



TUNING-FREE DECENTRALIZED NONCONVEX STOCHASTIC OPTIMIZATION

Jiaxiang Li, Xuxing Chen, Shiqian Ma* and Mingyi Hong

Dedicated to Professor R. T. Rockafellar on the occasion of his 90th birthday

Abstract: Existing decentralized algorithms usually require knowledge of problem parameters for updating local iterates. For example, the hyperparameters (such as learning rate) usually require the knowledge of Lipschitz constant of the global gradient or topological information of the communication networks, which are usually not accessible in practice. In this paper, we propose D-NASA, the first algorithm for decentralized nonconvex stochastic optimization that requires no prior knowledge of any problem parameters. We show that D-NASA has the optimal rate of convergence for nonconvex objectives under very mild conditions and enjoys the linear-speedup effect, i.e. the computation becomes faster as the number of nodes in the system increases. Extensive numerical experiments are conducted to support our findings.

Key words: *nonconvex optimization, decentralized optimization, tuning-free methods*

Mathematics Subject Classification: *90C15*

1 Introduction

Decentralized (distributed) optimization appears in many applications, such as machine learning [32, 33, 52], robotics [48], signal processing [19], and control systems [41, 59]. In machine learning, decentralized optimization arises naturally when the data is either stored in different physical locations, or split into different servers to boost training efficiency. Therefore the main concerns for decentralized algorithms are data privacy, algorithmic scalability and robustness. For example, starting from earlier works [32, 52], researchers seek to develop scalable decentralized algorithms for distributed training that are provably more efficient than centralized algorithms, usually reflected in an inverse dependency over the number of devices/nodes in their final convergence rate, known as *linear speedup*.

One obstacle of applying most of the developed decentralized algorithms in practice is that their hyperparameters (such as learning rate) usually depend on information of the problem in order to show a theoretical convergence, e.g. the Lipschitz constant of the global gradient, the spectral gap of the graph adjacency matrix or other topological information of the problem. Such information is usually hard to obtain due to either physical/privacy restrictions or computational constraints (e.g. due to excessive amount of data in machine

*Research supported in part by ONR grant N00014-24-1-2705, NSF grants DMS-2243650, CCF-2308597, CCF-2311275 and ECCS-2326591, and a startup fund from Rice University

learning applications), and tedious hyperparameter tuning is thus required. Nonetheless, in most of these works people demonstrate a decent performance in experiments of the proposed algorithms without strictly following the hyperparameter rules suggested in their theory. This gap between theory and application exists in centralized optimization problems and researchers have proposed different methods to mitigate it. However, this gap introduces more serious problems in the decentralized setting for several reasons: (i) In distributed settings, it is hard, if not impossible, for local devices to know the problem information of other devices, even if the network is fully connected [60]; (ii) The network architecture might be largely unknown for algorithmic design, especially when the data are distributed in different physical locations, thus it is difficult to compute network related constants such as eigenvalues of the graph Laplacian; (iii) The extra error introduced by the heterogeneity of the data distributions on each local nodes brings more challenges for convergence analysis [24, 52].

In this work, we close these gaps by designing problem-parameter-free algorithms, i.e., algorithms whose hyperparameters do not require problem information, for decentralized optimization. Specifically, consider the following stochastic decentralized optimization problem:

$$\min_x f(x) := \frac{1}{n} \sum_{i=1}^n f_i(x), \quad (1.1)$$

where each $f_i = \mathbb{E}_{\xi_i \sim \mathcal{D}_i} [F_i(x, \xi_i)]$ is stored on a local device/node/agent i , which is assumed to be L_i -Lipschitz smooth and possibly nonconvex, a standard assumption in the literature. Moreover, we assume that the Lipschitz constant is not available for the algorithmic design. Each local node i is only allowed to access the stochastic function $F_i(x, \xi_i)$ in the algorithm design. Note that for different node i the data distribution could be highly heterogeneous, i.e., each ξ_i follows completely different distributions $\mathcal{D}_i$. Also, each local agent is only connected to a limited amount of neighboring agents, forming an undirected connected graph, which is summarized by the doubly stochastic mixing matrix W (see Section 2). To be more specific, in this paper, “tuning-free” means that the hyperparameters of the algorithm (such as learning rate) do not depend on problems parameters such as L_i and W . The goal of this paper is thus to design such algorithms for solving (1.1).

The most straightforward method for decentralized optimization is decentralized stochastic gradient descent (D-SGD) where each local device runs stochastic gradient descent then communicates the update with their neighbors to form the next iterate. In [32], the authors provided a convergence analysis under the assumption of bounded heterogeneity, i.e., the gradient distributions across different devices are similar. To remove this assumption, another famous method is the decentralized gradient tracking algorithm (D-SGT, Algorithm 1, see [13, 25, 34, 42, 46, 47, 56]) which efficiently guarantees convergence without requiring bounded heterogeneity, also yields superior numerical performances. We thus first inspect the convergence of D-SGT algorithm under the tuning/parameter-free setting. Our analysis shows that D-SGT could converge when the hyperparameters are problem-parameter-free. Besides standard assumption of local Lipschitz smooth, however, this convergence result additionally requires the local functions to be *Lipschitz continuous* (i.e., having bounded gradient). To remove this restrictive assumption, we propose a new decentralized *normalized* averaged stochastic approximate gradient tracking (D-NASA, Algorithm 2) which enjoys tuning-free convergence without additional assumptions.

Our contributions are summarized as follows.

- **New analysis of D-SGT.** We investigate D-SGT (Algorithm 1) and point out that

one can use a learning rate that is tuning/parameter-free and still guarantee the convergence, at the expense of an additional assumption: local functions are Lipschitz continuous, i.e., bounded local function gradients. This is a rather strong requirement since it implies bounded heterogeneity among different nodes (see Section 3.1). The analysis also indicates that D-SGT can no longer achieve a linear speedup under this setting.

- **A new parameter-free algorithm.** We propose a fully tuning/parameter-free algorithm (D-NASA, Algorithm 2) based on certain normalization technique that does not require information of global Lipschitz constant or spectral gap of the topology of the problem. The convergence of D-NASA is guaranteed without any additional assumption. The convergence result matches the lower bound for nonconvex stochastic optimization and still enjoys the desired linear speedup.
- **Normalization controls consensus error.** D-NASA utilizes a novel control over the consensus error. Specifically, we notice that normalized update efficiently helps the control of the consensus error, and enables controlling the cumulative consensus error directly by stepsizes (see Section 3.2). This opens the door of adapting a wide class of normalization-based adaptive algorithms to the decentralized setting, and its fine-grained analysis is of independent interest.
- **Numerical evidences.** We conduct extensive numerical study to verify our findings. We observe linear speedup effect of D-NASA with stepsizes exactly predicted by our theory. We also show that D-NASA compares favorably with existing algorithms D-SGD, D-SGT and D-ASAGT in terms of convergence speed. We empirically demonstrate D-NASA does not require any parameter tuning for a wide range of Lipschitz smooth parameters and network topology. Without this technique, the stepsize tuning process can be time-consuming since the optimal choices of the hyperparameters vary drastically when datasets change.

Notation. We denote $\mathbf{X}^t := [x_1^t, \dots, x_n^t]$ which is the collection of local variables x_i^t at iteration t for $i = 1, \dots, n$ as column vectors. $\bar{x}^t := \frac{1}{n} \sum_{i=1}^n x_i^t$ is the average of all local variables. The same convention applies to $\mathbf{U}$, $\mathbf{V}$, $\mathbf{Z}$ and $\bar{u}^t$, $\bar{v}^t$, $\bar{z}^t$. Also denote by $\bar{\mathbf{X}}^t = \bar{x}^t \mathbf{1}^\top = \frac{1}{n} \mathbf{X}^t \mathbf{1} \mathbf{1}^\top$ the collection of average of local variables, where $\mathbf{1}$ is n -dimensional all one column vector. The same convention applies to $\bar{\mathbf{U}}$, $\bar{\mathbf{V}}$ and $\bar{\mathbf{Z}}$. We use $\|\cdot\|$ to represent the Euclidean vector norm and matrix Frobenius norm to simplify the notation. For matrix 2 norm (i.e., spectral norm) we use $\|\cdot\|_2$.

1.1 Related works

Decentralized optimization. While the study of decentralized optimization algorithms has a long history [49, 53, 57, 61], their distinctive advantages, such as robustness, scalability and privacy preserving, in comparison to centralized setting like [30], were not well understood both theoretically and empirically until the case study conducted by [32]. Despite its great success in characterizing the superiority of decentralized training over the centralized setting, the analysis therein relies on a bounded gradient heterogeneity assumption, which was later removed by follow-up works such as D² [52].

Motivated by the empirical success of decentralized training, another line of work focused on improving the convergence rates of decentralized algorithms. Vanilla decentralized gradient descent with a fixed stepsize is known to only converge to a neighborhood of the

Table 1: Comparison of D-NASA (Algorithm 2) with some widely-used decentralized stochastic nonconvex optimization algorithms: D-SGD [32], D² [52] and D-SGT [25]. ‘Other aspt’ refers to the additional assumptions required for theoretical convergence (Note that the parameters in the assumptions might not be available to the algorithm), where ‘Hetero’ stands for bounded heterogeneity, and all algorithms require the stochastic bounded variance and the deterministic gradients begin Lipschitz continuous; ‘Info Required’ refers to the problem parameters that the algorithm parameters (such as stepsizes) should depend on to achieve the sample complexity, where “smoothness” is the Lipschitz constant of the global gradient, and “variance” is the variance of the stochastic oracle; All algorithms in this table require $\mathcal{O}(n^{-1}\epsilon^{-4})$ oracle calls to achieve an ϵ -stationary point.

Algorithm	Other Aspt	Info Required
D-SGD	HETERO	SMOOTHNESS, VARIANCE
D²	NONE	SMOOTHNESS, NET-TOPOLOGY
D-SGT	NONE	SMOOTHNESS, NET-TOPOLOGY
D-NASA (OURS)	NONE	NONE

optimal solution even under the deterministic and strongly convex setting [61]. One important technique to mitigate this effect is gradient tracking, which was introduced in control community [13, 42, 47, 56] to improve the convergence rate in the deterministic setting. Later this method was revealed to be helpful to remove the bounded gradient heterogeneity assumption [25, 35, 46, 62] in convergence analysis. A more recent technique of moving-average updates (momentum) have been studied in both decentralized optimization and federated learning setting [9, 54] to further improve the rate of convergence.

It is worth noticing that the above works all require knowledge about the global problem to design their algorithms. Under the assumption that the local functions are Lipschitz continuous, NEXT [13] is able to achieve a tuning/parameter-free asymptotic convergence (in **deterministic** setting). We point out again that Lipschitz continuity of the objective functions is a strong assumption that implies boundedness of gradients and bounded heterogeneity (see Section 3.1).

Other interesting research topics in decentralized optimization include network topology [2, 24, 43], communication compression [23, 51], large-model training [16, 60], adaptive algorithms [8], etc.

Tuning-Free optimization. Tuning-Free, or frequently refer to as parameter-free optimization are the algorithms that require no/few information from the problem so that the algorithm converges without any tedious process of hyperparameter-tuning. For deterministic smooth optimization, one could show the convergence of gradient descent to either the optimal (convex) or the stationary point (nonconvex) when the stepsize η is smaller than $2/L$, where L is the Lipschitz smooth constant [44]. When problem parameters such as L are not available, one usually uses backtracking line-search to determine the stepsize. Recently, there is a line of research initiated by [37] that adaptively estimates the local curvature information in each iteration and does not require the knowledge of L . See [26, 27, 31, 38, 64] for more recent works on this subject. Currently, these adaptive methods are for deterministic problems and it remains an open direction to extend them to stochastic and decentralized settings.

For stochastic gradient descent for convex problems, the current convergence result re-

quires either a constant step upper bounded by $1/(2L)$, or a diminishing stepsize $\eta_t = \eta/\sqrt{t}$ with η still upper bounded by terms related to L [17]. Sufficiently small stepsize guarantees the convergence since the analysis resembles the gradient flow regime, yet this is usually inconsistent with empirical studies, which encourage the stepsize to be large as long as there is no divergence. The stepsize can be chosen up to 10^3 and 10^4 in some logistic regression problems (see Section C.2 in [18]), which indicates that optimal choices of stepsizes in SGD heavily depend on problem parameters.

For nonconvex stochastic optimization, various adaptive methods, such as AdaGrad [14, 40], AMSGrad-Norm [50], NSGD-M [11], are proved to be convergent without any knowledge of the parameters [15, 20, 58], which are thus believed to be more robust algorithms comparing to SGD. Another line of works for the stochastic/online convex optimization is to use the accumulative norm of the stochastic gradient to design adaptive stepsizes [6, 21]. These research results emphasize the optimal dependency on $\|x^0 - x^*\|$, i.e., the distance from the initial to the optimal point, and it is not clear how these works adapt to the nonconvex problems.

Tuning/Parameter-free stochastic optimization in decentralized setting is unexplored, although efforts have been dedicated to understand the adaptive step sizes in decentralized setting [7, 22, 63]. For example, [22, 63] consider the deterministic setting and [22] requires additional Hessian Lipschitz continuity as it considers Newton’s method. Another work [7] analyzes the decentralized AMSGrad provided that the stochastic gradients are bounded, which is not required in our algorithm. It is natural to ask whether one can achieve parameter-free decentralized training, given the unique challenges such as communication complexity and heterogeneous data distributions across agents. We provide an affirmative answer in this paper, and in Table 1 we make the comparison between our Algorithm 2 and existing well-known algorithms: D-SGD [32], D^2 [52] and D-SGT [25]. In particular, D-SGD in [32] requires the information of Lipschitz smoothness parameter and variance of stochastic gradients. D-SGT in [25] requires the Lipschitz smoothness parameter and λ_2, λ_n (see Section 2), which we summarize as ‘net-topology’. Our D-NASA (Algorithm 2) does not require any problem information to select the algorithm parameters.

2 Methodology

We now present the full methodology of our algorithm. First we recall the decentralized communication topology with a weighted undirected graph (V, W) . The vertex set $V = \{1, 2, \dots, n\}$ is the set of local device/nodes, and $W = (W_{i,j}) \in \mathbb{R}^{n \times n}$ is a symmetric doubly stochastic matrix known as weighted adjacency matrix, i.e., W satisfies the following properties: (1) $W_{i,j} \in [0, 1]$, $\forall i, j$; (2) $W_{i,j} = W_{j,i}$, $\forall i, j$, i.e., $W^\top = W$; and (3) $\sum_{j=1}^n W_{i,j} = 1$, $\forall i$, i.e., $W\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\top W = \mathbf{1}^\top$. Intuitively, $W_{i,j}$ represents how well the communication between node i and j is, and $W_{i,j} = 0$ if and only if i and j are not communicating. Note that we assume that the eigenvalues of W satisfy $1 = \lambda_1 > \lambda_2 \geq \dots \geq \lambda_n > -1$, and

$$\rho := \max\{|\lambda_2|, |\lambda_n|\} < 1, \quad (2.1)$$

which is standard in decentralized optimization literature [32, 52]. This ensures the communication graph is strongly connected, and after each round of communication with neighbors, the consensus error (i.e., $\sum_{i=1}^n \|a_i - \bar{a}\|^2$ where a_i is a vector owned by the i -th agent only)

We do not compare with NEXT [13], which only proves asymptotic convergence under deterministic setting.

decreases at a controllable rate. We assume W satisfies the above properties throughout the paper and thus will not explicitly state them in the theorems.

