Problem-Parameter-Free Decentralized Nonconvex Stochastic Optimization

Jiaxiang Li * Xuxing Chen † Shiqian Ma ‡ Mingyi Hong §

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.

1 Introduction

Decentralized (distributed) optimization appears in many applications, such as machine learning (Lian et al., 2017; Tang et al., 2018b; Lian et al., 2018), robotics (Queralta et al., 2020), signal processing (Hong et al., 2015), and control systems (Nedić & Liu, 2018; Yang et al., 2019). 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 Lian et al. (2017); Tang et al. (2018b), 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 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

*Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities. li003755@umn.edu

†Department of Mathematics, University of California, Davis. xuxchen@ucdavis.edu

‡Department of Computational Applied Math and Operations Research, Rice University. Research supported in part by NSF grants DMS-2243650, CCF-2308597, CCF-2311275 and ECCS-2326591, and a startup fund from Rice University. sqma@rice.edu

§Department of Electrical and Computer Engineering, University of Minnesota, Twin Cities. mhong@umn.edu

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 (Yuan et al., 2022); (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 (Tang et al., 2018b; Koloskova et al., 2020).

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:

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

where each $f_i = \mathsf{E}_{\xi_i \sim \mathcal{D}_i}[F_i(x,\xi_i)]$ 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, "problem-parameter-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 Lian et al. (2017), 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 Xu et al. (2015); Di Lorenzo & Scutari (2016); Nedic et al. (2017); Qu & Li (2017); Pu & Nedić (2021); Koloskova et al. (2021); Liu et al. (2023)) 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 problem-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 parameter-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 problem-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 problem-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 the stepsize 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 that 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 $\mathbf{X}^t := [x_1^t, ..., x_n^t]$ which is the collection of local variables x_i^t at iteration t for $i = 1, ..., n$ 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 (Tsitsiklis, 1984; Ram et al., 2009; Yan et al., 2012; Yuan et al., 2016), their distinctive advantages, such as robustness, scalability and privacy preserving, in comparison to centralized setting like Li et al. (2014), were not well understood both theoretically and empirically until the case study conducted by Lian et al. (2017). Despite its great success in characterizing the superiority of decentralized training over the centralized setting, the analysis therein replies on a bounded gradient heterogeneity assumption, which was later removed by follow-up works such as D^2 (Tang et al., 2018b).

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 optimal solution even under the deterministic and strongly convex setting (Yuan et al., 2016). One important technique to mitigate this effect is gradient tracking, which was introduced in control community (Xu et al., 2015; Di Lorenzo & Scutari, 2016; Nedic et al., 2017; Qu & Li, 2017) to improve the convergence rate in the deterministic setting. Later this method was revealed to be helpful to remove the bounded gradient heterogeneity assumption (Zhang & You, 2019; Lu et al., 2019; Pu & Nedić, 2021; Koloskova et al., 2021) in convergence analysis. A more recent technique of moving-average updates (momentum) have been studied in both decentralized optimization and federated learning setting (Xiao et al., 2023; Cheng et al., 2023) to further improve the rate of convergence.

Table 1: Comparison of D-NASA (Algorithm 2) with some widely-used decentralized stochastic nonconvex optimization algorithms: D-SGD (Lian et al., 2017), D^2 (Tang et al., 2018b) and D-SGT (Koloskova et al., 2021). '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 $O(n^{1}\epsilon^{4})$ oracle calls to achieve an ϵ-stationary point.

Algorithm	Other Aspt	Info Required
D-SGD	Hetero	Smoothness, variance
D^2	None	Smoothness, net-topology
D-SGT	None	Smoothness, net-topology
D-NASA (ours)	None	None

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 (Di Lorenzo & Scutari, 2016) is able to achieve a problem-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 (Neglia et al., 2020; Koloskova et al., 2020), communication compression (Tang et al., 2018a; Koloskova et al., 2019), large-model training (Gan et al., 2021; Yuan et al., 2022), adaptive algorithms (Chen et al., 2023), to name a few.

Parameter-free optimization (Problem-) Parameter-free optimization refers to the algorithms that require no/few information needed 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 (Nesterov et al., 2018). 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 Malitsky & Mishchenko (2019) that adaptively estimates the local curvature information in each iteration and does not require the knowledge of L. See Malitsky & Mishchenko (2023); Latafat et al. (2023a,b); Li & Lan (2023); Zhou et al. (2024) for more recent works on this subject. Currently, these adaptive methods are for deterministic problems and it remains an interesting direction to extend them to stochastic and decentralized settings.

For stochastic gradient descent for solving convex problems, the current convergence result requires 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 (Garrigos & Gower, 2023). 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 Grazzi et al. (2020)), which indicates that optimal choices of stepsizes in SGD heavily depend on problem parameters.

For nonconvex stochastic optimization, various adaptive methods, such as AdaGrad (Duchi et al., 2011; McMahan & Streeter, 2010), AMSGrad-Norm Reddi et al. (2019), NSGD-M Cutkosky & Mehta (2020), are proved to be convergent without any knowledge of the parameters Faw et al. (2022); Yang et al. (2023); Hübler et al. (2023), 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 (Carmon & Hinder, 2022; Ivgi et al., 2023). These research results emphasize the optimal dependency on $kx^0 \quad x\ k$, i.e., the distance from the initial to the optimal point, and it is not clear how these works adapt to the nonconvex problems.

Parameter-free stochastic optimization in decentralized setting is unexplored. It is natural to ask whether one can achieve parameter-free decentralized training, given the unique challenges such as communication complexity and heterogeneous data distribution 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 (Lian et al., 2017), D^2 (Tang et al., 2018b) and D-SGT (Koloskova et al., 2021)[1]. In particular, D-SGD in Lian et al. (2017) requires the information of Lipschitz smoothness parameter and variance of stochastic gradients. D-SGT in Koloskova et al. (2021) 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 $V = f1, 2, ..., ng$ is the set of local device/nodes, and $W = (W_{i,j})\ 2\ \mathsf{R}^{n\ \ n}$ is a symmetric doubly stochastic matrix known as weighted adjacency matrix, i.e., W satisfies the following properties: (1) $W_{i,j}\ 2\ [0, 1]$, $8i, j$; (2) $W_{i,j} = W_{j,i}$, $8i, j$, i.e., $W^> = W$; and (3) $\sum_{j=1}^n W_{i,j} = 1$, $8i$, i.e., $W\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^> W = \mathbf{1}^>$. 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 $1 = \lambda_1 > \lambda_2 \qquad \lambda_n > \quad 1$, and

$$\rho := \max fj\lambda_2 j, j\lambda_n jg < 1, \tag{2.1}$$

which is standard in decentralized optimization literature (Lian et al., 2017; Tang et al., 2018b). This ensures the communication graph is strongly connected, and after each round of communication with neighbors, the consensus error (i.e., $\sum_{i=1}^n ka_i \quad \bar{a}k^2$ where a_i is a vector owned by the i-th agent only) 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) (Zhang & You, 2019; Lu et al., 2019; Pu & Nedić, 2021; Koloskova et al., 2021) 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 $r\ f$ provided the consensus

