$$\leq \big(\mathbb{E}f(\bar{\mathbf{x}}^{(t)}) - f(\mathbf{x}^*)\big) + \frac{L}{2}\eta_t^2 \mathbb{E}\|\bar{\mathbf{g}}^{(t)}\|^2 - \eta_t \mathbb{E}\langle \nabla f(\bar{\mathbf{x}}^{(t)}), \bar{\mathbf{g}}^{(t)}\rangle, \tag{A13}$$

where we recall that $\bar{\mathbf{g}}^{(t)} = \frac{1}{n}\sum_{i=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})$. Next, we bound the second and third terms in (A13). To bound the third term, using Assumption 2.3, we have

$$\mathbb{E}\langle \nabla f(\bar{\mathbf{x}}^{(t)}), \bar{\mathbf{g}}^{(t)}\rangle$$

$$= \mathbb{E}\left\langle \nabla f(\bar{\mathbf{x}}^{(t)}), \frac{1}{n}\sum_{i=1}^n \nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\rangle$$

$$= \mathbb{E}\left[\mathbb{E}_{\{\xi_i^{(t)}\}}\left[\left\langle \nabla f(\bar{\mathbf{x}}^{(t)}), \frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\rangle | \mathcal{F}^t\right]\right]$$

$$= \mathbb{E}\left[\left\langle \nabla f(\bar{\mathbf{x}}^{(t)}), \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\xi_i^{(t)}}[\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})|\mathcal{F}^t]\right\rangle\right]$$

$$= \mathbb{E}\left\langle \nabla f(\bar{\mathbf{x}}^{(t)}), \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})\right\rangle$$

$$= \mathbb{E}\left\langle \nabla f(\bar{\mathbf{x}}^{(t)}), \frac{1}{n}\sum_{i=1}^{n}(\nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f_i(\bar{\mathbf{x}}^{(t)}))\right\rangle + \mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|^2$$

$$\geq \frac{1}{2}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|^2 - \frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f_i(\bar{\mathbf{x}}^{(t)})\|^2$$

$$\geq \frac{1}{2}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|^2 - \frac{L^2}{2n}\sum_{i=1}^{n}\mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\|^2$$

$$= \frac{1}{2}h_t - \frac{L^2}{2}V_t, \tag{A14}$$

where the last inequality is by L-smoothness assumption. To bound the second term in (A13), we have

$$\mathbb{E}\|\bar{\mathbf{g}}^{(t)}\|^2 = \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)})\right\|^2$$

$$= \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)}))\right\|^2 + \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\mathbf{x}_i^{(t)})\right\|^2$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\|\nabla F_i(\mathbf{x}_i^{(t)}, \xi_i^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})\|^2 + \mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^{n}(\nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f_i(\bar{\mathbf{x}}^{(t)})) + \nabla f(\bar{\mathbf{x}}^{(t)})\right\|^2.$$

Thus, using the bounded noise Assumption 2.3, we get

$$\mathbb{E}\|\bar{\mathbf{g}}^{(t)}\|^2 \leq \frac{\sigma^2}{n} + 2\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|^2 + \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f_i(\bar{\mathbf{x}}^{(t)})\|^2$$

$$\leq \frac{\sigma^2}{n} + 2\mathbb{E}\|\nabla f(\bar{\mathbf{x}}^{(t)})\|^2 + \frac{2L^2}{n}\sum_{i=1}^{n}\mathbb{E}\|\bar{\mathbf{x}}^{(t)} - \mathbf{x}_i^{(t)}\|^2$$

$$= \frac{\sigma^2}{n} + 2h_t + 2L^2V_t, \tag{A15}$$

where the second inequality holds by the L-smooth assumption. Finally, by substituting (A14) and (A15) into (A13), we obtain

$$e_{t+1} \leq e_t - \eta_t\left(\frac{1}{2}h_t - \frac{L^2}{2}V_t\right) + \frac{L\eta_t^2}{2}\left(\frac{\sigma^2}{n} + 2h_t + 2L^2V_t\right)$$

$$\leq e_t - \frac{\eta_t}{4}h_t + \frac{L\sigma^2}{2n}\eta_t^2 + L^2\eta_t V_t,$$

where the last inequality holds because $\eta_t \leq \frac{1}{4L}$.   ∎

We first establish the following technical lemma to prove the consensus descent lemma for the nonconvex functions.