Now we recall the decentralized stochastic gradient tracking (D-SGT) [25, 35, 46, 62] in Algorithm 1. The algorithm takes a gradient step at each local node, keeps a tracker u_i^t to approximate the global stochastic gradient, and executes a communication round in each iteration to achieve consensus among agents. A simple arithmetic verification shows that $\bar{u}^t = \bar{v}^t$ for all iteration number $t > 0$. This key mechanism guarantees that the averaged gradient tracker $\bar{u}^t$ is close to full gradient ∇f provided the consensus error $\sum_{i=1}^n \|x_i^t - \bar{x}^t\|^2$ is small. However, as we will show in Section 3, this popular D-SGT algorithm cannot achieve a tuning/parameter-free convergence with linear speedup even when we assume that the local functions are Lipschitz continuous, i.e., their gradients are bounded. We primarily use D-SGT to showcase the difficulties of applying these algorithms to modern machine learning applications, as we essentially still need to tune the algorithm parameters for a better performance in distributed training, which is largely impossible [60].

Algorithm 1: Decentralized stochastic gradient tracking (D-SGT)

```

1: Input:  $T, \{\eta_t\}, u_i^0 = v_i^0 = \nabla F_i(x_i^0, \xi_i^0)$ 
2: Output:  $\tilde{x} = x^T$  or uniformly from  $\{x^1, \dots, x^T\}$ 
3: for  $t = 0, \dots, T - 1$  do
4:   for each node  $i = 1, \dots, n$  (in parallel) do
5:      $x_i^{t+1} \leftarrow \sum_{j=1}^n W_{i,j}(x_j^t - \eta_t u_j^t),$ 
6:      $v_i^{t+1} \leftarrow \nabla F_i(x_i^{t+1}, \xi_i^{t+1})$ 
7:      $u_i^{t+1} \leftarrow \sum_{j=1}^n W_{i,j}u_j^t + v_i^{t+1} - v_i^t$ 
8:   end for
9: end for

```

To overcome this obstacle, we propose decentralized normalized averaged stochastic approximation (D-NASA) as in Algorithm 2, where we maintain z_i^t as a moving-average update of the tracker u_i^t and then utilize the normalized direction $z_i^t / \|z_i^t\|$ to update x_i^t . Another difference of D-NASA is that we update all the local operations and communicate at the end, simply for the ease of analysis. The moving-average technique, also known as momentum method, was recently introduced to distributed optimization and proven to mitigate client drift in federated learning [9] and achieve linear speedup in decentralized composite optimization [54]. In Section 3, our theory reveals that the normalized direction coupled with the moving-average provably achieves parameter-free decentralized optimization. The combination of normalization and moving-average was explored in [11, 20], yet it is unclear and highly non-trivial to understand if one can achieve parameter-free convergence when each local node is normalized only by its local norm of gradients.

3 Convergence Analysis

In this section we analyze the convergence properties of our algorithms. We have the following standard assumptions for our theoretical analysis of Algorithm 1 and 2.

Assumption 3.1. The function f_i is L_i -Lipschitz smooth, i.e.

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L_i \|x - y\|.$$

As a result, f is L -Lipschitz smooth with $L = \frac{1}{n} \sum_i L_i$.

Algorithm 2: Decentralized normalized averaged stochastic approximation (D-NASA)

```

1: Input:  $T, \{\eta_t\}, \{\alpha_t\}, x_i^0 = z_i^0 = v_i^0 = 0$ 
2: Output:  $\hat{x} = x^T$  or uniformly from  $\{x^1, \dots, x^T\}$ 
3: for  $t = 0, \dots, T - 1$  do
4:   for each node  $i = 1, \dots, n$  (in parallel) do
5:      $\tilde{x}_i^{t+1} \leftarrow x_i^t - \frac{\eta_t}{\|z_i^t\|} z_i^t$ 
6:      $v_i^{t+1} \leftarrow \nabla F_i(x_i^t, \xi_i^t)$ 
7:      $\tilde{u}_i^{t+1} \leftarrow u_i^t + v_i^{t+1} - v_i^t$ 
8:      $\tilde{z}_i^{t+1} \leftarrow (1 - \alpha_t) z_i^t + \alpha_t \tilde{u}_i^{t+1}$ 
9:   end for
10:   $[x_1^{t+1}, \dots, x_n^{t+1}] \leftarrow [\tilde{x}_1^{t+1}, \dots, \tilde{x}_n^{t+1}]W$ 
11:   $[u_1^{t+1}, \dots, u_n^{t+1}] \leftarrow [\tilde{u}_1^{t+1}, \dots, \tilde{u}_n^{t+1}]W$ 
12:   $[z_1^{t+1}, \dots, z_n^{t+1}] \leftarrow [\tilde{z}_1^{t+1}, \dots, \tilde{z}_n^{t+1}]W$ 
13: end for

```

Next, we also have the following standard assumption on the mean and variance of each local gradient estimator. Denote the filtration generated by the random variables sampled upon the t -th iteration as $\mathcal{F}_t$, i.e. $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_t := \sigma(\xi_i^k | i = 1, \dots, n, k = 0, \dots, t), \forall t \geq 1$, where σ is the σ -algebra generated by the random variables.

Assumption 3.2. The stochastic gradient estimator is unbiased and with bounded variance, i.e.,

$$\mathbb{E}_{\xi_i}[\nabla F_i(x, \xi_i)] = \nabla f_i(x), \quad \mathbb{E}_{\xi_i} \|\nabla F_i(x, \xi_i) - \nabla f_i(x)\|^2 \leq \sigma^2.$$

Moreover, we assume $\{\xi_i^{t+1} : i = 1, \dots, n\}$ are independent given $\mathcal{F}_t$.

Note that Assumption 3.2 is only imposed on each local stochastic function, and does not imply any bound for the difference between local and global functions. We now define the notion of stationarity.

Definition 3.1. For any $\epsilon > 0$, we say an algorithm finds an ϵ -stationary point, if an output sequence $\{\bar{x}^t\}_{t=0}^T$ generated by the algorithm satisfies $\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\| \leq \epsilon$.

We say that an algorithm achieves **linear speedup** if it takes $T \propto n^{-1}$ oracles calls at each node to achieve an ϵ -stationary point.

3.1 Parameter-free convergence theory for D-SGT

We first show the convergence analysis of the D-SGT algorithm in which the learning rate does not depend on problem parameters. However, this convergence result requires the following Lipschitz continuity assumption on functions f_i .

Assumption 3.3. The function f_i is G_i -Lipschitz continuous, i.e.

$$\|f_i(x) - f_i(y)\| \leq G_i \|x - y\|.$$

As a result, f is G -Lipschitz continuous with $G = \frac{1}{n} \sum_i G_i$.

Note that Assumption 3.3 is only used in the parameter-free convergence analysis for the D-SGT algorithm (Algorithm 1). This is a very strong assumption since in convex optimization, Lipschitz continuity of the objective functions (or bounded subgradient) can readily give a parameter-independent convergence result by taking the stepsize to be $\mathcal{O}(1/\sqrt{t})$ [5]. Moreover, it implies that each function f_i has bounded gradients, i.e., $\|\nabla f_i(x)\| \leq G_i$, which further indicates the bounded heterogeneity condition since $\|\nabla f_i(x) - \nabla f(x)\| \leq \|\nabla f_i(x)\| + \frac{1}{n} \sum_{j=1}^n \|\nabla f_j(x)\| \leq G_i + G$.

We point out that, even under such a strong assumption, we are not able to show a linear speed up effect for D-SGT. Specifically, we have the following theorem.

Theorem 3.2. *Suppose Assumptions 3.1, 3.2 and 3.3 hold, also take $\eta_t = \eta T^{-1/2}$ (constant) or $\eta_t = \eta t^{-1/2}$ (diminishing, $\eta_0 = 0$ for this case) for $\eta > 0$, the update of Algorithm 1 satisfies:*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 \leq \tilde{\mathcal{O}} \left(\frac{\Delta_0/\eta + (L\sigma^2/n + LG^2)\eta}{\sqrt{T}} + \frac{\tilde{\rho} L^2 \eta^2}{T} (\sigma^2 + G^2) \right).$$

Here $\tilde{\rho} = \frac{\rho}{1-\rho} \frac{1+\rho^2}{1-\rho^2} > 0$ is a parameter dependent on ρ in (2.1), $\Delta_0 = f(\bar{x}^0) - f^*$ is the initial function value gap and we omit higher-order and logarithmic terms in $\tilde{\mathcal{O}}$.

Remark 3.3. The rate of $\mathcal{O}(1/\sqrt{T})$ matches the lower bound for nonconvex stochastic optimization [3], yet it is worth noticing that we are not able to choose the parameter η to achieve a linear speedup effect (even if we have access to n , the number of nodes), due to the term related to G . We also remind the reader that if we assume the access of Lipschitz smooth constant L , one can achieve linear speedup for D-SGT as in [25, 55, 62]. This motivates the design of new algorithms that can achieve linear speedup under tuning/parameter-free setting for decentralized optimization, without the restrictive Assumption 3.3.

3.2 Parameter-free convergence theory for D-NASA

Now we analyze the convergence of D-NASA (Algorithm 2). Similar to Theorem 3.2, we consider both fixed and diminishing stepsizes. We first state the convergence result for fixed stepsizes as follows.

Theorem 3.4. *Suppose Assumptions 3.1 and 3.2 hold, also take $\alpha_t = \alpha T^{-1/2}$ and $\eta_t = \eta T^{-3/4}$ for any $\alpha, \eta > 0$, the update of Algorithm 2 satisfies:*

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\| &\leq \mathcal{O} \left(\left(\frac{\Delta_0}{\eta} + \frac{2L\eta}{\alpha} + \frac{2\sigma\sqrt{\alpha}}{\sqrt{n}} \right) \frac{1}{T^{1/4}} + \frac{\tilde{\rho}^2 \sqrt{10\sigma^2 + 5L^2} \alpha}{T^{1/2}} \right), \\ \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{n} [\|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 + \|\mathbf{Z}^t - \bar{\mathbf{Z}}^t\|^2] &\leq \mathcal{O} \left(\tilde{\rho} \frac{\eta}{T^{1/2}} + \tilde{\rho}^2 (10\sigma^2 + 5L^2) \frac{\alpha^4}{T} + 2\tilde{\rho}^3 \frac{\alpha^4 \eta^2}{T^{5/2}} \right), \end{aligned}$$

as stationarity and consensus error bounds, respectively. Note that we omit higher-order terms in $\mathcal{O}$, and $\tilde{\rho} = \max \left\{ \frac{1}{1-\sqrt{\frac{1+\rho^2}{2}}} \sqrt{\frac{1+\rho^2}{1-\rho^2}}, \frac{\rho}{1-\rho} \frac{1+\rho^2}{1-\rho^2} \right\}$. Taking $\alpha_t = \sqrt{n/T}$ and $\eta_t = n^{1/4}/T^{3/4}$ will give:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\| \leq \mathcal{O} \left(\frac{\Delta_0 + L + \sigma}{n^{1/4} T^{1/4}} + \frac{\tilde{\rho}^2 (\sigma + L) n^{1/2}}{T^{1/2}} \right).$$

We thus stick to this parameter choice in numerical experiments.

Remark 3.5. To make $1/T \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\| \leq \epsilon$, we need $T = \tilde{\mathcal{O}}(1/(n\epsilon^4))$, which matches the lower bounds as in [3, 36], and also indicates the linear speedup effect [32]. Note that the approximation error $\|\bar{z}^t - \nabla f(\bar{x}^t)\|$ also enjoys linear speedup effect. Readers might realize that this choice of parameter requires knowledge of the total number of nodes. We presume that it is impossible to achieve linear speedup if we have none of the problem information. Moreover, this choice of algorithm parameters still does not require global information about the loss function or the topological information about the communication graph, thus it is better than existing algorithms in the literature in decentralized optimization, as we have presented in Table 1.

Remark 3.6. If $\sigma = 0$, we notice that current analysis cannot achieve a $\mathcal{O}(1/\epsilon^2)$ rate of convergence. This suggest that our parameter-free convergence analysis does not achieve the optimal rate for the deterministic case. It remains an open question if parameter-free decentralized deterministic optimization could achieve optimal rate with linear speed up.

To free algorithm parameters even from the total number of iterations T , we present the result when we do not fix the total number of iterations in advance and the stepsize is diminishing in Theorem 3.7.

Theorem 3.7. *Suppose Assumptions 3.1 and 3.2 hold, also take $\alpha_t = \alpha t^{-1/2}$ and $\eta_t = \eta t^{-3/4}$ for any $0 < \alpha < 1$ and $0 < \eta \leq 1$, the update of Algorithm 2 satisfies:*

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\| &\leq \tilde{\mathcal{O}}\left(\frac{\Delta_0/\eta + L\eta/\alpha + L\tilde{\rho}\eta + \sigma\sqrt{\alpha}/\sqrt{n} + \tilde{\rho}^2(\sigma + L)\alpha + L\tilde{\rho}^3\eta\alpha + L\eta}{T^{1/4}}\right), \\ \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{n} \mathbb{E} [\|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 + \|\mathbf{Z}^t - \bar{\mathbf{Z}}^t\|^2] &\leq \mathcal{O}\left(\frac{\tilde{\rho}^2(\sigma^2 + L^2 + \tilde{\rho}\eta^2)}{T}\right). \end{aligned}$$

Note that we omit logarithmic factors in $\tilde{\mathcal{O}}$ and $\tilde{\rho}$ is defined in Theorem 3.4.

Remark 3.8. To ensure $1/T \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\| \leq \epsilon$, we need $T = \tilde{\mathcal{O}}(1/\epsilon^4)$, which again matches the lower bound as in [3, 36] up to logarithmic factors. Yet we are not able to achieve a concrete linear speedup effect with this choice of algorithm parameters. It is now an open question if one can achieve linear speedup without knowing the total number of iterations T and other problem parameters. Nevertheless, we show in the numerical experiments (see Figure 1) that the stepsize choices $\alpha = \sqrt{n}$ and $\eta = n^{1/4}$ can still achieve linear speedup empirically, and we thus stick to this choice of parameters in experiments.

Remark 3.9. Here we briefly discuss the dependency on ρ in the convergence rate. At the first glance, existing analyses [2, 24, 25, 43] may provide better dependency on the network topology characterized by ρ . However, none of these works consider the tuning/parameter-free setting, which requires additional investigation on achieving the optimal dependency on ρ .

4 Numerical Experiments

In this section, we test D-NASA (Algorithm 2) numerically and compare it with existing algorithms such as D-SGD [32], D-SGT (Algorithm 1) and D-ASAGT [54]. We follow the experimental setup in the code framework of [39] to test the algorithms on real datasets using mpi4py [12] and PyTorch [45].

[54] considers nonsmooth proximal version of the algorithm. In our numerical experiments we simply regard the nonsmooth proximal term as zero.

4.1 Synthetic data experiments

We first use synthetic data to verify the linear speedup effect of D-NASA (Algorithm 2). We consider a simple linear regression model with the data sample $\xi = (X, Y)$ generated by $Y = X^\top \theta_* + \epsilon$ where $X, \theta_* \in \mathbb{R}^d$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ are Gaussian noise. We solve the following least-square problem:

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{(X,Y) \sim \mathcal{D}_i} [(Y - X^\top \theta)^2]. \quad (4.1)$$

In our experiment, we set $d = 100$, data $X \sim \mathcal{N}(0, I_d)$ and $\sigma = 0.1$. We simulate streaming data samples with batch size = 1 for training and 10000 data samples per node for evaluations. We employ a ring topology for the network where self-weighting and neighbor weights are set to be 1/3. For D-NASA, we try both fixed stepsizes ($\alpha_t = \sqrt{n/T}$, $\eta_t = n^{1/4}/T^{3/4}$) and diminishing stepsizes ($\alpha_t = \min\{1, \sqrt{n/t}\}$, $\eta_t = n^{1/4}/t^{3/4}$) where the total number of iteration $T = 15000$ and $n \in \{5, 10, 20\}$. Figure 1 shows results of our experiment. It could be seen that with more number of nodes D-NASA is more efficient in terms of both test loss and the norm of the gradient (at the global point $\bar{x}^t$).