[1]We do not compare with NEXT (Di Lorenzo & Scutari, 2016), which only proves asymptotic convergence under deterministic setting.

error $\sum_{i=1}^n \left\|x_i^t \quad \bar{x}^t\right\|^2$ is small. However, as we will show in Section 3, this popular D-SGT algorithm cannot achieve a problem-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 Yuan et al. (2022).

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

Input: T, $\{\eta_t\}$, $u_i^0 = v_i^0 = \nabla F_i(x_i^0, \xi_i^0)$
Output: $\tilde{x} = x^T$ or uniformly from $\{x^1, ..., x^T\}$
for $t = 0, ..., T \quad 1$ **do**
  for each node $i = 1, ..., n$ (in parallel) **do**
    $x_i^{t+1} \quad \sum_{j=1}^n W_{i,j}(x_j^t \quad \eta_t u_j^t)$,
    $v_i^{t+1} \quad \nabla F_i(x_i^{t+1}, \xi_i^{t+1})$
    $u_i^{t+1} \quad \sum_{j=1}^n W_{i,j} u_j^t + v_i^{t+1} \quad v_i^t$
  end for
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 $z_i^t / \left\|z_i^t\right\|$ 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 (Cheng et al., 2023) and achieve linear speedup in decentralized composite optimization (Xiao et al., 2023). 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 Cutkosky & Mehta (2020); Hübler et al. (2023), yet it is unclear and highly non-trivial to understand if one can achieve parameter-free convergence when each of the 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.*

$$\|\nabla f_i(x) \quad \nabla f_i(y)\| \quad L_i\|x \quad y\|.$$

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

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

$$\mathcal{F}_t := \sigma(\xi_i^k | i = 1, ..., n, \ k = 0, ..., t), \ \forall t \quad 1$$

where σ is the σ-algebra generated by the random variables.

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

Input: T, $\{\eta_t\}$, $\{\alpha_t\}$, $x_i^0 = z_i^0 = v_i^0 = 0$
Output: $\tilde{x} = x^T$ or uniformly from $\{x^1, ..., x^T\}$
for $t = 0, ..., T$ 1 **do**
  for each node $i = 1, ..., n$ (in parallel) **do**
    $\tilde{x}_i^{t+1}$ x_i^t $\frac{\eta_t}{\max\{\|z_i^t\|, \epsilon\}} z_i^t$
    v_i^{t+1} $\nabla F_i(x_i^t, \xi_i^t)$
    $\tilde{u}_i^{t+1}$ $u_i^t + v_i^{t+1}$ v_i^t
    $\tilde{z}_i^{t+1}$ $(1$ $\alpha_t) z_i^t + \alpha_t \tilde{u}_i^{t+1}$
  end for
  `# Communication`
  $[x_1^{t+1}, ..., x_n^{t+1}]$ $[\tilde{x}_1^{t+1}, ..., \tilde{x}_n^{t+1}] W$
  $[u_1^{t+1}, ..., u_n^{t+1}]$ $[\tilde{u}_1^{t+1}, ..., \tilde{u}_n^{t+1}] W$
  $[z_1^{t+1}, ..., z_n^{t+1}]$ $[\tilde{z}_1^{t+1}, ..., \tilde{z}_n^{t+1}] W$
end for

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

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

Moreover, we assume $\{\xi_i^{t+1} : i = 1, ..., n\}$ 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 for this paper.

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*

$$\frac{1}{T} \sum_{t=0}^{T} \frac{1}{} \mathbb{E} \|\nabla f(\bar{x}^t)\| \quad \epsilon.$$

We say that an algorithm achieves **linear speedup** if it takes $T \;/\; n$ ^{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.*

$$\|\nabla f_i(x) \quad \nabla f_i(y)\| \quad G_i \|x \quad y\|.$$

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 $O(1/\sqrt{t})$ (Boyd et al., 2003). 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 3.1. *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:*

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

Here $\tilde{\rho} > 0$ 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 $\tilde{O}$.

Remark 3.1. *The rate of $O(1/\sqrt{T})$ matches the lower bound for nonconvex stochastic optimization Arjevani et al. (2023), 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 Zhang & You (2019); Xin et al. (2021); Koloskova et al. (2021). This motivates the design of new algorithms that can achieve linear speedup under problem-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 the result for D-SGT, we provide both the result for fixed and diminishing stepsizes. Our analysis depends on a key observation over the control of the consecutive consensus error. By the update of Algorithm 2 one can get:

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

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

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

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. Now we state the convergence result for a fixed stepsize as follows.

Theorem 3.2. *Suppose Assumptions 3.1 and 3.2 hold and we take $\alpha_t = \sqrt{n/T}$ and $\eta_t = n^{1/4}/T^{3/4}$ in Algorithm 2. The following bounds hold:*

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

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

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{n}\left[\|\mathbf{X}^t \quad \bar{\mathbf{X}}^t\|^2 + \|\mathbf{Z}^t \quad \bar{\mathbf{Z}}^t\|^2\right]$$
$$\mathcal{O}\left(\tilde{\rho}\frac{n^{1/4}}{T^{1/2}} + \tilde{\rho}^2(\sigma^2+L^2)\frac{n^2}{T}\right).$$

Here $\tilde{\rho} > 0$ is a parameter dependent on ρ in (2.1), $\Delta_0 = f(\bar{x}^0) \quad f$ is the initial function value gap and we omit higher-order terms in $\mathcal{O}$. Note that the above three bounds correspond to stationarity, approximation to gradient and consensus errors.