Lemma A.6: *Let Assumptions 2.3 hold. Then,*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)})\|^2 \leq 3(L^2 V_t + B^2 h_t + \sigma^2 + G^2).$$

Proof: We can write,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)})\|^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big\|\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)}) + \nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f_i(\bar{\mathbf{x}}^{(t)}) + \nabla f_i(\bar{\mathbf{x}}^{(t)})\big\|^2$$

$$\leq \frac{3}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla F_i(\mathbf{x}_i^{(t)},\xi_i^{(t)}) - \nabla f_i(\mathbf{x}_i^{(t)})\|^2 + \frac{3}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(\mathbf{x}_i^{(t)}) - \nabla f_i(\bar{\mathbf{x}}^{(t)})\|^2$$

$$+ \frac{3}{n}\sum_{i=1}^{n}\mathbb{E}\|\nabla f_i(\bar{\mathbf{x}}^{(t)})\|^2$$

$$\leq 3\sigma^2 + 3L^2 V_t + 3(G^2 + B^2 h_t).$$

where the last inequality is obtained using the L-smoothness assumption and Assumption 2.3. ∎

Proof of Lemma A.4: By following the same steps as in the proof of Lemma A.2, we can write

$$nV_t = \mathbb{E}\sum_{i=1}^{n}\|\mathbf{x}_i^{(t)} - \bar{\mathbf{x}}^t\|^2$$

$$\leq (t - \tau_{k(t)})\sum_{i=1}^{n}\sum_{j=\tau_{k(t)}}^{t-1}\eta_j^2\mathbb{E}\|\nabla F_i(\mathbf{x}_i^{(j)},\xi_i^{(j)})\|^2$$

$$\leq n(t - \tau_{k(t)})\sum_{j=\tau_{k(t)}}^{t-1} 3\eta_j^2\big(L^2 V_j + B^2 h_j + \sigma^2 + G^2\big)$$

$$\leq nH_{k(t)+1}\sum_{j=\tau_{k(t)}}^{t-1} 3\eta_j^2\big(L^2 V_j + B^2 h_j + \sigma^2 + G^2\big)$$

$$\leq nH_{k(t)+1}\sum_{j=\tau_{k(t)}}^{t-1} 3\eta_j^2\big(L^2 V_j + B^2 h_j + \sigma^2 + G^2\big),$$

where in the second inequality, we have used Lemma A.6. Since by the choice of step size we may assume $\eta_j \leq \frac{1}{3LH_{k(t)+1}}$, for any time instance in the time interval $[\tau_{k(t)}, \tau_{k(t)+1})$, we have shown that

$$V_t \leq \frac{1}{3H_{k(t)+1}}\sum_{j=\tau_{k(t)}}^{t-1} V_j + H_{k(t)+1}\sum_{j=\tau_{k(t)}}^{t-1} 3\eta_j^2\big(B^2 h_j + \sigma^2 + G^2\big).$$

Finally, if we recursively unroll $V_j, j = \tau_{k(t)}, \ldots, t-1$ as in the proof of Lemma A.2, we obtain

$$V_t \leq (1 + \frac{1}{3H_{k(t)}+1})^{H_{k(t)}+1} \times H_{k(t)}+1 \sum_{j=\tau_{k(t)}}^{t-1} 3\eta_j^2 (B^2 h_j + \sigma^2 + G^2)$$

$$\leq H_{k(t)}+1 \sum_{j=\tau_{k(t)}}^{t-1} 6\eta_j^2 (B^2 h_j + \sigma^2 + G^2).$$

∎

Lemma A.7: *Let* $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$. *Then, for any* $\mathbf{x}' \in \mathbb{R}^d$, $\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{x}'\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 + n\|\bar{\mathbf{x}} - \mathbf{x}'\|^2$. *In particular,* $\sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 \leq \sum_{i=1}^{n} \|\mathbf{x}_i\|^2$.