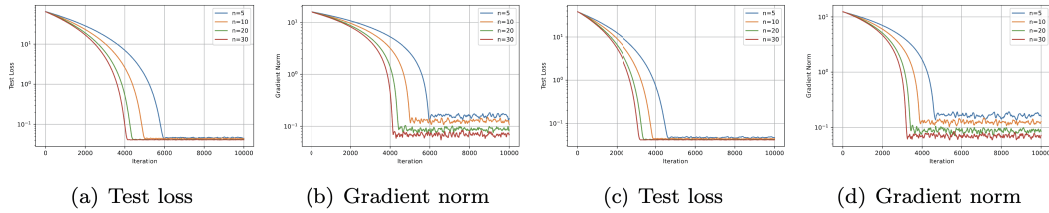


Figure 1: The convergence curve of Algorithm 2 to solve (4.1) with different choice of number of nodes $n \in \{5, 10, 20, 30\}$. The left two figures correspond to fixed stepsizes ($\alpha_t = \sqrt{n/T}$, $\eta_t = n^{1/4}/T^{3/4}$) and right two corresponds to diminishing stepsizes ($\alpha_t = \min\{1, \sqrt{n/t}\}$, $\eta_t = n^{1/4}/t^{3/4}$), respectively.

Next, we compare D-NASA on (4.1) with the other three algorithms. We still set $d = 100$, yet with a spike model with $X \sim \mathcal{N}(0, \text{diag}(100, 1, \dots, 1))$ where only the first entry has a large variance, in order to make the Lipschitz smooth constant of (4.1) large. It is worth noting that despite that conservative constant stepsize choices (usually $\mathcal{O}(\sqrt{n/T})$) can lead to linear speedup effect in decentralized training theoretically [32, 52], this choice is usually for the sake of proof simplicity (see footnote on Page 6 of [32]). In practice it is tempting to choose diminishing stepsize in the learning rate scheduler, since the model training often benefits from large stepsizes, a phenomenon that has attracted a lot of attention recently in deep learning community [10, 29]. We thus compare D-SGD, D-SGT, D-ASAGT with D-NASA using diminishing stepsizes with a tunable hyperparameter.

For D-SGD and D-SGT, we test the algorithm with diminishing stepsizes $\eta_t = \eta \sqrt{n/t}$ as suggested by [25, 32]; For D-ASAGT, we test the algorithm with stepsizes $\eta_t = \eta \sqrt{n/t}$ and $\alpha_t = \min\{\sqrt{n/t}, 0.3\}$ as suggested in their experiments [54]; For D-NASA we fix $\eta_t = n^{1/4}/t^{3/4}$ and $\alpha_t = \min\{1, \sqrt{n/t}\}$ based on our theoretical analysis. We conduct a simple grid search for D-SGD, D-SGT and D-ASAGT to determine and use the best choices

Due to page limit, we refer to the appendix for the experiment on different choices of η_t .

of η for each algorithms. The convergence result is shown in Figure 2. We observe that for all algorithms, the test losses decrease with oscillations, presenting the catapults [29] and Edge of Stability (EOS) [10] phenomena, two closely related large-step-size regimes in which the training converges non-monotonically with oscillations and usually generalize better than small-step-size settings [1, 4, 10, 29]. Furthermore, we observe that the test loss of our algorithm is much lower than other baselines, indicating superior generalization performance. We emphasize that different from the large-step-size training setup in the literature, our Algorithm presents the catapults and EOS without any hyperparameter tuning, while other algorithms even require tedious tuning to prevent divergent training in this ill-conditioned problem.

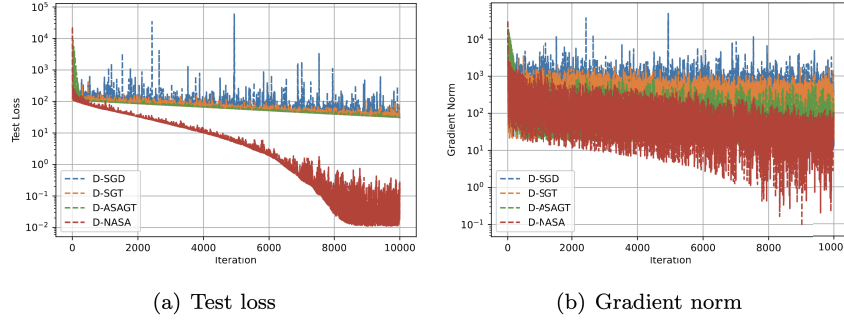


Figure 2: Convergence curve for D-SGD, D-SGT, D-ASAGT and D-NASA for solving (4.1) under the spike model.

4.2 Real-world data experiments

We utilize the code framework in [39] where we compare D-NASA with D-SGD, D-SGT, D-ASAGT for solving the classification problem:

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^n \frac{1}{|\mathcal{D}_i|} \sum_{(x,y) \in \mathcal{D}_i} \ell(f(x; \theta), y) \quad (4.2)$$

on MNIST, a9a and miniboone datasets. Here ℓ denotes the cross-entropy loss, and f represents a neural network parameterized by θ with x being its input data. $\mathcal{D}_i$ is the training set only available to agent i . We use a 2-layer perceptron model on a9a and miniboone, and the LeNet architecture [28] for the MNIST dataset. We take $n = 8$ which connect in the form of a random graph ($\rho = 0.375$) for all three datasets. The data is divided evenly to $n = 8$ devices (CPUs) and using mpi4py interface to communicate the computation results. The batch-sizes are fixed to be 32.

Similar to the synthetic data, for D-SGD and D-SGT, again we test the algorithm with diminishing stepsizes $\eta_t = \eta \sqrt{n/t}$; For D-ASAGT, we test the algorithm with $\eta_t = \eta \sqrt{n/t}$ and $\alpha_t = \min\{\sqrt{n/t}, 0.3\}$; For D-NASA we again take $\eta_t = n^{1/4}/t^{3/4}$ and $\alpha_t = \sqrt{n/t}$ based on our theory. Figure 3 shows the test accuracy under different stepsizes $\eta \in$

Available at <https://www.openml.org>

To make sure that the graph is connected, we set the probability of each two node being connected as 0.8. We refer to Appendix A for more graph designs.

$\{0.005, 0.01, 0.5, 1, 5, 10, 50, 100\}$, where the dashed horizontal line is the result for D-NASA. We can see that D-NASA yields comparable numerical results without tuning any parameters, and other three algorithms can also work well under certain parameter choices.

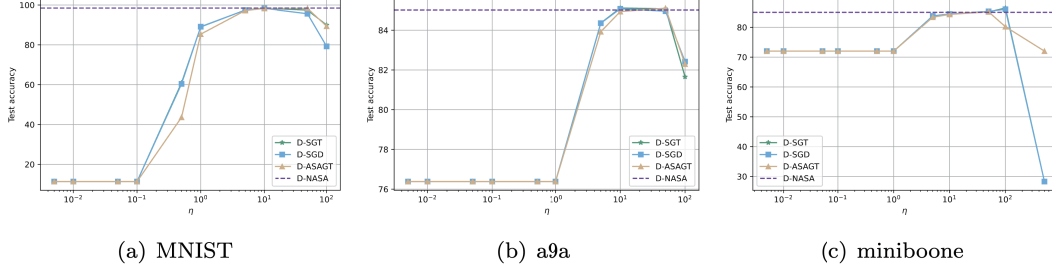


Figure 3: The testing accuracy of the outputs from different algorithms with respect to different choices of learning rates. The test accuracy is defined as the ratio of correct predictions in the output of the model to all test data points, and the vertical axis uses % as the unit.

5 Conclusion

In this paper we propose D-NASA, a tuning/parameter-free decentralized stochastic optimization algorithm and give its finite-time convergence analysis. Moreover, we showcase that in comparison to other baselines, our algorithm demonstrates superior generalization without tedious hyperparameter tuning process, thus having great potential for large scale machine learning problems.

References

- [1] K. Ahn, S. Bubeck, S. Chewi, Y.T. Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability, in: *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [2] S.A. Alghunaim and K. Yuan, A unified and refined convergence analysis for non-convex decentralized learning. *IEEE Transactions on Signal Processing* 70 (2022) 3264–3279.
- [3] Y. Arjevani, Y. Carmon, J.C. Duchi, D.J. Foster, N. Srebro and B. Woodworth, Lower bounds for non-convex stochastic optimization. *Mathematical Programming* 199 (2023) 165–214.
- [4] S. Arora, Z. Li and A. Panigrahi, Understanding gradient descent on the edge of stability in deep learning, in] *International Conference on Machine Learning*, PMLR, 2022, pp. 948–1024.
- [5] S. Boyd, L. Xiao and A. Mutapcic, Subgradient methods, *lecture notes of EE392o, Stanford University, Autumn Quarter, 2004(01)*, 2003.

It can be seen that D-SGD, D-SGT and D-ASAGT all seem to work well when η is around 10. We believe the main reason is that all the datasets we tested are normalized and the Lipschitz smooth constants are fairly similar for all three datasets.

- [6] Y. Carmon and O. Hinder, Making SGD parameter-free. in: *Conference on Learning Theory*, PMLR, 2022, pp. 2360–2389.
- [7] X. Chen, B. Karimi, W. Zhao and P. Li, Convergent adaptive gradient methods in decentralized optimization, 2020.
- [8] X. Chen, B. Karimi, W. Zhao and Ping Li, On the convergence of decentralized adaptive gradient methods. in: *Asian Conference on Machine Learning*, PMLR, 2023, pp. 217–232.
- [9] Z. Cheng, X. Huang, P. Wu and K. Yuan, Momentum benefits non-iid federated learning simply and provably, in: *The Twelfth International Conference on Learning Representations*, 2024.
- [10] J. Cohen, S. Kaur, Y. Li, J.Z. Kolter and A. Talwalkar, Gradient descent on neural networks typically occurs at the edge of stability, in: *International Conference on Learning Representations*, 2021.
- [11] A. Cutkosky and H. Mehta, Momentum improves normalized SGD, in: *International conference on machine learning*, PMLR, 2020, pp. 2260–2268.
- [12] L. Dalcin and Y.L.L. Fang, mpi4py: Status update after 12 years of development, *Computing in Science & Engineering* 23 (2021) 47–54.
- [13] P.D. Lorenzo and G. Scutari, Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks* 2 (2016) 120–136.
- [14] J. Duchi, E. Hazan and Y. Singer, Adaptive subgradient methods for online learning and stochastic optimization, *Journal of Machine Learning Research* 12 (2011) 2121–2159.
- [15] M. Faw, I. Tziotis, C. Caramanis, A. Mokhtari, S. Shakkottai and R. Ward, The power of adaptivity in SGD: Self-tuning step sizes with unbounded gradients and affine variance, in: *Conference on Learning Theory*, PMLR, 2022, pp. 313–355.
- [16] S. Gan, J. Jiang, B. Yuan, C. Zhang, X. Lian, R. Wang, J. Chang, C. Liu, H. Shi, S. Zhang, X. Li, T. Sun, S. Yang and J. Liu, Bagua: scaling up distributed learning with system relaxations, *Proc. VLDB Endow.* 15 (2021) 804–813.
- [17] G. Garrigos and R.M. Gower, Handbook of convergence theorems for (stochastic) gradient methods, *arXiv preprint arXiv:2301.11235*, 2023.
- [18] R. Grazzi, L. Franceschi, M. Pontil and S. Salzo, On the iteration complexity of hypergradient computation. in: *International Conference on Machine Learning*, PMLR, 2020, pp. 3748–3758.
- [19] M. Hong, M. Razaviyayn, Z.-Q. Luo and J.-S. Pang, A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing, *IEEE Signal Processing Magazine* 33 (2015) 57–77.
- [20] F. Hübler, J. Yang, X. Li and N. He, Parameter-agnostic optimization under relaxed smoothness. in: *OPT 2023: Optimization for Machine Learning*, 2023.
- [21] M. Ivgi, O. Hinder and Y. Carmon, DoG is SGD’s best friend: A parameter-free dynamic step size schedule, in *International Conference on Machine Learning*. PMLR, 2023, pp. 14465–14499.

- [22] D. Jakovetic, N. Krejic and G. Malaspina, Distributed inexact newton method with adaptive step sizes, *arXiv preprint arXiv:2305.13985*, 2023.
- [23] A. Koloskova, T. Lin, S.U. Stich and M. Jaggi, Decentralized deep learning with arbitrary communication compression, in: *International Conference on Learning Representations*, 2020.
- [24] 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.
- [25] A. Koloskova, T. Lin and S.U. Stich, An improved analysis of gradient tracking for decentralized machine learning, in: *Advances in Neural Information Processing Systems*, 34, 2021.
- [26] P. Latafat, A. Themelis and P. Patrinos, On the convergence of adaptive first order methods: proximal gradient and alternating minimization algorithms, in: *6th Annual Learning for Dynamics & Control Conference*, PMLR, 2024, pp. 197–208.
- [27] P. Latafat, A. Themelis, L. Stella and P. Patrinos, Adaptive proximal algorithms for convex optimization under local lipschitz continuity of the gradient. *arXiv preprint arXiv:2301.04431*, 2023.
- [28] Y. LeCun et al, Lenet-5, convolutional neural networks, URL: <http://yann.lecun.com/exdb/lenet>, 20(5):14, 2015.
- [29] A. Lewkowycz, Y. Bahri, E. Dyer, J.S.-D. and G. Gur-Ari, The large learning rate phase of deep learning: the catapult mechanism, *arXiv preprint arXiv:2003.02218*, 2020.
- [30] M. Li, D.G. Andersen, J.W. Park, A.J. Smola, A. Ahmed, V. Josifovski, J. Long, E.J. Shekita and B.-Y. Su, Scaling distributed machine learning with the parameter server, in: *11th USENIX Symposium on operating systems design and implementation (OSDI 14)*, 2014, pp. 583–598.
- [31] T. Li and G. Lan, A simple uniformly optimal method without line search for convex optimization, *arXiv preprint arXiv:2310.10082*, 2023.
- [32] 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*, 30, 2017.
- [33] X. Lian, W. Zhang, C. Zhang and J. Liu, Asynchronous decentralized parallel stochastic gradient descent. in *International Conference on Machine Learning*, PMLR, 2018, pp. 3043–3052.
- [34] Y. Liu, T. Lin, A. Koloskova and S.U. Stich, Decentralized gradient tracking with local steps, *Optimization Methods and Software* (2024) 1–28. <https://doi.org/10.1080/10556788.2024.2322095>.
- [35] S. Lu, X. Zhang, H. Su, and M. Hong, Gnsd: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization, in: *2019 IEEE Data Science Workshop (DSW)*, IEEE, 2019, pp. 315–321.