Remark 3.2. *To make $1/T\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{x}^t)\| \quad \epsilon$, we need $T = \tilde{\mathcal{O}}(1/(n\epsilon^4))$, which matches the lower bounds as in Lu & De Sa (2021); Arjevani et al. (2023), and also indicates the linear speedup effect (Lian et al., 2017). Note that the approximation error $\|\bar{z}^t \quad \nabla f(\bar{x}^t)\|$ also enjoys linear speedup effect. Readers might realize that this choice of parameter requires prior knowledge of the total number of nodes. We presume that it is impossible to achieve linear speedup if we are using 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.*

To free the algorithm parameters even from the total number of iterations T, we also present the result when we do not fix the total number of iterations in advance and the stepsize will be diminishing in Theorem 3.3.

Theorem 3.3. *Suppose Assumptions 3.1 and 3.2 hold, also take $\alpha_t = \sqrt{n/t}$ and $\eta_t = n^{1/4}/t^{3/4}$ for any t (take $\eta_0 = \alpha_0 = 0$), the update of Algorithm 2 satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{x}^t)\| \quad \tilde{\mathcal{O}}\Bigg(\frac{\Delta_0 + L + \sigma}{n^{1/4}T^{1/4}}$$
$$+ \frac{L\tilde{\rho}n^{1/4} + \tilde{\rho}^2(\sigma+L)n^{1/2} + L\tilde{\rho}^3 n^{3/4}}{T^{1/4}}\Bigg),$$

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

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{n}\mathbb{E}\left[\|\mathbf{X}^t \quad \bar{\mathbf{X}}^t\|^2 + \|\mathbf{Z}^t \quad \bar{\mathbf{Z}}^t\|^2\right]$$
$$\mathcal{O}\left(\frac{\tilde{\rho}^2(\sigma^2 + L^2 + \tilde{\rho}n^{1/2})}{T}\right),$$

where $\tilde{\rho}$ and Δ_0 are the same as Theorem 3.2 and we omit logarithmic factors in $\tilde{O}$.

Remark 3.3. *To ensure $1/T \sum_{t=0}^{T} \mathsf{E}k\, r\, f(\bar{x}^t)k$ ϵ, we need $T = \tilde{O}(1/\epsilon^4)$, which again matches the lower bound as in Lu & De Sa (2021); Arjevani et al. (2023) up to logarithmic factors. Yet we are not able to achieve a concrete linear speedup effect with this choice of algorithm parameters. This might root back to our estimation of certain error terms in the proof (see Lemma B.6). Nevertheless, we show in the numerical experiments (see Figure 1) that the stepsize choices $\alpha = \sqrt{p/n}$ and $\eta = n^{1/4}$ can still achieve linear speedup empirically, and we thus stick to this choice of parameters in experiments.*

4 Numerical experiments

In this section, we test D-NASA (Algorithm 2) numerically and compare it with existing algorithms such as D-SGD Lian et al. (2017), D-SGT (Algorithm 1) and D-ASAGT Xiao et al. (2023)[2]. We follow the experimental setup in the code framework of Mancino-Ball et al. (2023) to test the algorithms on real datasets using mpi4py (Dalcin & Fang, 2021) and PyTorch (Paszke et al., 2019).

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 where the data sample at each node $\xi = (X, Y)$ is generated by $Y = X^\top \theta_\star + \epsilon$ where $X, \theta_\star$ 2 R^d and ϵ $N(0, \sigma^2)$ are Gaussian noise. We solve the following least-square problem:

$$\min_{\theta 2 \mathsf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \mathsf{E}_{(X,Y)\ D_i}[(Y \quad X^\top \theta)^2]. \tag{4.1}$$

In our experiment, we set $d = 100$, data X $N(0, I_d)$ 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 ($\alpha_t = \sqrt{n/t}$, $\eta_t = n^{1/4}/t^{3/4}$) where the total number of iteration $T = 15000$ and n 2 $\mathsf{f}5, 10, 20\mathsf{g}$. 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 $\bar{x}^t$).

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 $N(0, \mathrm{diag}(100, 1, \quad, 1))$ 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 the fact that conservative constant stepsize choices (usually $O(\sqrt{n/T})$) can lead to linear speedup effect in decentralized training theoretically (Lian et al., 2017; Tang et al., 2018b), this choice is usually for the sake of proof simplicity (see footnote on Page 6 of Lian et al. (2017)). 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 (Lewkowycz et al., 2020; Cohen et al., 2021). 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 Lian et al. (2017); Koloskova et al. (2021); For D-ASAGT, we test the algorithm with stepsizes $\eta_t = \eta\sqrt{n/t}$ and $\alpha_t = \min\mathsf{f}\sqrt{n/t}, 0.3\mathsf{g}$ as suggested in their experiments Xiao et al. (2023);

[2] Xiao et al. (2023) considers nonsmooth proximal version of the algorithm. In our numerical experiments we simply regard the nonsmooth proximal term as zero.