Proof: We have,

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{x}'\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \mathbf{x}'\|^2$$

$$= \sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 + n\|\bar{\mathbf{x}} - \mathbf{x}'\|^2 - \sum_{i=1}^{n} \langle \mathbf{x}_i - \bar{\mathbf{x}}, \bar{\mathbf{x}} - \mathbf{x}' \rangle$$

$$= \sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2 + n\|\bar{\mathbf{x}} - \mathbf{x}'\|^2.$$

The second inequality holds by choosing $\mathbf{x}' = 0$. ∎

Lemma A.8: *Let f be a L-smooth convex function. Then, for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, *we have*

$$f(\mathbf{x}) - f(\mathbf{y}) + \frac{L}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 \geq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle,$$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq 2L(f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle).$$

Proof: The first inequality is an immediate consequence of the *L*-smoothness property. To show the second inequality, let us define $\mathbf{z} = \mathbf{y} - \frac{1}{L}(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))$. Then,

$$f(\mathbf{z}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle,$$

$$f(\mathbf{z}) \leq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{z} - \mathbf{y} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{z}\|^2.$$

Therefore,

$$f(\mathbf{x}) + \left\langle \nabla f(\mathbf{x}), \mathbf{y} - \frac{1}{L}(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})) - \mathbf{x} \right\rangle$$

$$\leq f(\mathbf{y}) - \left\langle \nabla f(\mathbf{y}), \frac{1}{L}(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})) \right\rangle$$

$$+ \frac{1}{2L} \|(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))\|^2.$$

Rearranging the terms completes the proof. ∎

A.5 Choice of β in Proposition 3.1

Here we prove that in Proposition 3.1, we can choose $\beta = a\lceil\frac{24L}{\mu}\rceil^s \cdot \frac{12L}{\mu} + 1$ in order to satisfy condition 1) in Corollary 3.1, i.e. $H_i \le \frac{\mu(\beta+\sum_{j=1}^{i-1} H_j)}{12L}$, $\forall i$. Since $a = \mathcal{O}(n^{-\frac{s+1}{2}} T^{\frac{1-s}{2}})$, the overall convergence rate is still $\mathcal{O}(\frac{1}{nT})$.

Proof: Let $k = \lceil\frac{24L}{\mu}\rceil$, then $\beta \ge H_k \cdot \frac{12L}{\mu} + 1$. For all $i \le k$, we have

$$H_i \le H_k < \frac{\mu\beta}{12L} < \frac{\mu(\beta + \sum_{j=1}^{i-1} H_j)}{12L}.$$

For all $k \le i \le T$, we would prove by induction that $a \cdot i^s \le \frac{\mu(\beta+\sum_{j=1}^{i-1} H_j)}{12L}$, thus concluding the proof.

In fact, for the base case $i = k$, we have

$$a \cdot \left\lceil\frac{24L}{\mu}\right\rceil^s < \frac{\mu\beta}{12L} < \frac{\mu(\beta + \sum_{j=1}^{k-1} H_j)}{12L}.$$

For inductive step, assume for some $k \le i \le T$, we have $a \cdot i^s \le \frac{\mu(\beta+\sum_{j=1}^{i-1} H_j)}{12L}$, then

$$a \cdot (i+1)^s \le \frac{\mu(\beta + \sum_{j=1}^{i} H_j)}{12L}$$

$$\Leftarrow a \cdot (i+1)^s - a \cdot i^s \le \frac{\mu H_i}{12L}$$

$$\Leftarrow a \cdot (i+1)^s - a \cdot i^s \le \frac{\mu a \cdot i^s}{24L}$$

$$\Leftarrow \left(\frac{i+1}{i}\right)^s \le \frac{\mu}{24L} + 1$$

$$\Leftarrow i \ge \frac{1}{(\frac{\mu}{24L} + 1)^{\frac{1}{s}} - 1}$$

$$\Leftarrow i \ge \frac{1}{1 + \frac{\mu}{12Ls} - 1} \Leftarrow i \ge \lceil\frac{24L}{\mu}\rceil = k.$$

By induction we conclude that for all $k \le i \le T$, we also have $H_i \le a \cdot i^s \le \frac{\mu(\beta+\sum_{j=1}^{i-1} H_j)}{12L}$. ∎