- [36] Y. Lu and C. De Sa, Optimal complexity in decentralized training, in: *International Conference on Machine Learning*, PMLR, 2021, pp. 7111–7123.
- [37] Y. Malitsky and K. Mishchenko, Adaptive gradient descent without descent, ICML’20. JMLR.org, 2020.
- [38] Y. Malitsky and K. Mishchenko, Adaptive proximal gradient method for convex optimization, *arXiv preprint arXiv:2308.02261*, 2023.
- [39] G. Mancino-Ball, S. Miao, Y. Xu and J. Chen, Proximal stochastic recursive momentum methods for nonconvex composite decentralized optimization, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 9055–9063.
- [40] H.B. McMahan and M. Streeter, Adaptive bound optimization for online convex optimization, *arXiv preprint arXiv:1002.4908*, 2010.
- [41] A. Nedić and J. Liu, Distributed optimization for control. *Annual Review of Control, Robotics, and Autonomous Systems*, 1 (2018) 77–103.
- [42] A. Nedic, A. Olshevsky and W. Shi, Achieving geometric convergence for distributed optimization over time-varying graphs, *SIAM Journal on Optimization* 27 (2017) 2597–2633.
- [43] G. Neglia, C. Xu, D. Towsley and G. Calbi, Decentralized gradient methods: does topology matter? in: *International Conference on Artificial Intelligence and Statistics*, PMLR, 2020, pp. 2348–2358.
- [44] Yurii Nesterov, *Lectures on Convex Optimization*, vol. 137. Springer, 2018.
- [45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimeshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: *Advances in neural information processing systems*, 32, 2019.
- [46] S. Pu and A. Nedić, Distributed stochastic gradient tracking methods, *Mathematical Programming* 187 (2021) 409–457.
- [47] G. Qu and N. Li, Harnessing smoothness to accelerate distributed optimization, *IEEE Transactions on Control of Network Systems* 5 (2017) 1245–1260.
- [48] J.P. Queralta, J. Taipalmaa, B.C. Pullinen, V.K. Sarker, T.N. Gia, H. Tenhunen, M. Gabbouj, J. Raitoharju and T. Westerlund, Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision, *Ieee Access* 8 (2020) 191617–191643.
- [49] S.S. Ram, A. Nedić and V.V. Veeravalli, Asynchronous gossip algorithms for stochastic optimization. in: *Proceedings of the 48th IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, IEEE, 2009, pp. 3581–3586.
- [50] S.J. Reddi, S. Kale and S. Kumar, On the convergence of Adam and beyond, *arXiv preprint arXiv:1904.09237*, 2019.

- [51] H. Tang, S. Gan, C. Zhang, T. Zhang and J. Liu, Communication compression for decentralized training, in: *Advances in Neural Information Processing Systems*, 31, 2018.
- [52] H. Tang, X. Lian, M. Yan, C. Zhang and J. Liu, D^2 : Decentralized training over decentralized data. in: *International Conference on Machine Learning*, PMLR, 2018, pp. 4848–4856.
- [53] J.N. Tsitsiklis, *Problems in decentralized decision making and computation*, PhD thesis, Massachusetts Institute of Technology, 1984.
- [54] T. Xiao, X. Chen, K. Balasubramanian and S. Ghadimi, A one-sample decentralized proximal algorithm for non-convex stochastic composite optimization, in: *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence, Proceedings of Machine Learning Research*, J.E. Robin, I. Shpitser (eds), vol. 216, PMLR, 31 Jul–04 Aug 2023, pp. 2324–2334.
- [55] R. Xin, U.A. Khan and S. Kar, An improved convergence analysis for decentralized online stochastic non-convex optimization, *IEEE Transactions on Signal Processing* 69 (2021) 1842–1858.
- [56] J. Xu, S. Zhu, Y.C. Soh and L. Xie, Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes, in: *2015 54th IEEE Conference on Decision and Control (CDC)*, IEEE, 2015, pp. 2055–2060.
- [57] F. Yan, S. Sundaram, S.V.N. Vishwanathan and Y. Qi, Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties, *IEEE Transactions on Knowledge and Data Engineering*, 25 (2012) 2483–2493.
- [58] J. Yang, X. Li, I. Fatkhullin and N. He, Two sides of one coin: the limits of untuned SGD and the power of adaptive methods, in: *Thirty-seventh Conference on Neural Information Processing Systems*, 2023, pp. 74257–74288.
- [59] T. Yang, Xi. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin and K.H. Johansson, A survey of distributed optimization. *Annual Reviews in Control* 47 (2019) 278–305.
- [60] B. Yuan, Y. He, J. Davis, T. Zhang, T. Dao, B. Chen, P.S. Liang, C. Re and C. Zhang, Decentralized training of foundation models in heterogeneous environments, in: *Advances in Neural Information Processing Systems*, 35, 2022, pp. 25464–25477.
- [61] K. Yuan, Q. Ling and W. Yin, On the convergence of decentralized gradient descent. *SIAM Journal on Optimization* 26 (2016) 1835–1854.
- [62] J. Zhang and K. You, Decentralized stochastic gradient tracking for non-convex empirical risk minimization, *arXiv preprint arXiv:1909.02712*, 2019.
- [63] X. Zhang, S. Liu and N. Zhao, An accelerated algorithm for distributed optimization with barzilai-borwein step sizes, *Signal Processing* 202 (2023): 108748,.
- [64] D. Zhou, S. Ma and J. Yang, AdaBB: Adaptive Barzilai-Borwein method for convex optimization, *arXiv preprint arXiv:2401.08024*, 2024.

Appendix

A Details of Experiments

Our experiments are performed on Amazon AWS EC2 g5.4xlarge cluster which consists of 16 vCPUs with 64 GiB memory and NVIDIA A10G GPU with 24 GiB memory. All the experiments are conducted on CPU where 8 CPU are used to imitate 8 different nodes. The network topology is chosen as Figure 4.

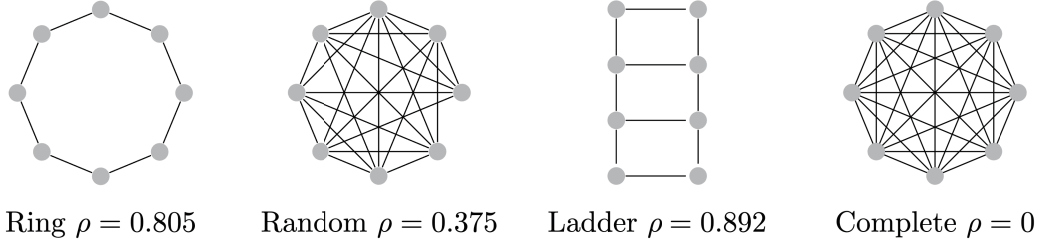


Figure 4: Network topology for $n = 8$. The four graphs represent the ring, (an instance of) the random, the ladder and the complete graph.

In addition to the theoretical analysis in Section 3.1, we conduct further experiments to show that D-SGT cannot achieve linear-speedup in the **parameter-free** setting. We set the step size to be exactly $\sqrt{n/T}$ for D-SGT, and observe from Figure 5 that indeed D-SGT could not achieve linear speedup in this case.

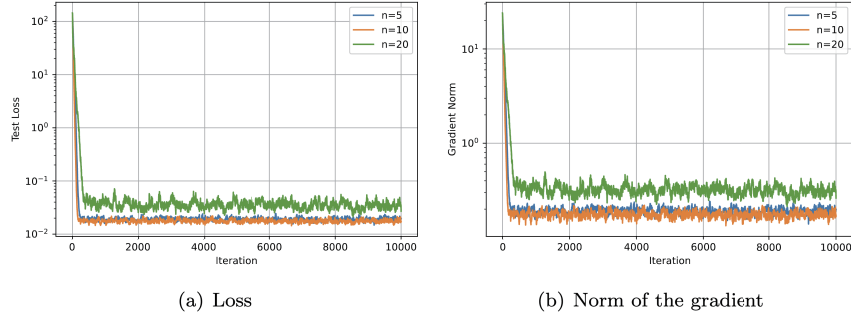


Figure 5: The loss and norm of the stochastic gradient for D-SGT when the step size is exactly $\sqrt{n/T}$. The data is generated same as Figure 1.

In our experiment in Section 4.2, we fix the stepsizes as $\eta_t = n^{1/4}/t^{3/4}$ and $\alpha_t = \sqrt{n/t}$ based on our theoretical analysis. The readers may wonder what would happen if we take $\eta_t = \eta n^{1/4} T^{-3/4}$ with different η . We tested this experiment in Figure and found that indeed our choice of parameter is optimal, i.e. $\eta = 1$ in general gives the optimal performance in terms of the testing accuracy.

We now present the additional results for testing D-SGD, D-SGT, D-ASAGT, D-NASA on (4.2) with a9a data over different network topology as specified in Figure 4. The hyper-parameters follow exactly the same as in Section 4.2. The results are presented in Figure

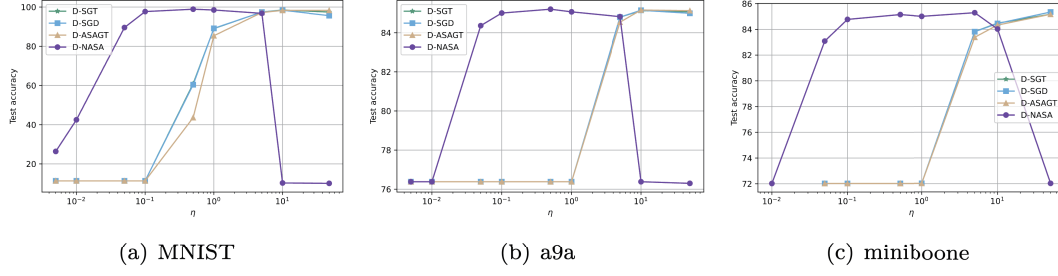


Figure 6: The effect of different η in $\eta_t = \eta n^{1/4} T^{-3/4}$. Testing setting is exactly same as Figure 3 and all other algorithms' curve are the same as Figure 3.

7. It can be seen that D-NASA achieves competitive testing accuracy under almost every network topology choice.

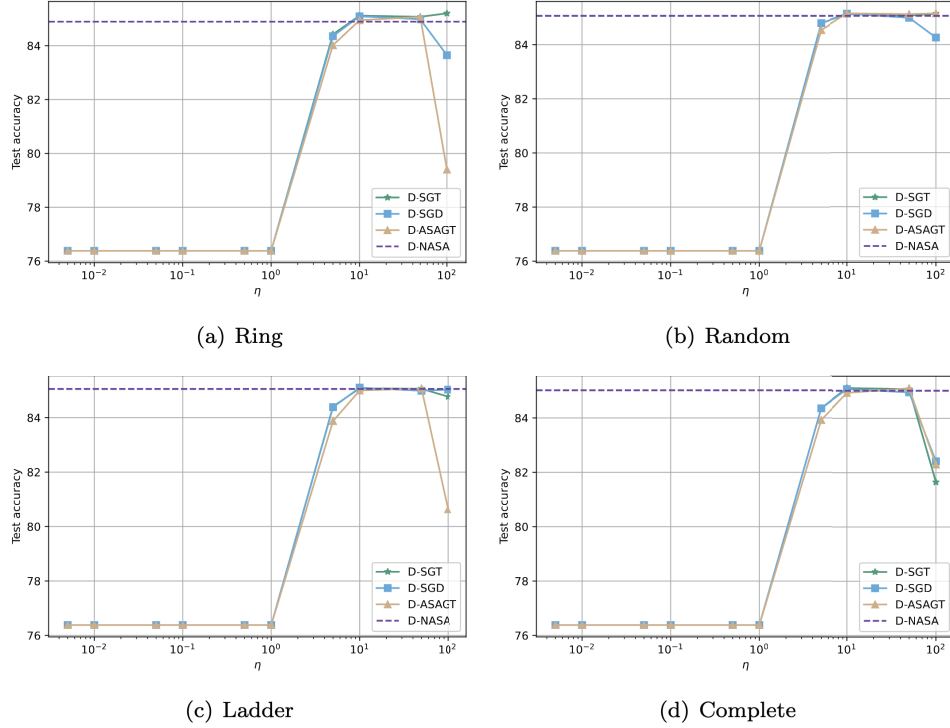


Figure 7: The testing accuracy of the outputs from different algorithms with respect to different choices of learning rates for a9a dataset. The four figures corresponds to four different network graphs as in Figure 4.

We also include the figures of the loss, accuracy, and stationarity curves of all algorithms. Figure 8 and 9 shows the training/testing curve with respect to training epoch or CPU time when applying the four algorithms to (4.2) with MNIST and a9a dataset. We show each algorithm with the **best choice of stepsizes** in the light of Figure 3. One can see that

D-NASA achieves competitive rate of convergence with exactly the same stepsize choice as our theory, without tuning any parameter.

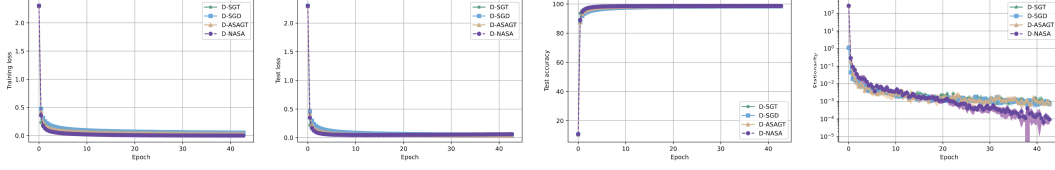


Figure 8: The convergence curve of D-SGD, D-SGT, D-ASAGT, D-NASA for the MNIST dataset, while the first three are at their **best stepsizes** (after the grid search as in Figure 3), and D-NASA follows the stepsize choice as in Remark 3.8, i.e. $\eta_t = n^{1/4}/t^{3/4}$ and $\alpha_t = n^{1/2}/t^{1/2}$. The four columns are the curves for training loss, testing loss, testing accuracy and stationarity, respectively. The experiments are repeated and averaged for 10 times.

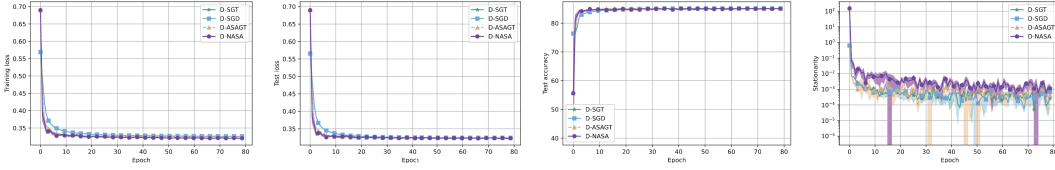


Figure 9: The convergence curve of D-SGD, D-SGT, D-ASAGT, D-NASA for the a9a dataset, while the first three are at their **best stepsizes** (after the grid search as in Figure 3), and D-NASA again follow the same choice as Figure 8. The experiments are repeated and averaged for 10 times.

B Convergence Analysis

B.1 Parameter-free convergence theory for D-SGT

From the update of Algorithm 1, we have that:

$$\begin{aligned} \mathbf{X}^{t+1} &= \mathbf{X}^t - \eta_t \mathbf{U}^t W, \quad \bar{x}^{t+1} = \bar{x}^t - \eta_t \bar{u}^t, \\ \bar{u}^t &= \bar{v}^t = \frac{1}{n} \sum_{i=1}^n \nabla F_i(x_i^t, \xi_i^t). \end{aligned} \quad (\text{B.1})$$

The following descent lemma characterizes the difference between the function values of two consecutive iterates for Algorithm 1:

Lemma B.1. *Suppose Assumption 3.1 and 3.3 holds. Algorithm 1 satisfies:*

$$\mathbb{E}[f(\bar{x}^{t+1})] - \mathbb{E}[f(\bar{x}^t)] \leq -\frac{3\eta_t}{4} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 + 2\eta_t^2 L \frac{\sigma^2}{n} + (\eta_t + 2\eta_t^2 L) \frac{L^2}{n} \mathbb{E} \|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 + \eta_t^2 L G^2,$$

where the expectation is taken conditioned on $\mathcal{F}_{t-1}$.

Proof. By the L -Lipschitz smooth of f (Assumption 3.1) we get:

$$\begin{aligned}
& f(\bar{x}^{t+1}) - f(\bar{x}^t) \\
& \leq \nabla f(\bar{x}^t)^\top (\bar{x}^{t+1} - \bar{x}^t) + \frac{L}{2} \|\bar{x}^{t+1} - \bar{x}^t\|^2 = -\eta_t \nabla f(\bar{x}^t)^\top \bar{u}^t + \frac{\eta_t^2 L}{2} \|\bar{u}^t\|^2 \\
& = -\eta_t \|\nabla f(\bar{x}^t)\|^2 - \eta_t \nabla f(\bar{x}^t)^\top (\bar{v}^t - \nabla f(\bar{x}^t)) + \frac{\eta_t^2 L}{2} \|\bar{v}^t\|^2 \\
& = -\eta_t \|\nabla f(\bar{x}^t)\|^2 - \eta_t \nabla f(\bar{x}^t)^\top (\bar{v}^t - h^t + h^t - \nabla f(\bar{x}^t)) + \frac{\eta_t^2 L}{2} \|\bar{v}^t\|^2
\end{aligned}$$

where $h^t := \frac{1}{n} \sum_{i=1}^n \nabla f_i(x_i^t)$.