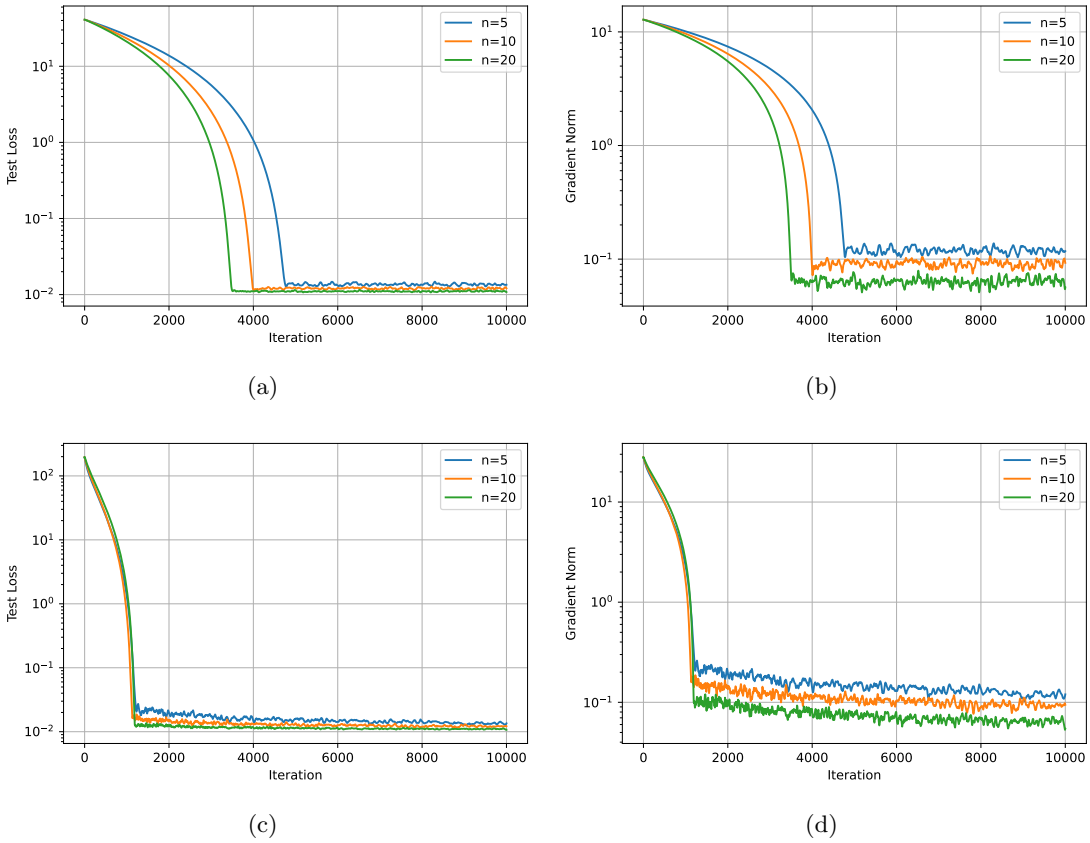
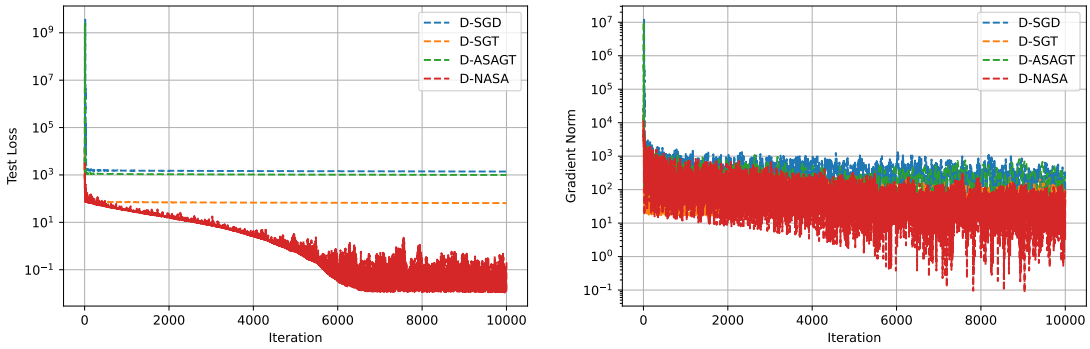


Figure 1: The convergence curve of Algorithm 2 to solve (4.1) with different choice of number of nodes/devices $n \in \{5, 10, 20\}$. The above two figures correspond to fixed stepsizes ($\alpha_t = \sqrt{n/T}$, $\eta_t = n^{1/4}/T^{3/4}$) and below two corresponds to diminishing stepsizes ($\alpha_t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

For D-NASA we take $\eta_t = n^{1/4}/t^{3/4}$ and $\alpha_t = \sqrt{n/t}$ based on our theoretical analysis. We conduct a simple grid search for D-SGD, D-SGD and D-ASAGT to determine and use the best choices of η for each algorithms. The convergence result is shown in Figure 2. Among all algorithms, the test loss of D-NASA decreases with oscillations, presenting the catapults (Lewkowycz et al., 2020) and Edge of Stability (EOS) (Cohen et al., 2021) phenomena, two closely related large-stepsize regimes in which the training converges non-monotonically with oscillations and usually generalize better than small-stepsize settings (Lewkowycz et al., 2020; Cohen et al., 2021; Arora et al., 2022; Ahn et al., 2022). 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-stepsize training setup in the literature, our Algorithm presents the catapults and EOS without any hyperparameter tuning.



(a) (b)

Figure 2: Convergence curve for D-SGD, D-SGD, 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 Mancino-Ball et al. (2023) where we compare D-NASA with D-SGD, D-SGT, D-ASAGT for solving the classification problem:

$$\min_{\theta 2 \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{jD_i j} \sum_{(x,y) 2 D_i} \ell(f(x;\theta), y) \tag{4.2}$$

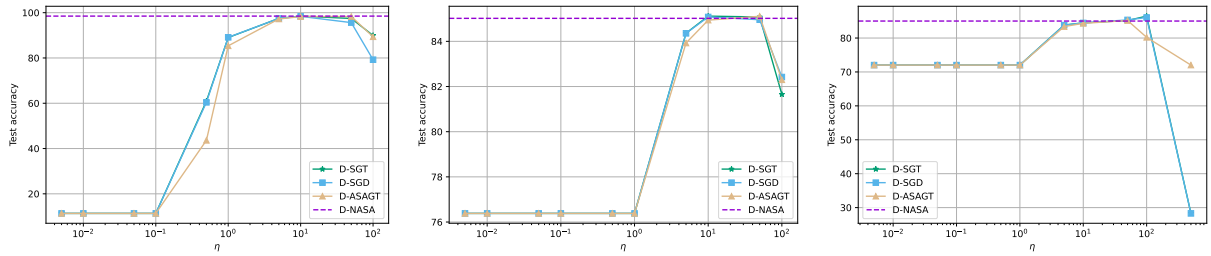
on MNIST, a9a and miniboone datasets[3]. Here ℓ denotes the cross-entropy loss, and f represents a neural network parameterized by θ with x being its input data. D_i is the training set only available to agent i. We use a 2-layer perception model on a9a and miniboone, and the LeNet architecture LeCun et al. (2015) for the MNIST dataset. We take $n = 8$ which connect in the form of a random graph ($\rho = 0.375$) for all three datasets[4]. 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 $\alpha_t = \min f\sqrt{n/t}, 0.3g$; 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 η 2 $f0.005, 0.01, 0.5, 1, 5, 10, 50, 100g$, 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[5].

[3]Available at https://www.openml.org

[4]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 due to page limits.

[5]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 constant are fairly similar for all three datasets.



(a) MNIST (b) a9a (c) miniboone

Figure 3: The testing accuracy of the outputs from different algorithms with respect to different choices of learning rates.

5 Conclusion

In this paper we propose D-NASA, a problem-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. It would be interesting to explore parameter-free convergence in convex, also the nonsmooth regimes.

References

Ahn, K., Bubeck, S., Chewi, S., Lee, Y. T., Suarez, F., and Zhang, Y. Learning threshold neurons via the" edge of stability". *arXiv preprint arXiv:2212.07469*, 2022. (Cited on page 11.)