Now taking the expectation conditioned on $\mathcal{F}_{t-1}$, we get

$$\begin{aligned}
& \mathbb{E}[f(\bar{x}^{t+1}) - f(\bar{x}^t)] \\
& \leq -\eta_t \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 - \eta_t \mathbb{E}[\nabla f(\bar{x}^t)^\top (h^t - \nabla f(\bar{x}^t))] + \frac{\eta_t^2 L}{2} \mathbb{E} \|\bar{v}^t\|^2 \\
& \leq -\eta_t \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 + \eta_t / 4 \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 + \eta_t \mathbb{E} \|h^t - \nabla f(\bar{x}^t)\|^2 + \eta_t^2 L \mathbb{E} \|\bar{v}^t - \nabla f(\bar{x}^t)\|^2 + \eta_t^2 L G^2 \\
& \leq -\frac{3\eta_t}{4} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 + 2\eta_t^2 L \mathbb{E} \|\bar{v}^t - h^t\|^2 + (\eta_t + 2\eta_t^2 L) \mathbb{E} \|h^t - \nabla f(\bar{x}^t)\|^2 + \eta_t^2 L G^2
\end{aligned}$$

where the second inequality is by $\mathbb{E}[\bar{v}^t] = h^t$, Cauchy-Schwarz, $a^\top b \leq 1/\gamma \|a\|^2 + \gamma \|b\|^2$ and Assumption 3.3, and the third is by $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$.

Now taking the conditional expectation over $\mathcal{F}_{t-1}$ we get:

$$\mathbb{E} \|\bar{v}^t - h^t\|^2 = \mathbb{E} \left\| \frac{1}{n} \sum_i (\nabla F_i(x_i^t, \xi_i^t) - \nabla f(x_i^t)) \right\|^2 = \frac{1}{n} \sum_i \mathbb{E} \|\nabla F_i(x_i^t, \xi_i^t) - \nabla f(x_i^t)\|^2 \leq \frac{\sigma^2}{n}$$

due to Assumption 3.2.

As for the term $\|h^t - \nabla f(\bar{x}^t)\|^2$, we have

$$\|h^t - \nabla f(\bar{x}^t)\|^2 = \left\| \frac{1}{n} \left(\sum_{i=1}^n \nabla f_i(x_i^t) - \nabla f_i(\bar{x}^t) \right) \right\|^2 \leq \frac{L^2}{n} \sum_i \|x_i^t - \bar{x}^t\|^2 = \frac{L^2}{n} \|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2.$$

□

We have the following lemma about the consensus error, that is, the average distance of each node to the global average.

Lemma B.2. *For the update of Algorithm 1, we have:*

$$\begin{aligned}
\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}^{t+1}\|^2 & \leq \frac{1+\rho}{2} \|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 + \eta_t^2 \frac{1+\rho^2}{1-\rho^2} \|\mathbf{U}^t - \bar{\mathbf{U}}^t\|^2, \\
\|\mathbf{U}^{t+1} - \bar{\mathbf{U}}^{t+1}\|^2 & \leq \frac{1+\rho}{2} \|\mathbf{U}^t - \bar{\mathbf{U}}^t\|^2 + \frac{1+\rho^2}{1-\rho^2} \|\mathbf{V}^{t+1} - \bar{\mathbf{V}}^t\|^2.
\end{aligned}$$

Proof. Since

$$\begin{aligned}
\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}^{t+1}\|^2 &= \|(\mathbf{X}^t - \eta_t \mathbf{U}^t)W - (\bar{x}^t - \eta_t \bar{u}^t)\mathbf{1}^\top\|^2 \\
&= \|(\mathbf{X}^t - \eta_t \mathbf{U}^t)W - \frac{1}{n}(\mathbf{X}^t - \eta_t \mathbf{U}^t)\mathbf{1}\mathbf{1}^\top\|^2 = \|(\mathbf{A}^t - \mathbf{A}^t \frac{\mathbf{1}\mathbf{1}^\top}{n})(W - \frac{\mathbf{1}\mathbf{1}^\top}{n})\|^2 \\
&\leq \|\mathbf{A}^t - \mathbf{A}^t \frac{\mathbf{1}\mathbf{1}^\top}{n}\|^2 \|W - \frac{\mathbf{1}\mathbf{1}^\top}{n}\|_2^2 \\
&\leq \rho^2 \|\mathbf{A}^t - \mathbf{A}^t \frac{\mathbf{1}\mathbf{1}^\top}{n}\|^2 = \rho^2 \|(\mathbf{X}^t - \bar{x}^t \mathbf{1}^\top) - \eta_t(\mathbf{U}^t - \bar{u}^t \mathbf{1}^\top)\|^2 \\
&\leq \rho^2 (1 + \frac{1}{c}) \|\mathbf{X}^t - \bar{x}^t \mathbf{1}^\top\|^2 + \rho^2 \eta_t^2 (1 + c) \|\mathbf{U}^t - \bar{u}^t \mathbf{1}^\top\|^2
\end{aligned}$$

where $\mathbf{A}^t := \mathbf{X}^t - \eta_t \mathbf{U}^t$. Taking $c = \frac{2\rho^2}{1-\rho^2} \geq 0$ gives the desired result. For the consensus error of $\mathbf{U}^t$ we could get it in a similar way. $\square$

With the analysis of one step of the consensus error, we are readily to analyze the cumulative consensus error for the final convergence. To do this, we need the following technical lemma:

Lemma B.3 (Lemma 3.3 in [54]). *Suppose we are given three sequences $\{a_n\}_{n=0}^\infty$, $\{c_n\}_{n=0}^\infty$, $\{\tau_n\}_{n=0}^\infty$, and a constant $r \in (0, 1)$ such that $a_k, b_k \geq 0$, $0 = c_{-1} \leq c_{k+1} \leq c_k$ and*

$$a_{k+1} \leq r a_k + b_k,$$

then we have

$$\sum_{k=0}^K c_k a_k \leq \frac{1}{1-r} \left(c_0 a_0 + \sum_{k=0}^K c_k b_k \right)$$

for any positive integer K .

Note that Lemma 3.3 in [54] requires $c_k \leq 1$ but it's not necessary in the proof.

Now we are ready to analyze the cumulative consensus error for Algorithm 1 as follows:

Lemma B.4. *For the update of Algorithm 1, under Assumption 3.3 and 3.2, we have:*

$$\sum_{t=0}^{T-1} \frac{1}{n} \eta_t^\tau \mathbb{E} \|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 \leq 10\tilde{\rho} \sum_{t=0}^{T-1} \eta_t^{\tau+2} (\sigma^2 + G^2),$$

where $\tau = 0, 1$ or 2 and

$$\tilde{\rho} := \frac{\rho}{1-\rho} \frac{1+\rho^2}{1-\rho^2}.$$

Note that $\tilde{\rho}$ is greater than 0.

Proof. First by applying Lemma B.2 and B.3 with $a_k = \frac{1}{n} \|\mathbf{X}^k - \bar{\mathbf{X}}^k\|^2$, $b_k = \eta_k^2 \frac{1+\rho^2}{1-\rho^2} \frac{1}{n} \|\mathbf{U}^k - \bar{\mathbf{U}}^k\|^2$ and $r = (1 + \rho)/2$, we get

$$\sum_{t=0}^{T-1} \frac{1}{n} \mathbb{E} \|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 \leq \tilde{\rho} \sum_{t=0}^{T-1} \eta_t^2 \frac{1}{n} \mathbb{E} \|\mathbf{U}^t - \bar{\mathbf{U}}^t\|^2. \quad (\text{B.2})$$

Second, by applying Lemma B.2 and B.3 again we get

$$\sum_{t=0}^{T-1} \eta_t^2 \frac{1}{n} \mathbb{E} \|\mathbf{U}^t - \bar{\mathbf{U}}^t\|^2 \leq \tilde{\rho} \sum_{t=0}^{T-1} \eta_t^2 \frac{1}{n} \mathbb{E} \|\mathbf{V}^{t+1} - \mathbf{V}^t\|^2. \quad (\text{B.3})$$

Now we inspect the term $\mathbf{V}^{t+1} - \mathbf{V}^t$ following [54]. We first have

$$\begin{aligned} \mathbf{V}^{t+1} - \mathbf{V}^t &= \mathbf{V}^{t+1} - \mathbb{E}[\mathbf{V}^{t+1} \mid \mathcal{F}_t] - (\mathbf{V}^t - \mathbb{E}[\mathbf{V}^t \mid \mathcal{F}_{t-1}]) \\ &\quad + \mathbb{E}[\mathbf{V}^{t+1} \mid \mathcal{F}_t] - \nabla \mathbf{F}(\bar{x}^{t+1}) + \nabla \mathbf{F}(\bar{x}^{t+1}) - \nabla \mathbf{F}(\bar{x}^t) + \nabla \mathbf{F}(\bar{x}^t) - \mathbb{E}[\mathbf{V}^t \mid \mathcal{F}_{t-1}] \end{aligned}$$

where we use the notation $\nabla \mathbf{F}(x) := [\nabla f_1(x), \dots, \nabla f_n(x)]$ being the matrix of column gradient vectors. We thus have

$$\begin{aligned} &\mathbb{E} \|\mathbf{V}^{t+1} - \mathbf{V}^t\|^2 \\ &\leq 5 \left\{ \mathbb{E} \|\mathbf{V}^{t+1} - \mathbb{E}[\mathbf{V}^{t+1} \mid \mathcal{F}_t]\|^2 + \mathbb{E} \|\mathbf{V}^t - \mathbb{E}[\mathbf{V}^t \mid \mathcal{F}_{t-1}]\|^2 + \sum_{i=1}^n \mathbb{E} \|\nabla f_i(x_i^{t+1}) - \nabla f_i(\bar{x}^{t+1})\|^2 \right. \\ &\quad \left. + \sum_{i=1}^n \mathbb{E} \|\nabla f_i(\bar{x}^{t+1}) - \nabla f_i(\bar{x}^t)\|^2 + \sum_{i=1}^n \mathbb{E} \|\nabla f_i(x_i^t) - \nabla f_i(\bar{x}^t)\|^2 \right\} \leq 10n\sigma^2 + 60nG^2 \end{aligned}$$

where the first inequality uses Cauchy-Schwarz inequality, and the second utilizes Lipschitz continuity of each f_i . Plug this back to (B.3) gives the result. For $k > 0$ we can get the result in the exact same manner. $\square$

Now we are ready to present our final convergence for Algorithm 1, which we restate it here:

Theorem B.5. *Suppose Assumptions 3.1, 3.3 and 3.2 hold, also take $\eta_t = \eta T^{-1/2}$ for $\eta > 0$, the update of Algorithm 1 satisfies:*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 \leq \mathcal{O} \left(\frac{\Delta_0/\eta + (L\sigma^2/n + LG^2)\eta}{\sqrt{T}} + \frac{\tilde{\rho}L^2\eta^2}{T}(\sigma^2 + G^2) \right).$$

If we take $(\eta_0 = 0)$ $\eta_t = \eta t^{-1/2}$ for $\eta > 0$, the update of Algorithm 1 satisfies:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 \leq \tilde{\mathcal{O}} \left(\frac{\Delta_0/\eta + (L\sigma^2/n + LG^2)\eta}{\sqrt{T}} + \frac{\tilde{\rho}L^2\eta^2}{T}(\sigma^2 + G^2) \right).$$

Note that we hide higher-order terms in $\mathcal{O}$ and log terms in $\tilde{\mathcal{O}}$.

Proof. From Lemma B.1 we know that

$$\frac{3\eta_t}{4} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 \leq \mathbb{E}[f(\bar{x}^t)] - \mathbb{E}[f(\bar{x}^{t+1})] + 2\eta_t^2 L \frac{\sigma^2}{n} + (\eta_t + 2\eta_t^2 L) \frac{L^2}{n} \mathbb{E} \|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 + \eta_t^2 LG^2$$

and sum up the above equation from $t = 0$ to $T - 1$ gives

$$\begin{aligned} \sum_{t=0}^{T-1} \frac{3\eta_t}{4} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 &\leq \mathbb{E}[f(\bar{x}^0)] - f^* + \frac{2L\sigma^2}{n} \sum_{t=0}^{T-1} \eta_t^2 + \frac{L^2}{n} \sum_{t=0}^{T-1} (\eta_t + 2\eta_t^2 L) \mathbb{E} \|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 + LG^2 \sum_{t=0}^{T-1} \eta_t^2 \\ &\leq \mathbb{E}[f(\bar{x}^0)] - f^* + \frac{2L\sigma^2}{n} \sum_{t=0}^{T-1} \eta_t^2 + 10\tilde{\rho}L^2 \sum_{t=0}^{T-1} (\eta_t^3 + 2\eta_t^4 L)(\sigma^2 + G^2) + LG^2 \sum_{t=0}^{T-1} \eta_t^2 \end{aligned}$$

where we used Lemma B.4 for the second line.

Now for the constant stepsize $\eta_t = \eta T^{-1/2}$, it's very straightforward to check that $\sum_t \eta_t^2 = \eta^2$, $\sum_t \eta_t^3 = \eta^3/\sqrt{T}$ and $\sum_t \eta_t^4 = \eta^4/T$, therefore we get the following convergence result:

$$\sum_{t=0}^{T-1} \frac{3\eta}{4\sqrt{T}} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 \leq \Delta_0 + \left(\frac{2L\sigma^2}{n} + LG^2\right)\eta^2 + 10\tilde{\rho}L^2\left(\frac{\eta^3}{\sqrt{T}} + 2L\frac{\eta^4}{T}\right)(\sigma^2 + G^2)$$

i.e.

$$\frac{\eta}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 \leq \mathcal{O}\left(\frac{\Delta_0}{\sqrt{T}} + \frac{(L\sigma^2/n + LG^2)\eta^2}{\sqrt{T}} + \tilde{\rho}L^2\left(\frac{\eta^3}{T} + L\frac{\eta^4}{T^{3/2}}\right)(\sigma^2 + G^2)\right)$$

where $\Delta_0 := \mathbb{E}[f(\bar{x}^0)] - f^*$. This gives the first result in the theorem.

For the diminishing stepsize ($\eta_0 = 0$) $\eta_t = \eta t^{-1/2}$ for $\eta > 0$, it's again very straightforward to check that $\sum_t \eta_t^2 \leq \eta^2 \log(T)$, $\sum_t \eta_t^3 \leq \eta^3/\sqrt{T}$ and $\sum_t \eta_t^4 \leq \eta^4/T$, therefore we get the following convergence result:

$$\sum_{t=0}^{T-1} \frac{3\eta}{4\sqrt{T}} \mathbb{E} \|\nabla f(\bar{x}^t)\|^2 \leq \Delta_0 + \left(\frac{2L\sigma^2}{n} + LG^2\right)\eta^2 \log(T) + 10\tilde{\rho}L^2\left(\frac{\eta^3}{\sqrt{T}} + 2L\frac{\eta^4}{T}\right)(\sigma^2 + G^2)$$

which results in the second line of the result. $\square$

B.2 Parameter-free convergence theory for D-NASA

Our analysis depends on a key observation over the control of the consecutive consensus error. By the update of Algorithm 2 one can get:

$$\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}^{t+1}\|^2 \leq \frac{1+\rho}{2} \|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 + \eta_t^2 \frac{1+\rho^2}{1-\rho^2} \|\hat{\mathbf{Z}}^t - \bar{\mathbf{Z}}^t\|^2,$$

where $\hat{\mathbf{Z}}^t := \left[\frac{z_1^t}{\|z_1^t\|}, \dots, \frac{z_n^t}{\|z_n^t\|}\right]$ is the collection of column vectors of normalized z_i^t . Now the key observation is that the consensus error of $\hat{\mathbf{Z}}^t$ is always bounded:

$$\|\hat{\mathbf{Z}}^t - \bar{\mathbf{Z}}^t\|^2 = \sum_{i=1}^n \left\| \frac{z_i^t}{\|z_i^t\|} - \frac{1}{n} \sum_{i=1}^n \frac{z_i^t}{\|z_i^t\|} \right\|^2 \leq n.$$

Therefore the consecutive consensus error for $\mathbf{X}$ becomes:

$$\frac{1}{n} \|\mathbf{X}^{t+1} - \bar{\mathbf{X}}^{t+1}\|^2 \leq \frac{1+\rho}{2} \frac{1}{n} \|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 + \frac{1+\rho^2}{1-\rho^2} \eta_t^2,$$

and the cumulative consensus error is controlled directly by our stepsize choice η_t . This indicates that a careful stepsize choice will result in a bounded consensus error, regardless of any problem parameter.

From the update of Algorithm 2 we have that:

$$\begin{aligned} \mathbf{X}^{t+1} &= \mathbf{X}^t - \eta_t \hat{\mathbf{Z}}^t, \quad \bar{x}^{t+1} = \bar{x}^t - \frac{1}{n} \sum_{i=1}^n \frac{\eta_t}{\|z_i^t\|} z_i^t, \\ \mathbf{Z}^{t+1} &= (1 - \alpha_t) \mathbf{Z}^t W + \alpha_t \mathbf{U}^{t+1} W, \quad \bar{z}^{t+1} = (1 - \alpha_t) \bar{z}^t + \alpha_t \bar{u}^{t+1}, \\ \bar{u}^{t+1} &= \bar{v}^{t+1} = \frac{1}{n} \sum_{i=1}^n \nabla F_i(x_i^t, \xi_i^t), \end{aligned} \tag{B.4}$$

where

$$\hat{\mathbf{Z}}^t := \left[\frac{z_1^t}{\|z_1^t\|}, \dots, \frac{z_n^t}{\|z_n^t\|} \right]$$

is the collections of column vectors where each column is normalized z_i^t .

The following descent lemma characterizes the difference between the function value of two consecutive iterates for Algorithm 2:

Lemma B.6. *Suppose Assumption 3.1 holds. Algorithm 2 satisfies:*

$$f(\bar{x}^{t+1}) - f(\bar{x}^t) \leq -\eta_t \|\nabla f(\bar{x}^t)\| + 2\eta_t \|\bar{z}^t - \nabla f(\bar{x}^t)\| + \frac{\eta_t}{n} \sum_{i=1}^n \|z_i^t - \bar{z}^t\| + \frac{\eta_t^2 L}{2}. \quad (\text{B.5})$$

Proof. By the L -Lipschitz smooth of f (Assumption 3.1) we get:

$$\begin{aligned} f(\bar{x}^{t+1}) - f(\bar{x}^t) &\leq \nabla f(\bar{x}^t)^\top (\bar{x}^{t+1} - \bar{x}^t) + \frac{L}{2} \|\bar{x}^{t+1} - \bar{x}^t\|^2 \\ &= -\eta_t \nabla f(\bar{x}^t)^\top \left(\frac{1}{n} \sum_{i=1}^n \frac{z_i^t}{\|z_i^t\|} \right) + \frac{\eta_t^2 L}{2} \left\| \frac{1}{n} \sum_{i=1}^n \frac{z_i^t}{\|z_i^t\|} \right\|^2 \\ &\leq -\eta_t \nabla f(\bar{x}^t)^\top \left(\frac{1}{n} \sum_{i=1}^n \frac{z_i^t}{\|z_i^t\|} \right) + \frac{\eta_t^2 L}{2} \\ &= -\eta_t (\nabla f(\bar{x}^t) - \bar{z}^t)^\top \left(\frac{1}{n} \sum_{i=1}^n \frac{z_i^t}{\|z_i^t\|} \right) - \eta_t (\bar{z}^t)^\top \left(\frac{1}{n} \sum_{i=1}^n \frac{z_i^t}{\|z_i^t\|} \right) + \frac{\eta_t^2 L}{2} \\ &= -\eta_t (\nabla f(\bar{x}^t) - \bar{z}^t)^\top \left(\frac{1}{n} \sum_{i=1}^n \frac{z_i^t}{\|z_i^t\|} \right) - \eta_t (\bar{z}^t)^\top \left(\frac{1}{n} \sum_{i=1}^n \frac{z_i^t}{\|z_i^t\|} - \frac{\bar{z}^t}{\|\bar{z}^t\|} \right) - \eta_t \|\bar{z}^t\| + \frac{\eta_t^2 L}{2} \\ &\leq 2\eta_t \|\nabla f(\bar{x}^t) - \bar{z}^t\| - \eta_t \|\nabla f(\bar{x}^t)\| + \eta_t \|\bar{z}^t\| \left\| \frac{1}{n} \sum_{i=1}^n \frac{z_i^t}{\|z_i^t\|} - \frac{\bar{z}^t}{\|\bar{z}^t\|} \right\| + \frac{\eta_t^2 L}{2} \end{aligned}$$

where the second and third inequalities are by Cauchy-Schwarz inequality. It remains to bound the second last term in the last line. We have

$$\begin{aligned} \|\bar{z}^t\| \left\| \frac{1}{n} \sum_{i=1}^n \frac{z_i^t}{\|z_i^t\|} - \frac{\bar{z}^t}{\|\bar{z}^t\|} \right\| &= \frac{\|\bar{z}^t\|}{n} \left\| \sum_{i=1}^n \frac{\|\bar{z}^t\| - \|z_i^t\|}{\|\bar{z}^t\| \|z_i^t\|} z_i^t \right\| \\ &\leq \frac{\|\bar{z}^t\|}{n} \sum_{i=1}^n \frac{|\|\bar{z}^t\| - \|z_i^t\||}{\|\bar{z}^t\| \|z_i^t\|} \|z_i^t\| = \frac{1}{n} \sum_{i=1}^n |\|\bar{z}^t\| - \|z_i^t\|| \leq \frac{1}{n} \sum_{i=1}^n \|z_i^t - \bar{z}^t\| \end{aligned}$$

which concludes the proof. $\square$

We have the following dual convergence.

Lemma B.7. *We have*

$$\bar{z}^{t+1} - \nabla f(\bar{x}^{t+1}) = (1 - \alpha_t)(\bar{z}^t - \nabla f(\bar{x}^t)) + \alpha_t(\delta_1^t + \delta_2^t + \delta_3^t), \quad (\text{B.6})$$

where

$$\begin{aligned}\delta_1^t &= \frac{\nabla f(\bar{x}^t) - \nabla f(\bar{x}^{t+1})}{\alpha_t}, \\ \delta_2^t &= \frac{1}{n} \sum_{i=1}^n \nabla f_i(x_i^t) - \nabla f(\bar{x}^t), \\ \delta_3^t &= \frac{1}{n} \sum_{i=1}^n (v_i^t - \nabla f_i(x_i^t)).\end{aligned}$$

Consequently, we get:

$$\mathbb{E}\|\bar{z}^t - \nabla f(\bar{x}^t)\| \leq L \sum_{\tau=1}^t \beta_{(\tau+1):t} \eta_\tau + L \sum_{\tau=0}^t \beta_{(\tau+1):t} \alpha_\tau \sqrt{\frac{1}{n} \sum_{i=1}^n \mathbb{E}\|x_i^\tau - \bar{x}^\tau\|^2} + \sigma \sqrt{\frac{1}{n} \sum_{\tau=0}^t \beta_{(\tau+1):t}^2 \alpha_\tau^2} \quad (\text{B.7})$$

where we have the following conventions:

$$\beta_t := 1 - \alpha_t \quad \text{and} \quad \beta_{a:b} := \prod_{i=a}^b \beta_i.$$

Proof. By the update we know that $\bar{u}^t = \bar{v}^t = \frac{1}{n} \sum_{i=1}^n v_i^t$, thus

$$\begin{aligned}\bar{z}^{t+1} - \nabla f(\bar{x}^{t+1}) &= (1 - \alpha_t) \bar{z}^t + \alpha_t \bar{u}^t - \nabla f(\bar{x}^{t+1}) \\ &= (1 - \alpha_t) (\bar{z}^t - \nabla f(\bar{x}^t)) + \alpha_t (\delta_1^t + \delta_2^t + \delta_3^t)\end{aligned}$$

Now repeat the above recursive relation we get

$$\begin{aligned}\bar{z}^t - \nabla f(\bar{x}^t) &= (1 - \alpha_{t-1}) (\bar{z}^{t-1} - \nabla f(\bar{x}^{t-1})) + \alpha_{t-1} (\delta_1^{t-1} + \delta_2^{t-1} + \delta_3^{t-1}) \\ &= \dots \\ &= \beta_{1:t} (\bar{z}^0 - \nabla f(\bar{x}^0)) + \sum_{\tau=1}^t \beta_{(\tau+1):t} \alpha_\tau (\delta_1^\tau + \delta_2^\tau + \delta_3^\tau) \\ &= \sum_{\tau=1}^t \beta_{(\tau+1):t} \alpha_\tau \delta_1^\tau + \sum_{\tau=1}^t \beta_{(\tau+1):t} \alpha_\tau \delta_2^\tau + \sum_{\tau=0}^t \beta_{(\tau+1):t} \alpha_\tau \delta_3^\tau.\end{aligned}$$

Therefore we get

$$\begin{aligned}\mathbb{E}\|\bar{z}^t - \nabla f(\bar{x}^t)\| &\leq \sum_{\tau=1}^t \beta_{(\tau+1):t} \alpha_\tau \mathbb{E}\|\delta_1^\tau\| + \sum_{\tau=1}^t \beta_{(\tau+1):t} \alpha_\tau \mathbb{E}\|\delta_2^\tau\| + \mathbb{E}\left\| \sum_{\tau=0}^t \beta_{(\tau+1):t} \alpha_\tau \delta_3^\tau \right\| \\ &\leq L \sum_{\tau=1}^t \beta_{(\tau+1):t} \eta_\tau + L \sum_{\tau=0}^t \beta_{(\tau+1):t} \alpha_\tau \frac{1}{n} \sum_{i=1}^n \mathbb{E}\|x_i^\tau - \bar{x}^\tau\| + \sqrt{\sum_{\tau=0}^t \beta_{(\tau+1):t}^2 \alpha_\tau^2 \mathbb{E}\|\delta_3^\tau\|^2} \\ &\leq L \sum_{\tau=1}^t \beta_{(\tau+1):t} \eta_\tau + L \sum_{\tau=0}^t \beta_{(\tau+1):t} \alpha_\tau \sqrt{\frac{1}{n} \sum_{i=1}^n \mathbb{E}\|x_i^\tau - \bar{x}^\tau\|^2} + \sigma \sqrt{\frac{1}{n} \sum_{\tau=0}^t \beta_{(\tau+1):t}^2 \alpha_\tau^2}\end{aligned}$$

where the second term is due to smoothness of each f_i , also the last term is by $\mathbb{E}[\langle \delta_3^{\tau_1}, \delta_3^{\tau_2} \rangle] =$

0 for any $\tau_1 \neq \tau_2$ (due to unbiased assumption) and

$$\begin{aligned} \sqrt{\sum_{\tau=0}^t \beta_{(\tau+1):t}^2 \alpha_\tau^2 \mathbb{E} \|\delta_3^\tau\|^2} &= \sqrt{\sum_{\tau=0}^t \beta_{(\tau+1):t}^2 \alpha_\tau^2 \frac{1}{n^2} \sum_{i=1}^n \mathbb{E} \|\nabla F_i(x_i^t, \xi_i^t) - \nabla f_i(x_i^t)\|^2} \\ &\leq \sqrt{\sum_{\tau=0}^t \beta_{(\tau+1):t}^2 \alpha_\tau^2 \frac{\sigma^2}{n}} \end{aligned}$$

since all the cross inner-product terms vanish due to unbiased assumption. $\square$

We have the following lemma about the consensus error, that is, the average distance of each node to the global average.

Lemma B.8. *For the update of Algorithm 2, we have:*

$$\begin{aligned} \|\mathbf{X}^{t+1} - \bar{\mathbf{X}}^{t+1}\|^2 &\leq \frac{1+\rho}{2} \|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 + \eta_t^2 \frac{1+\rho^2}{1-\rho^2} \|\hat{\mathbf{Z}}^t - \bar{\mathbf{Z}}^t\|^2, \\ \|\hat{\mathbf{Z}}^t - \bar{\mathbf{Z}}^t\|^2 &\leq n, \\ \|\mathbf{Z}^{t+1} - \bar{\mathbf{Z}}^{t+1}\|^2 &\leq \frac{1+\rho}{2} \|\mathbf{Z}^t - \bar{\mathbf{Z}}^t\|^2 + \alpha_t^2 \frac{1+\rho^2}{1-\rho^2} \|\mathbf{U}^t - \bar{\mathbf{U}}^t\|^2, \\ \|\mathbf{U}^{t+1} - \bar{\mathbf{U}}^{t+1}\|^2 &\leq \frac{1+\rho}{2} \|\mathbf{U}^t - \bar{\mathbf{U}}^t\|^2 + \frac{1+\rho^2}{1-\rho^2} \|\mathbf{V}^{t+1} - \mathbf{V}^t\|^2, \\ \mathbb{E} \|\mathbf{V}^{t+1} - \mathbf{V}^t\|^2 &\leq 10n\sigma^2 + 5nL^2 + 5L^2 \mathbb{E} \left[\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}^{t+1}\|^2 + \|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 \right]. \end{aligned}$$

Proof. Since

$$\begin{aligned} \|\mathbf{X}^{t+1} - \bar{\mathbf{X}}^{t+1}\|^2 &= \|(\mathbf{X}^t - \eta_t \hat{\mathbf{Z}}^t)W - (\bar{x}^t - \eta_t \bar{z}^t)\mathbf{1}^\top\|^2 \\ &= \|(\mathbf{X}^t - \eta_t \hat{\mathbf{Z}}^t)W - \frac{1}{n}(\mathbf{X}^t - \eta_t \hat{\mathbf{Z}}^t)\mathbf{1}\mathbf{1}^\top\|^2 = \|(\mathbf{A}^t - \mathbf{A}^t \frac{\mathbf{1}\mathbf{1}^\top}{n})(W - \frac{\mathbf{1}\mathbf{1}^\top}{n})\|^2 \\ &\leq \|\mathbf{A}^t - \mathbf{A}^t \frac{\mathbf{1}\mathbf{1}^\top}{n}\|^2 \|W - \frac{\mathbf{1}\mathbf{1}^\top}{n}\|_2^2 \\ &\leq \rho^2 \|\mathbf{A}^t - \mathbf{A}^t \frac{\mathbf{1}\mathbf{1}^\top}{n}\|^2 = \rho^2 \|(\mathbf{X}^t - \bar{x}^t \mathbf{1}^\top) - \eta_t(\hat{\mathbf{Z}}^t - \bar{z}^t \mathbf{1}^\top)\|^2 \\ &\leq \rho^2 (1 + \frac{1}{c}) \|\mathbf{X}^t - \bar{x}^t \mathbf{1}^\top\|^2 + \rho^2 \eta_t^2 (1+c) \|\hat{\mathbf{Z}}^t - \bar{z}^t \mathbf{1}^\top\|^2 \end{aligned}$$

where $\mathbf{A}^t := \mathbf{X}^t - \eta_t \hat{\mathbf{Z}}^t$. Taking $c = \frac{2\rho^2}{1-\rho^2} \geq 0$ gives the desired result. For the consensus error of $\mathbf{Z}^t$ and $\mathbf{U}^t$ we get it in similar ways. It remains to bound the consensus error of $\hat{\mathbf{Z}}^t$ and the term $\mathbf{V}^{t+1} - \mathbf{V}^t$. For the consensus error of $\hat{\mathbf{Z}}^t$, we have

$$\|\hat{\mathbf{Z}}^t - \bar{\mathbf{Z}}^t\|^2 = \sum_{i=1}^n \left\| \frac{z_i^t}{\|z_i^t\|} - \frac{1}{n} \sum_{i=1}^n \frac{z_i^t}{\|z_i^t\|} \right\|^2 \leq \sum_{i=1}^n \left\| \frac{z_i^t}{\|z_i^t\|} \right\|^2 \leq n$$

where we use

$$\frac{1}{n} \sum_{i=1}^n \|v^i\|^2 - \frac{1}{n} \sum_{i=1}^n v^i \|^2 = \frac{1}{n} \sum_{i=1}^n \|v^i\|^2 - \left\| \frac{1}{n} \sum_{i=1}^n v^i \right\|^2 \leq \frac{1}{n} \sum_{i=1}^n \|v^i\|^2$$

for any sequence of vectors $v^1, \dots, v^n$.