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214, 2023. (Cited on pages 8, 9, and 10.)

Arora, S., Li, Z., and Panigrahi, A. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pp. 948–1024. PMLR, 2022. (Cited on page 11.)

Boyd, S., Xiao, L., and Mutapcic, A. Subgradient methods. *lecture notes of EE392o, Stanford University, Autumn Quarter*, 2004(01), 2003. (Cited on page 8.)

Carmon, Y. and Hinder, O. Making SGD parameter-free. In *Conference on Learning Theory*, pp. 2360–2389. PMLR, 2022. (Cited on page 5.)

Chen, X., Karimi, B., Zhao, W., and Li, P. On the convergence of decentralized adaptive gradient methods. In *Asian Conference on Machine Learning*, pp. 217–232. PMLR, 2023. (Cited on page 4.)

Cheng, Z., Huang, X., and Yuan, K. Momentum benefits non-iid federated learning simply and provably. *arXiv preprint arXiv:2306.16504*, 2023. (Cited on pages 3 and 6.)

Cohen, J. M., Kaur, S., Li, Y., Kolter, J. Z., and Talwalkar, A. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021. (Cited on pages 10 and 11.)

Cutkosky, A. and Mehta, H. Momentum improves normalized SGD. In *International conference on machine learning*, pp. 2260–2268. PMLR, 2020. (Cited on pages 5 and 6.)

Dalcin, L. and Fang, Y.-L. L. mpi4py: Status update after 12 years of development. *Computing in Science & Engineering*, 23(4):47–54, 2021. (Cited on page 10.)

Di Lorenzo, P. and Scutari, G. Next: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016. (Cited on pages 2, 3, 4, and 5.)

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011. (Cited on page 5.)

Faw, M., Tziotis, I., Caramanis, C., Mokhtari, A., Shakkottai, S., and Ward, R. The power of adaptivity in SGD: Self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, pp. 313–355. PMLR, 2022. (Cited on page 5.)

Gan, S., Lian, X., Wang, R., Chang, J., Liu, C., Shi, H., Zhang, S., Li, X., Sun, T., Jiang, J., et al. Bagua: scaling up distributed learning with system relaxations. *arXiv preprint arXiv:2107.01499*, 2021. (Cited on page 4.)

Garrigos, G. and Gower, R. M. Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*, 2023. (Cited on page 4.)

Grazzi, R., Franceschi, L., Pontil, M., and Salzo, S. On the iteration complexity of hypergradient computation. In *International Conference on Machine Learning*, pp. 3748–3758. PMLR, 2020. (Cited on page 5.)

Hong, M., Razaviyayn, M., Luo, Z.-Q., and Pang, J.-S. A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing. *IEEE Signal Processing Magazine*, 33(1):57–77, 2015. (Cited on page 1.)

Hübler, F., Yang, J., Li, X., and He, N. Parameter-agnostic optimization under relaxed smoothness. In *OPT 2023: Optimization for Machine Learning*, 2023. (Cited on pages 5, 6, 31, and 32.)

Ivgi, M., Hinder, O., and Carmon, Y. DoG is SGD's best friend: A parameter-free dynamic step size schedule. In *International Conference on Machine Learning*. PMLR, 2023. (Cited on page 5.)

Koloskova, A., Lin, T., Stich, S. U., and Jaggi, M. Decentralized deep learning with arbitrary communication compression. *arXiv preprint arXiv:1907.09356*, 2019. (Cited on page 4.)

Koloskova, A., Loizou, N., Boreiri, S., Jaggi, M., and Stich, S. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pp. 5381–5393. PMLR, 2020. (Cited on pages 2 and 4.)

Koloskova, A., Lin, T., and Stich, S. U. An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems*, 34, 2021. (Cited on pages 2, 3, 4, 5, 8, and 10.)

Latafat, P., Themelis, A., and Patrinos, P. On the convergence of adaptive first order methods: proximal gradient and alternating minimization algorithms. *arXiv preprint arXiv:2311.18431*, 2023a. (Cited on page 4.)

Latafat, P., Themelis, A., Stella, L., and Patrinos, P. Adaptive proximal algorithms for convex optimization under local lipschitz continuity of the gradient. *arXiv preprint arXiv:2301.04431*, 2023b. (Cited on page 4.)

LeCun, Y. et al. Lenet-5, convolutional neural networks. *URL: http://yann. lecun. com/exdb/lenet*, 20(5):14, 2015. (Cited on page 12.)

Lewkowycz, A., Bahri, Y., Dyer, E., Sohl-Dickstein, J., and Gur-Ari, G. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020. (Cited on pages 10 and 11.)

Li, M., Andersen, D. G., Park, J. W., Smola, A. J., Ahmed, A., Josifovski, V., Long, J., Shekita, E. J., and Su, B.-Y. Scaling distributed machine learning with the parameter server. In *11th USENIX Symposium on operating systems design and implementation (OSDI 14)*, pp. 583–598, 2014. (Cited on page 3.)

Li, T. and Lan, G. A simple uniformly optimal method without line search for convex optimization. *arXiv preprint arXiv:2310.10082*, 2023. (Cited on page 4.)

Lian, X., Zhang, C., Zhang, H., Hsieh, C.-J., Zhang, W., and Liu, J. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017. (Cited on pages 1, 2, 3, 4, 5, 9, and 10.)

Lian, X., Zhang, W., Zhang, C., and Liu, J. Asynchronous decentralized parallel stochastic gradient descent. In *International Conference on Machine Learning*, pp. 3043–3052. PMLR, 2018. (Cited on page 1.)

Liu, Y., Lin, T., Koloskova, A., and Stich, S. U. Decentralized gradient tracking with local steps. *arXiv preprint arXiv:2301.01313*, 2023. (Cited on page 2.)

Lu, S., Zhang, X., Sun, H., and Hong, M. Gnsd: A gradient-tracking based nonconvex stochastic algorithm for decentralized optimization. In *2019 IEEE Data Science Workshop (DSW)*, pp. 315–321. IEEE, 2019. (Cited on pages 3 and 5.)

Lu, Y. and De Sa, C. Optimal complexity in decentralized training. In *International Conference on Machine Learning*, pp. 7111–7123. PMLR, 2021. (Cited on pages 9 and 10.)

Malitsky, Y. and Mishchenko, K. Adaptive gradient descent without descent. *arXiv preprint arXiv:1910.09529*, 2019. (Cited on page 4.)

Malitsky, Y. and Mishchenko, K. Adaptive proximal gradient method for convex optimization. *arXiv preprint arXiv:2308.02261*, 2023. (Cited on page 4.)

Mancino-Ball, G., Miao, S., Xu, Y., and Chen, J. Proximal stochastic recursive momentum methods for nonconvex composite decentralized optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9055–9063, 2023. (Cited on pages 10 and 12.)

McMahan, H. B. and Streeter, M. Adaptive bound optimization for online convex optimization. *arXiv preprint arXiv:1002.4908*, 2010. (Cited on page 5.)

Nedić, A. and Liu, J. Distributed optimization for control. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:77–103, 2018. (Cited on page 1.)

Nedic, A., Olshevsky, A., and Shi, W. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017. (Cited on pages 2 and 3.)

Neglia, G., Xu, C., Towsley, D., and Calbi, G. Decentralized gradient methods: does topology matter? In *International Conference on Artificial Intelligence and Statistics*, pp. 2348–2358. PMLR, 2020. (Cited on page 4.)

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018. (Cited on page 4.)

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. (Cited on page 10.)

Pu, S. and Nedić, A. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187(1):409–457, 2021. (Cited on pages 2, 3, and 5.)

Qu, G. and Li, N. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017. (Cited on pages 2 and 3.)

Queralta, J. P., Taipalmaa, J., Pullinen, B. C., Sarker, V. K., Gia, T. N., Tenhunen, H., Gabbouj, M., Raitoharju, J., and Westerlund, T. Collaborative multi-robot search and rescue: Planning, coordination, perception, and active vision. *Ieee Access*, 8:191617–191643, 2020. (Cited on page 1.)

Ram, S. S., Nedić, A., and Veeravalli, V. V. Asynchronous gossip algorithms for stochastic optimization. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pp. 3581–3586. IEEE, 2009. (Cited on page 3.)

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of Adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019. (Cited on page 5.)

Tang, H., Gan, S., Zhang, C., Zhang, T., and Liu, J. Communication compression for decentralized training. *Advances in Neural Information Processing Systems*, 31, 2018a. (Cited on page 4.)

Tang, H., Lian, X., Yan, M., Zhang, C., and Liu, J. D^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pp. 4848–4856. PMLR, 2018b. (Cited on pages 1, 2, 3, 4, 5, and 10.)

Tsitsiklis, J. N. *Problems in decentralized decision making and computation*. PhD thesis, Massachusetts Institute of Technology, 1984. (Cited on page 3.)

Xiao, T., Chen, X., Balasubramanian, K., and Ghadimi, S. A one-sample decentralized proximal algorithm for non-convex stochastic composite optimization. In Evans, R. J. and Shpitser, I. (eds.), *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pp. 2324–2334. PMLR, 31 Jul–04 Aug 2023. URL https://proceedings.mlr.press/v216/xiao23a.html. (Cited on pages 3, 6, 10, 22, and 23.)

Xin, R., Khan, U. A., and Kar, S. An improved convergence analysis for decentralized online stochastic non-convex optimization. *IEEE Transactions on Signal Processing*, 69:1842–1858, 2021. (Cited on page 8.)

Xu, J., Zhu, S., Soh, Y. C., and Xie, L. Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 2055–2060. IEEE, 2015. (Cited on pages 2 and 3.)

Yan, F., Sundaram, S., Vishwanathan, S., and Qi, Y. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, 25(11):2483–2493, 2012. (Cited on page 3.)

Yang, J., Li, X., Fatkhullin, I., and He, N. 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. (Cited on page 5.)

Yang, T., Yi, X., Wu, J., Yuan, Y., Wu, D., Meng, Z., Hong, Y., Wang, H., Lin, Z., and Johansson, K. H. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019. (Cited on page 1.)

Yuan, B., He, Y., Davis, J., Zhang, T., Dao, T., Chen, B., Liang, P. S., Re, C., and Zhang, C. Decentralized training of foundation models in heterogeneous environments. *Advances in Neural Information Processing Systems*, 35:25464–25477, 2022. (Cited on pages 2, 4, and 6.)

Yuan, K., Ling, Q., and Yin, W. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016. (Cited on page 3.)

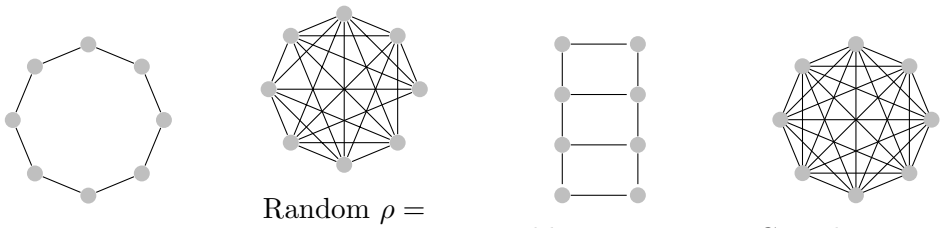
Zhang, J. and You, K. Decentralized stochastic gradient tracking for non-convex empirical risk minimization. *arXiv preprint arXiv:1909.02712*, 2019. (Cited on pages 3, 5, and 8.)

Zhou, D., Ma, S., and Yang, J. AdaBB: Adaptive Barzilai-Borwein method for convex optimization. *arXiv preprint arXiv:2401.08024*, 2024. (Cited on page 4.)

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.



Ring $\rho = 0.805$ | Random $\rho = 0.375$ | Ladder $\rho = 0.892$ | Complete $\rho = 0$

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

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 hyperparameters follow exactly the same as in Section 4.2. The results are presented in Figure 5. It can be seen that D-NASA achieves competitive testing accuracy under almost every network topology choice.

We also include the figures of the loss, accuracy, and stationarity curves of all algorithms. Figure 6 and 7 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.