Now we inspect the term $\mathbf{V}^{t+1} - \mathbf{V}^t$ similar to the proof of Lemma B.4. We again have

$$\begin{aligned} \mathbf{V}^{t+1} - \mathbf{V}^t &= \mathbf{V}^{t+1} - \mathbb{E}[\mathbf{V}^{t+1} \mid \mathcal{F}_t] - (\mathbf{V}^t - \mathbb{E}[\mathbf{V}^t \mid \mathcal{F}_{t-1}]) \\ &\quad + \mathbb{E}[\mathbf{V}^{t+1} \mid \mathcal{F}_t] - \nabla \mathbf{F}(\bar{x}^{t+1}) + \nabla \mathbf{F}(\bar{x}^{t+1}) - \nabla \mathbf{F}(\bar{x}^t) + \nabla \mathbf{F}(\bar{x}^t) - \mathbb{E}[\mathbf{V}^t \mid \mathcal{F}_{t-1}] \end{aligned}$$

where we use the notation $\nabla \mathbf{F}(x) := [\nabla f_1(x), \dots, \nabla f_n(x)]$ being the matrix of column gradient vectors. We thus have

$$\begin{aligned} &\mathbb{E} \|\mathbf{V}^{t+1} - \mathbf{V}^t\|^2 \\ &\leq 5 \left\{ \mathbb{E} \|\mathbf{V}^{t+1} - \mathbb{E}[\mathbf{V}^{t+1} \mid \mathcal{F}_t]\|^2 + \mathbb{E} \|\mathbf{V}^t - \mathbb{E}[\mathbf{V}^t \mid \mathcal{F}_{t-1}]\|^2 + \sum_{i=1}^n \mathbb{E} \|\nabla f_i(x_i^{t+1}) - \nabla f_i(\bar{x}^{t+1})\|^2 \right. \\ &\quad \left. + \sum_{i=1}^n \mathbb{E} \|\nabla f_i(\bar{x}^{t+1}) - \nabla f_i(\bar{x}^t)\|^2 + \sum_{i=1}^n \mathbb{E} \|\nabla f_i(x_i^t) - \nabla f_i(\bar{x}^t)\|^2 \right\} \\ &\leq 5 \left(2n\sigma^2 + nL^2 + L^2 \mathbb{E} [\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}^{t+1}\|^2 + \|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2] \right) \end{aligned}$$

where the first inequality uses Cauchy-Schwarz inequality, and the second utilizes Lipschitz smoothness of each f_i also note that $\|\nabla f_i(\bar{x}^{t+1}) - \nabla f_i(\bar{x}^t)\| \leq L\|\bar{x}^t - \bar{x}^{t+1}\| \leq L$. $\square$

Now we are ready to analyze the cumulative consensus error for Algorithm 2 as follows:

Lemma B.9. *For the update of Algorithm 2, if decreasing sequences such that $0 \leq \alpha_{t+1} \leq \alpha_t$ and $0 \leq \eta_{t+1} \leq \eta_t$, we have:*

$$\begin{aligned} &\sum_{t=0}^{T-1} \frac{1}{n} \mathbb{E} \|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 \leq \tilde{\rho} \sum_{t=0}^{T-1} \eta_t^2, \\ &\sum_{\tau=0}^t \alpha_\tau \sqrt{\frac{1}{n} \mathbb{E} \|\mathbf{X}^\tau - \bar{\mathbf{X}}^\tau\|^2} \leq \tilde{\rho} \sum_{\tau=0}^t \alpha_\tau \eta_\tau, \\ &\sum_{t=0}^{T-1} \frac{1}{n} \mathbb{E} \|\mathbf{Z}^t - \bar{\mathbf{Z}}^t\|^2 \leq \tilde{\rho}^2 (10\sigma^2 + 5L^2) \sum_{t=0}^{T-1} \alpha_t^2 + 2\tilde{\rho}^3 \sum_{t=0}^T \alpha_t^2 \eta_t^2, \\ &\sum_{\tau=0}^t \eta_\tau \sqrt{\frac{1}{n} \mathbb{E} \|\mathbf{Z}^\tau - \bar{\mathbf{Z}}^\tau\|^2} \leq \tilde{\rho}^2 \sqrt{10\sigma^2 + 5L^2} \sum_{\tau=0}^t \eta_\tau \alpha_\tau + 2\sqrt{5}L\tilde{\rho}^3 \sum_{\tau=0}^{t+1} \eta_\tau^2 \alpha_\tau, \end{aligned}$$

where

$$\tilde{\rho} := \max \left\{ \frac{1}{1 - \sqrt{\frac{1+\rho}{2}}} \sqrt{\frac{1+\rho^2}{1-\rho^2}}, \frac{\rho}{1-\rho} \frac{1+\rho^2}{1-\rho^2} \right\}.$$

Note that $\tilde{\rho}$ is greater than 0.

Proof. The first line is by Lemma B.8 and B.3 by taking $a_\tau = \frac{1}{n} \mathbb{E} \|\mathbf{X}^\tau - \bar{\mathbf{X}}^\tau\|^2$, $b_\tau = \eta_\tau^2(1+\rho^2)/(1-\rho^2)$, $c_\tau = 1$ and $r = (1+\rho)/2$ in Lemma B.3 directly.

For the second line, by Lemma B.8 we get

$$\begin{aligned}
\sqrt{\frac{1}{n}\mathbb{E}\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}^{t+1}\|^2} &\leq \sqrt{\frac{1+\rho}{2}\frac{1}{n}\mathbb{E}\|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 + \eta_t^2\frac{1+\rho^2}{1-\rho^2}\frac{1}{n}\mathbb{E}\|\hat{\mathbf{Z}}^t - \bar{\mathbf{Z}}^t\|^2} \\
&\leq \sqrt{\frac{1+\rho}{2}}\sqrt{\frac{1}{n}\mathbb{E}\|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2} + \eta_t\sqrt{\frac{1+\rho^2}{1-\rho^2}}\sqrt{\frac{1}{n}\mathbb{E}\|\hat{\mathbf{Z}}^t - \bar{\mathbf{Z}}^t\|^2} \\
&\leq \sqrt{\frac{1+\rho}{2}}\sqrt{\frac{1}{n}\mathbb{E}\|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2} + \eta_t\sqrt{\frac{1+\rho^2}{1-\rho^2}}
\end{aligned}$$

where the second inequality is by $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and third is by $\|\hat{\mathbf{Z}}^t - \bar{\mathbf{Z}}^t\|^2 \leq n$. Now taking $a_\tau = \sqrt{\frac{1}{n}\mathbb{E}\|\mathbf{X}^\tau - \bar{\mathbf{X}}^\tau\|^2}$, $b_\tau = \eta_\tau\sqrt{(1+\rho^2)/(1-\rho^2)}$, $c_\tau = \alpha_\tau$ and $r = \sqrt{(1+\rho)/2}$ as in Lemma B.3 will give the first line of the result. Note that here $a_0 = 0$ due to the initialization of our algorithm.

Now to the third line, again by Lemma B.8 we get

$$\begin{aligned}
\sum_{t=0}^{T-1} \frac{1}{n}\mathbb{E}\|\mathbf{Z}^t - \bar{\mathbf{Z}}^t\|^2 &\leq \tilde{\rho} \sum_{t=0}^{T-1} \alpha_t^2 \frac{1}{n}\mathbb{E}\|\mathbf{U}^t - \bar{\mathbf{U}}^t\|^2 \\
\sum_{t=0}^{T-1} \alpha_t^2 \frac{1}{n}\mathbb{E}\|\mathbf{U}^t - \bar{\mathbf{U}}^t\|^2 &\leq \tilde{\rho} \sum_{t=0}^{T-1} \alpha_t^2 \frac{1}{n}\mathbb{E}\|\mathbf{V}^{t+1} - \mathbf{V}^t\|^2
\end{aligned}$$

by using Lemma B.8 for two times. Also since

$$\begin{aligned}
\sum_{t=0}^{T-1} \alpha_t^2 \frac{1}{n}\mathbb{E}\|\mathbf{V}^{t+1} - \mathbf{V}^t\|^2 &\leq (10\sigma^2 + 5L^2) \sum_{t=0}^{T-1} \alpha_t^2 + 2 \sum_{t=0}^T \alpha_t^2 \frac{1}{n}\mathbb{E}\|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 \\
\sum_{t=0}^T \alpha_t^2 \frac{1}{n}\mathbb{E}\|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 &\leq \tilde{\rho} \sum_{t=0}^T \alpha_t^2 \eta_t^2
\end{aligned}$$

where for the last line above we again use Lemma B.8. Combining all above equations gives the second line of the theorem.

As for the forth line, note that from Lemma B.8 and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, we have:

$$\begin{aligned}
\sqrt{\frac{1}{n}\mathbb{E}\|\mathbf{Z}^{t+1} - \bar{\mathbf{Z}}^{t+1}\|^2} &\leq \sqrt{\frac{1+\rho}{2}}\sqrt{\frac{1}{n}\mathbb{E}\|\mathbf{Z}^t - \bar{\mathbf{Z}}^t\|^2} + \alpha_t\sqrt{\frac{1+\rho^2}{1-\rho^2}}\sqrt{\frac{1}{n}\mathbb{E}\|\mathbf{U}^t - \bar{\mathbf{U}}^t\|^2}, \\
\sqrt{\frac{1}{n}\mathbb{E}\|\mathbf{U}^{t+1} - \bar{\mathbf{U}}^{t+1}\|^2} &\leq \sqrt{\frac{1+\rho}{2}}\sqrt{\frac{1}{n}\mathbb{E}\|\mathbf{U}^t - \bar{\mathbf{U}}^t\|^2} + \sqrt{\frac{1+\rho^2}{1-\rho^2}}\sqrt{\frac{1}{n}\mathbb{E}\|\mathbf{V}^{t+1} - \mathbf{V}^t\|^2}, \\
\sqrt{\frac{1}{n}\mathbb{E}\|\mathbf{V}^{t+1} - \mathbf{V}^t\|^2} &\leq \sqrt{10\sigma^2 + 5L^2} + \sqrt{5}L \left[\sqrt{\frac{1}{n}\mathbb{E}\|\mathbf{X}^{t+1} - \bar{\mathbf{X}}^{t+1}\|^2} + \sqrt{\frac{1}{n}\mathbb{E}\|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2} \right].
\end{aligned}$$

Repeating the proof of the third line gives the fourth line. $\square$

Now we are ready to show our final convergence for constant stepsizes, which we restate as follows:

Theorem B.10. *Suppose Assumptions 3.1 and 3.2 hold, also take $\alpha_t = \alpha T^{-1/2}$ and $\eta_t = \eta T^{-3/4}$ for any $\alpha, \eta > 0$, the update of Algorithm 2 satisfies:*

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\| &\leq \mathcal{O} \left(\left(\frac{\Delta_0}{\eta} + \frac{2L\eta}{\alpha} + \frac{2\sigma\sqrt{\alpha}}{\sqrt{n}} \right) \frac{1}{T^{1/4}} + \frac{\tilde{\rho}^2 \sqrt{10\sigma^2 + 5L^2\alpha}}{T^{1/2}} \right), \\ \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\bar{z}^t - \nabla f(\bar{x}^t)\| &\leq \mathcal{O} \left(\left(\frac{2L\eta}{\alpha} + \frac{2\sigma\sqrt{\alpha}}{\sqrt{n}} \right) \frac{1}{T^{1/4}} + 2L\tilde{\rho}\eta \frac{1}{T^{1/2}} \right), \\ \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{n} [\|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 + \|\mathbf{Z}^t - \bar{\mathbf{Z}}^t\|^2] &\leq \mathcal{O} \left(\tilde{\rho} \frac{\eta}{T^{1/2}} + \tilde{\rho}^2 (10\sigma^2 + 5L^2) \frac{\alpha^4}{T} + 2\tilde{\rho}^3 \frac{\alpha^4 \eta^2}{T^{5/2}} \right). \end{aligned}$$

The above three bounds correspond to stationarity, approximation to gradient and consensus errors. Note that we hide some higher-order terms in $\mathcal{O}$.

Proof. By Lemma B.6, we have

$$\begin{aligned} \eta_t \mathbb{E} \|\nabla f(\bar{x}^t)\| &\leq \mathbb{E}[f(\bar{x}^t) - f(\bar{x}^{t+1})] + 2\eta_t \mathbb{E} \|\bar{z}^t - \nabla f(\bar{x}^t)\| + \mathbb{E} \left[\frac{\eta_t}{n} \sum_{i=1}^n \|z_i^t - \bar{z}^t\| \right] + \frac{\eta_t^2 L}{2} \\ &\leq \mathbb{E}[f(\bar{x}^t) - f(\bar{x}^{t+1})] + 2\eta_t \mathbb{E} \|\bar{z}^t - \nabla f(\bar{x}^t)\| + \eta_t \sqrt{\frac{1}{n} \sum_{i=1}^n \mathbb{E} \|z_i^t - \bar{z}^t\|^2} + \frac{\eta_t^2 L}{2} \end{aligned}$$

where in the second equality we used $\mathbb{E} X^2 \geq (\mathbb{E} X)^2$.