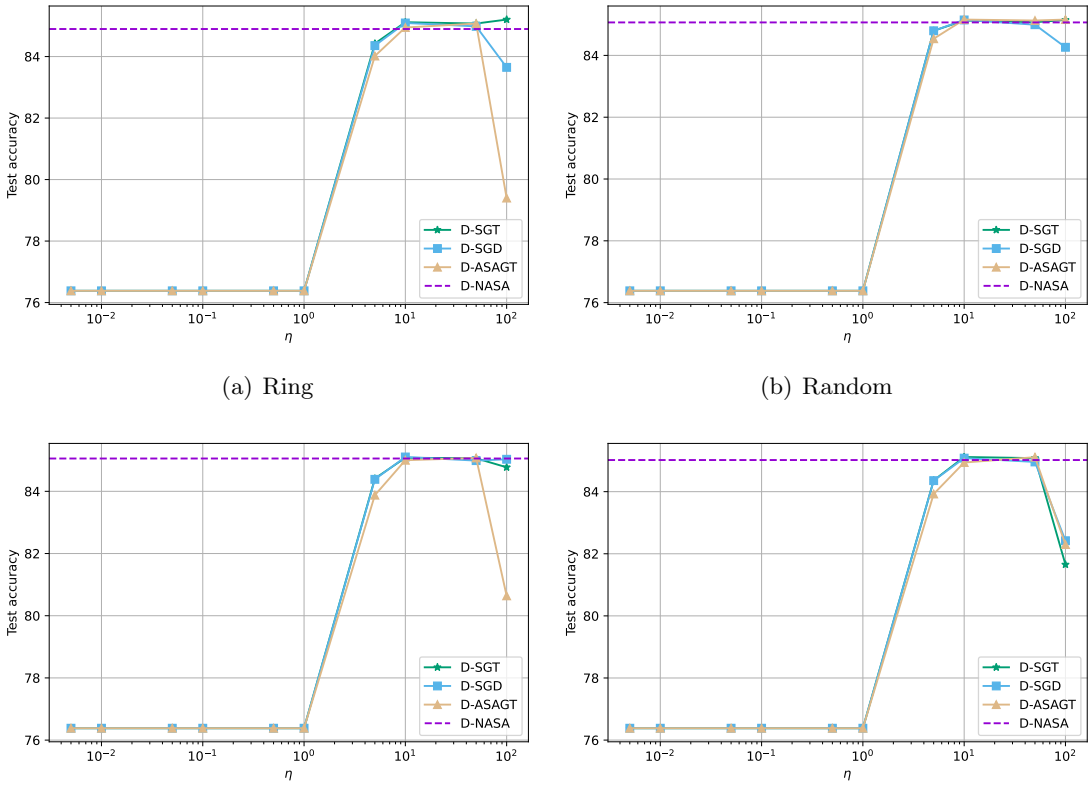
B Convergence analysis

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

From the update of Algorithm 1, we have that:

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

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



(a) Ring (b) Random

(c) Ladder (d) Complete

Figure 5: 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.

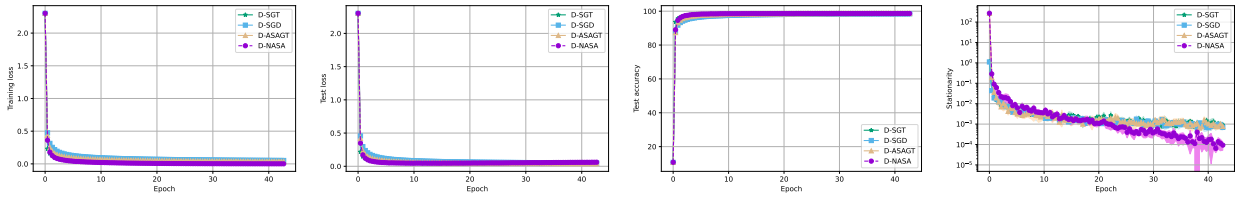


Figure 6: The convergence curve of D-SGD, D-SGT, D-ASAGT, D-NASA for the MNIST dataset, while the first three are at their **best stepsizes** (after the grid search as in Figure 3), and D-NASA follows the stepsize choice as in Remark 3.3, i.e. $\eta_t = n^{1/4}/t^{3/4}$ and $\alpha_t = n^{1/2}/t^{1/2}$. 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.

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

$$\mathsf{E}[f(\bar{x}^{t+1})] \quad \mathsf{E}[f(\bar{x}^t)] \qquad \frac{3\eta_t}{4}\mathsf{E}\left\| r\, f(\bar{x}^t)\right\|^2 + 2\eta_t^2 L\frac{\sigma^2}{n} + (\eta_t + 2\eta_t^2 L)\frac{L^2}{n}\mathsf{E}k\mathbf{X}^t \quad \bar{\mathbf{X}}^t k^2 + \eta_t^2 LG^2$$

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

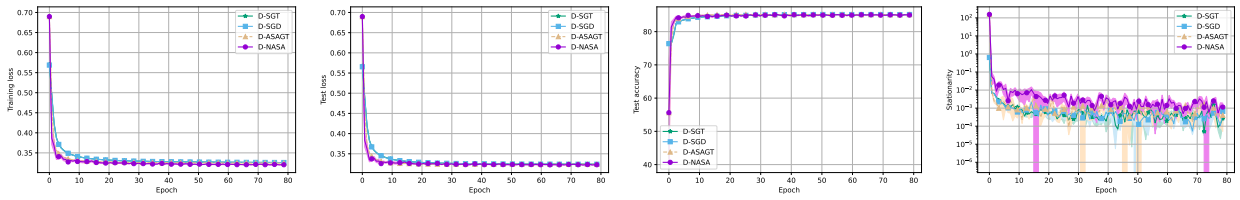


Figure 7: The convergence curve of D-SGD, D-SGT, D-ASAGT, D-NASA for the a9a dataset, while the first three are at their **best stepsizes** (after the grid search as in Figure 3), and D-NASA again follow the same choice as Figure 6. The experiments are repeated and averaged for 10 times.

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

$$\begin{aligned} & f(\bar{x}^{t+1}) \quad f(\bar{x}^t) \\ & r\, f(\bar{x}^t)^\top(\bar{x}^{t+1} \quad \bar{x}^t) + \frac{L}{2}\|\bar{x}^{t+1} \quad \bar{x}^t\|^2 = \quad \eta_t\, r\, f(\bar{x}^t)^\top \bar{u}^t + \frac{\eta_t^2 L}{2}\|\bar{u}^t\|^2 \\ = \quad & \eta_t \left\| r\, f(\bar{x}^t)\right\|^2 \quad \eta_t\, r\, f(\bar{x}^t)^\top(\bar{v}^t \quad r\, f(\bar{x}^t)) + \frac{\eta_t^2 L}{2}\|\bar{v}^t\|^2 \\ = \quad & \eta_t \left\| r\, f(\bar{x}^t)\right\|^2 \quad \eta_t\, r\, f(\bar{x}^t)^\top(\bar{v}^t \quad h^t + h^t \quad r\, f(\bar{x}^t)) + \frac{\eta_t^2 L}{2}\|\bar{v}^t\|^2 \end{aligned}$$