Now sum up from $t = 0$ to $T - 1$ and using Lemma B.6 and B.9, we get

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t \mathbb{E} \|\nabla f(\bar{x}^t)\| &\leq \Delta_0 + 2 \sum_{t=0}^{T-1} \eta_t \mathbb{E} \|\bar{z}^t - \nabla f(\bar{x}^t)\| + \sum_{t=0}^{T-1} \eta_t \sqrt{\frac{1}{n} \sum_{i=1}^n \mathbb{E} \|z_i^t - \bar{z}^t\|^2} + \frac{L}{2} \sum_{t=0}^{T-1} \eta_t^2 \\ &\leq \Delta_0 + 2 \sum_{t=0}^{T-1} \eta_t \left(L \sum_{\tau=1}^t \beta_{(\tau+1):t} \eta_\tau + L\tilde{\rho} \sum_{\tau=0}^t \beta_{(\tau+1):t} \alpha_\tau \eta_\tau + \sigma \sqrt{\frac{1}{n} \sum_{\tau=0}^t \beta_{(\tau+1):t}^2 \alpha_\tau^2} \right) \\ &\quad + \tilde{\rho}^2 \sqrt{10\sigma^2 + 5L^2} \sum_{t=0}^{T-1} \eta_t \alpha_t + 2\sqrt{5}L\tilde{\rho}^3 \sum_{t=0}^{T-1} \eta_t^2 \alpha_t + \frac{L}{2} \sum_{t=0}^{T-1} \eta_t^2 \end{aligned} \tag{B.8}$$

Now we inspect the each of the terms on the right hand side. By our choice of η_t and α_t , it's straightforward to verify that

$$\sum_{t=1}^T \alpha_t \eta_t = \frac{\alpha \eta}{T^{1/4}}, \quad \sum_{t=1}^T \alpha_t \eta_t^2 = \frac{\alpha \eta^2}{T}, \quad \sum_{t=1}^T \eta_t^2 = \frac{\eta^2}{T^{1/2}} \tag{B.9}$$

and

$$\begin{aligned}
\sum_{t=0}^{T-1} \eta_t \sum_{\tau=1}^t \beta_{(\tau+1):t} \eta_\tau &= \sum_{t=0}^{T-1} \frac{\eta}{T^{3/4}} \sum_{\tau=1}^t \left(1 - \frac{\alpha}{T^{1/2}}\right)^{t-\tau} \frac{\eta}{T^{3/4}} \\
&= \frac{\eta^2}{T^{3/2}} \sum_{t=0}^{T-1} \sum_{\tau=1}^t \left(1 - \frac{\alpha}{T^{1/2}}\right)^{t-\tau} \\
&= \frac{\eta^2}{T^{3/2}} \sum_{t=0}^{T-1} \left(1 - \frac{\alpha}{T^{1/2}}\right)^t \frac{T^{1/2}}{\alpha} \left[\left(1 - \frac{\alpha}{T^{1/2}}\right)^{-t+1} - 1\right] \\
&\leq \frac{\eta^2}{\alpha T} \sum_{t=0}^{T-1} \left(1 - \frac{\alpha}{T^{1/2}}\right) \leq \frac{\eta^2}{\alpha}
\end{aligned}$$

Similarly

$$\begin{aligned}
\sum_{t=0}^{T-1} \eta_t \sum_{\tau=0}^t \beta_{(\tau+1):t} \alpha_\tau \eta_\tau &\leq \frac{\eta^2}{T^{1/2}} \\
\sum_{t=0}^{T-1} \eta_t \sqrt{\sum_{\tau=0}^t \beta_{(\tau+1):t}^2 \alpha_\tau^2} &\leq \eta \sqrt{\alpha}
\end{aligned}$$

Now plugging everything back we get:

$$\begin{aligned}
\frac{\eta}{T^{3/4}} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\| &\leq \Delta_0 + 2L \frac{\eta^2}{\alpha} + 2L\tilde{\rho} \frac{\eta^2}{T^{1/2}} + 2 \frac{\sigma}{\sqrt{n}} \eta \sqrt{\alpha} \\
&\quad + \tilde{\rho}^2 \sqrt{10\sigma^2 + 5L^2} \frac{\alpha\eta}{T^{1/4}} + 2\sqrt{5}L\tilde{\rho}^3 \frac{\alpha\eta^2}{T} + \frac{L}{2} \frac{\eta^2}{T^{1/2}}
\end{aligned}$$

i.e.

$$\begin{aligned}
\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\| &\leq \left(\frac{\Delta_0}{\eta} + \frac{2L\eta}{\alpha} + \frac{2\sigma\sqrt{\alpha}}{\sqrt{n}}\right) \frac{1}{T^{1/4}} + (2L\tilde{\rho}\eta + \frac{L\eta}{2}) \frac{1}{T^{3/4}} \\
&\quad + \frac{\tilde{\rho}^2 \sqrt{10\sigma^2 + 5L^2} \alpha}{T^{1/2}} + \frac{2\sqrt{5}L\tilde{\rho}^3 \alpha^2 \eta}{T^{7/4}}
\end{aligned}$$

The first line of the theorem is obtained by neglecting the higher-order terms. Note that we also proved the second line of the theorem since (B.8) already contains the bound for $\|\bar{z}^t - \nabla f(\bar{x}^t)\|$.

It remains to bound the consensus error (third line), which follows directly from Lemma B.9 and (B.9). $\square$

We also present the result when we don't fix the total number of iterations in advance. We have the following useful technical lemma. Most of the result in this Lemma is from Lemma 11 in [20].

Lemma B.11. *Let $q \in (0, 1)$, $p \geq 0$ and $t > 0$. Further let positive integers a, b s.t. $2 \leq a \leq b$, then we have that for any $\alpha > 0$,*

$$\prod_{t=a}^b (1 - \alpha t^{-q}) \leq \exp \left(\frac{\alpha}{1-q} (a^{1-q} - b^{1-q}) \right).$$

If in addition $p \geq q$, we have

$$\sum_{t=a}^b t^{-p} \prod_{\tau=a}^t (1 - \alpha \tau^{-q}) \leq \frac{(a-1)^{q-p} \exp\left(\alpha \frac{a^{1-q} - (a-1)^{1-q}}{1-q}\right) - b^{q-p} \exp\left(\alpha \frac{a^{1-q} - b^{1-q}}{1-q}\right)}{(\alpha + (p-q)b^{q-1})}.$$

and in particular,

$$\sum_{t=a}^b t^{-p} \prod_{\tau=a}^t (1 - \alpha \tau^{-q}) \leq \frac{(a-1)^{q-p}}{\alpha} \exp\left(\alpha \frac{a^{1-q} - (a-1)^{1-q}}{1-q}\right) = \mathcal{O}\left(\frac{a^{q-p}}{\alpha}\right).$$

If further in addition $p < 1$, $\alpha \geq 2$, we have

$$\sum_{t=2}^b t^{-p} \prod_{\tau=t+1}^b (1 - \alpha \tau^{-q}) \leq \frac{2}{\alpha} \exp\left(\frac{\alpha}{1-q}\right) (b+1)^{q-p}$$

Proof. For the first equation we get

$$\prod_{t=a}^b (1 - \alpha t^{-q}) \leq \exp\left(-\alpha \sum_{\tau=a}^b t^{-q}\right) \leq \exp\left(-\alpha \int_a^{b+1} t^{-q} dt\right) = \exp\left(\frac{\alpha}{1-q} (a^{1-q} - (b+1)^{1-q})\right).$$

Now for the second line, using the above result we get

$$\begin{aligned} \sum_{t=a}^b t^{-p} \prod_{\tau=a}^t (1 - \alpha \tau^{-q}) &\leq \exp\left(\frac{\alpha a^{1-q}}{1-q}\right) \sum_{t=a}^b t^{-p} \exp\left(-\frac{\alpha(t+1)^{1-q}}{1-q}\right) \\ &\leq \exp\left(\frac{\alpha a^{1-q}}{1-q}\right) \int_{a-1}^b t^{-p} \exp\left(-\frac{\alpha t^{1-q}}{1-q}\right) dt \\ &= \exp\left(\frac{\alpha a^{1-q}}{1-q}\right) \int_{a-1}^b t^{-q} t^{q-p} \exp\left(-\frac{\alpha t^{1-q}}{1-q}\right) dt \end{aligned} \quad (\text{B.10})$$

The above integral can be calculated by integration by parts, specifically:

$$\begin{aligned} &\int_{a-1}^b t^{q-p} t^{-q} \exp\left(-\frac{\alpha t^{1-q}}{1-q}\right) dt \\ &= \left[-\frac{t^{q-p}}{\alpha} \exp\left(-\frac{\alpha t^{1-q}}{1-q}\right) \right]_{t=a-1}^{t=b} + (q-p) \int_{a-1}^b \frac{t^{q-p-1}}{\alpha} \exp\left(-\frac{\alpha t^{1-q}}{1-q}\right) dt. \end{aligned}$$

Finally, since the integrand is monotonically decreasing and $p \geq q$, we have

$$(q-p) \int_{a-1}^b t^{q-p-1} \exp\left(-\frac{\alpha t^{1-q}}{1-q}\right) dt \leq (q-p) b^{q-1} \int_{a-1}^b t^{-p} \exp\left(-\frac{\alpha t^{1-q}}{1-q}\right) dt$$

which is exactly the integral we started with. The results in the second and third lines are thus by rearranging terms.

Now for the last line of the lemma, and we use the similar technique as the second line,

both adopted from the proof of Lemma 10 of [20]. We have

$$\begin{aligned} \sum_{t=2}^b t^{-p} \prod_{\tau=t+1}^b (1 - \alpha \tau^{-q}) &\leq \exp \left(-\alpha \sum_{\tau=1}^b \tau^{-q} \right) \sum_{t=2}^b t^{-p} \exp \left(\alpha \sum_{\tau=1}^t \tau^{-q} \right) \\ &\leq \exp \left(-\alpha \int_1^{b+1} \tau^{-q} d\tau \right) \sum_{t=2}^b t^{-p} \exp \left(\alpha \int_0^t \tau^{-q} d\tau \right) \\ &\leq \exp \left(\alpha \frac{1 - (b+1)^{1-q}}{1-q} \right) \sum_{t=2}^b t^{-p} \exp \left(\alpha \frac{t^{1-q}}{1-q} \right) \end{aligned}$$

Note that the summation in the last line above is the same as (B.10). Repeat the proof of the second line gives the result. $\square$

Lemma B.12. Suppose we take $\alpha_t = \alpha t^{-1/2}$ and $\eta_t = \eta t^{-3/4}$ ($\alpha_0 = \eta_0 = 0$) with $0 < \alpha < 1$ and $\eta > 0$, then we have

$$\sum_{t=1}^T \alpha_t \eta_t \leq \mathcal{O}(\alpha \eta), \quad \sum_{t=1}^T \alpha_t \eta_t^2 \leq \mathcal{O}(\alpha \eta^2), \quad \sum_{t=1}^T \eta_t^2 \leq \mathcal{O}(\eta^2),$$

and

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t \sum_{\tau=1}^t \beta_{(\tau+1):t} \eta_\tau &\leq \mathcal{O} \left(\frac{\eta^2}{\alpha} \log(T) \right), \\ \sum_{t=0}^{T-1} \eta_t \sum_{\tau=0}^t \beta_{(\tau+1):t} \alpha_\tau \eta_\tau &\leq \mathcal{O}(\eta^2 \log(T)), \\ \sum_{t=0}^{T-1} \eta_t \sqrt{\sum_{\tau=0}^t \beta_{(\tau+1):t}^2 \alpha_\tau^2} &\leq \mathcal{O}(\sqrt{\alpha} \eta \log(T)), \end{aligned}$$

where $\beta_t = 1 - \alpha_t$.

Proof. The first line is directly by the integral test of the series in the form $\sum_t t^{-p}$ with $p > 1$.

The proof of the latter three resembles the proof of Lemma 11 in [20]. We prove the second line of this Lemma as a show case and refer to Lemma 11 in [20] for the detail of the proof of the last two lines. For the second line, we have

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t \sum_{\tau=1}^t \beta_{(\tau+1):t} \eta_\tau &= \eta^2 + \eta^2 \sum_{t=1}^{T-1} t^{-3/4} \sum_{\tau=2}^t \prod_{\xi=\tau+1}^t (1 - \alpha \xi^{-1/2}) \tau^{-3/4} \\ &\leq \eta^2 + \eta^2 \sum_{t=1}^{T-1} t^{-3/4} \frac{2}{\alpha} \exp(2\alpha) (t+1)^{-1/4} \\ &= \eta^2 + \frac{2\eta^2}{\alpha} \exp(2\alpha) \sum_{t=1}^{T-1} t^{-1} \leq \eta^2 + \frac{2\eta^2}{\alpha} \exp(2\alpha) \log(T) \end{aligned}$$

where we use Lemma B.11 for the first inequality. $\square$

Now we are ready to present the final convergence result. We restate the convergence theorem as follows:

Theorem B.13. Suppose Assumptions 3.1 and 3.2 hold, also take $\alpha_t = \alpha t^{-1/2}$ and $\eta_t = \eta t^{-3/4}$ for any $0 < \alpha < 1$ and $0 < \eta \leq 1$, the update of Algorithm 2 satisfies:

$$\begin{aligned} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\| &\leq \tilde{\mathcal{O}}\left(\frac{\Delta_0/\eta + L\eta/\alpha + L\tilde{\rho}\eta + \sigma\sqrt{\alpha}/\sqrt{n} + \tilde{\rho}^2(\sigma + L)\alpha + L\tilde{\rho}^3\eta\alpha + L\eta}{T^{1/4}}\right), \\ \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\bar{z}^t - \nabla f(\bar{x}^t)\| &\leq \tilde{\mathcal{O}}\left(\frac{L\eta/\alpha + L\tilde{\rho}\eta + \sigma\sqrt{\alpha}/\sqrt{n}}{T^{1/4}}\right), \\ \frac{1}{T} \sum_{t=0}^{T-1} \frac{1}{n} \mathbb{E} [\|\mathbf{X}^t - \bar{\mathbf{X}}^t\|^2 + \|\mathbf{Z}^t - \bar{\mathbf{Z}}^t\|^2] &\leq \mathcal{O}\left(\frac{\tilde{\rho}^2(\sigma^2 + L^2 + \tilde{\rho}\eta^2)}{T}\right). \end{aligned}$$

The above three bounds correspond to stationarity, approximation to gradient and consensus errors. Note that we hide logarithmic factors in $\tilde{\mathcal{O}}$.

Proof. Same as the proof of Theorem B.10, we get

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t \mathbb{E} \|\nabla f(\bar{x}^t)\| &\leq \Delta_0 + 2L \sum_{t=0}^{T-1} \eta_t \sum_{\tau=1}^t \beta_{(\tau+1):t} \eta_\tau + 2L\tilde{\rho} \sum_{t=0}^{T-1} \eta_t \sum_{\tau=0}^t \beta_{(\tau+1):t} \alpha_\tau \eta_\tau \\ &\quad + 2\frac{\sigma}{\sqrt{n}} \sum_{t=0}^{T-1} \eta_t \sqrt{\sum_{\tau=0}^t \beta_{(\tau+1):t}^2 \alpha_\tau^2} \\ &\quad + \tilde{\rho}^2 \sqrt{10\sigma^2 + 5L^2} \sum_{t=0}^{T-1} \eta_t \alpha_t + 2\sqrt{5}L\tilde{\rho}^3 \sum_{t=0}^T \eta_t^2 \alpha_t + \frac{L}{2} \sum_{t=0}^{T-1} \eta_t^2 \end{aligned}$$

Now we inspect the each of the terms on the right hand side. By our choice of η_t and α_t and using Lemma B.12 we get:

$$\sum_{t=0}^{T-1} \eta_t \mathbb{E} \|\nabla f(\bar{x}^t)\| \leq \tilde{\mathcal{O}}\left(\Delta_0 + L\frac{\eta^2}{\alpha} + L\tilde{\rho}\eta^2 + \frac{\sigma}{\sqrt{n}}\sqrt{\alpha}\eta + \tilde{\rho}^2(\sigma + L)\eta\alpha + L\tilde{\rho}^3\eta^2\alpha + L\eta^2\right)$$

where we hide the logarithmic factor in $\tilde{\mathcal{O}}$.

Now since $\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f(\bar{x}^t)\| \leq T^{-1/4} \sum_{t=0}^{T-1} t^{-3/4} \mathbb{E} \|\nabla f(\bar{x}^t)\|$, we yield the desired result in the theorem statement. $\square$

Manuscript received 26 October 2024
revised 2 February 2025
accepted for publication 5 February 2025

JIAXIANG LI

Department of Electrical and Computer Engineering
University of Minnesota, Twin Cities
E-mail address: li003755@umn.edu

XUXING CHEN

Department of Mathematics, University of California, Davis
E-mail address: xuxchen@ucdavis.edu

SHIQIAN MA

Department of Computational Applied Math and
Operations Research, Rice University
E-mail address: sqma@rice.edu

MINGYI HONG

Department of Electrical and Computer Engineering
University of Minnesota, Twin Cities
E-mail address: mhong@umn.edu