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

Now taking the expectation conditioned on $F_{t\ 1}$, we get

$$\begin{aligned} & \mathsf{E}[f(\bar{x}^{t+1}) \quad f(\bar{x}^t)] \\ & \eta_t \mathsf{E}\left\| r\, f(\bar{x}^t)\right\|^2 \quad \eta_t \mathsf{E}[\, r\, f(\bar{x}^t)^\top(h^t \quad r\, f(\bar{x}^t))] + \frac{\eta_t^2 L}{2}\mathsf{E}\|\bar{v}^t\|^2 \\ & \eta_t \mathsf{E}\left\| r\, f(\bar{x}^t)\right\|^2 + \eta_t/4\mathsf{E}\| r\, f(\bar{x}^t)\|^2 + \eta_t \mathsf{E}\|h^t \quad r\, f(\bar{x}^t)\|^2 + \eta_t^2 L\mathsf{E}\|\bar{v}^t \quad r\, f(\bar{x}^t)\|^2 + \eta_t^2 L G^2 \\ & \frac{3\eta_t}{4}\mathsf{E}\left\| r\, f(\bar{x}^t)\right\|^2 + 2\eta_t^2 L\mathsf{E}\|\bar{v}^t \quad h^t\|^2 + (\eta_t + 2\eta_t^2 L)\mathsf{E}\|h^t \quad r\, f(\bar{x}^t)\|^2 + \eta_t^2 L G^2 \end{aligned}$$

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

Now taking the conditional expectation over $F_{t\ 1}$ we get:

$$\mathsf{E}\|\bar{v}^t \quad h^t\|^2 = \mathsf{E}\|\frac{1}{n}\sum_i (\, r\, F_i(x_i^t, \xi_i^t) \quad r\, f(x_i^t))\|^2 = \frac{1}{n}\sum_i \mathsf{E}\| r\, F_i(x_i^t, \xi_i^t) \quad r\, f(x_i^t)\|^2 \quad \frac{\sigma^2}{n}$$

due to Assumption 3.2.

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

$$\|h^t \quad r\, f(\bar{x}^t)\|^2 = \|\frac{1}{n}(\sum_{i=1}^n r\, f_i(x_i^t) \quad r\, f_i(\bar{x}^t))\|^2 \quad \frac{L^2}{n}\sum_i \|x_i^t \quad \bar{x}^t\|^2 = \frac{L^2}{n}\|\mathbf{X}^t \quad \bar{\mathbf{X}}^t\|^2.$$

□

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:*

$$\begin{aligned} k\mathbf{X}^{t+1} \quad \bar{\mathbf{X}}^{t+1}k^2 \quad \frac{1+\rho}{2}k\mathbf{X}^t \quad \bar{\mathbf{X}}^t k^2 + \eta_t^2\frac{1+\rho^2}{1 \quad \rho^2}k\mathbf{U}^t \quad \bar{\mathbf{U}}^t k^2, \\ k\mathbf{U}^{t+1} \quad \bar{\mathbf{U}}^{t+1}k^2 \quad \frac{1+\rho}{2}k\mathbf{U}^t \quad \bar{\mathbf{U}}^t k^2 + \frac{1+\rho^2}{1 \quad \rho^2}k\mathbf{V}^{t+1} \quad \mathbf{V}^t k^2. \end{aligned}$$

Proof. Since

$$\begin{aligned} & k\mathbf{X}^{t+1} \quad \bar{\mathbf{X}}^{t+1}k^2 = k(\mathbf{X}^t \quad \eta_t\mathbf{U}^t)W \quad (\bar{x}^t \quad \eta_t\bar{u}^t)\mathbf{1}^\top k^2 \\ =& k(\mathbf{X}^t \quad \eta_t\mathbf{U}^t)W \quad \frac{1}{n}(\mathbf{X}^t \quad \eta_t\mathbf{U}^t)\mathbf{1}\mathbf{1}^\top k^2 = k(\mathbf{A}^t \quad \mathbf{A}^t\frac{\mathbf{1}^\top}{n})(W \quad \frac{\mathbf{1}\mathbf{1}^\top}{n})k^2 \\ & k\mathbf{A}^t \quad \mathbf{A}^t\frac{\mathbf{1}^\top}{n}k^2 kW \quad \frac{\mathbf{1}\mathbf{1}^\top}{n}k_2^2 \\ & \rho^2 k\mathbf{A}^t \quad \mathbf{A}^t\frac{\mathbf{1}^\top}{n}k^2 = \rho^2 k(\mathbf{X}^t \quad \bar{x}^t\mathbf{1}^\top) \quad \eta_t(\mathbf{U}^t \quad \bar{u}^t\mathbf{1}^\top)k^2 \\ & \rho^2(1+\frac{1}{c})k\mathbf{X}^t \quad \bar{x}^t\mathbf{1}^\top k^2 + \rho^2\eta_t^2(1+c)k\mathbf{U}^t \quad \bar{u}^t\mathbf{1}^\top k^2 \end{aligned}$$

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

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 Xiao et al. (2023))**.** *Suppose we are given three sequences* $\{a_n\}_{n=0}^{1}$, $\{c_n\}_{n=0}^{1}$, $\{\tau_n\}_{n=0}^{1}$, *and a constant* $r \in (0,1)$ *such that* $a_k, b_k \quad 0$, $0 = c_{\ 1} \quad c_{k+1} \quad c_k \quad 1$ *and*

$$a_{k+1} \quad ra_k + b_k$$

then we have

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

for any positive integer K.

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:*

$$\sum_{t=0}^{T} \frac{1}{n}\eta_t^\tau \mathsf{E}k\mathbf{X}^t \quad \bar{\mathbf{X}}^t k^2 \quad 10\tilde{\rho}\sum_{t=0}^{T} \eta_t^{\tau+2}(\sigma^2 + G^2)$$

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

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

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

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

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

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

Now we inspect the term $\mathbf{V}^{t+1} - \mathbf{V}^t$ following Xiao et al. (2023). We first have

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

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

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

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. □

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

Theorem B.1. *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:*

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

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

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

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

Proof. From Lemma B.1 we know that

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

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

$$\sum_{t=0}^{T} \frac{3\eta_t}{4} \mathsf{E} \left\| r f(\bar{x}^t) \right\|^2 \quad \mathsf{E}[f(\bar{x}^0)] \quad f \; + \frac{2L\sigma^2}{n} \sum_{t=0}^{T} \eta_t^2 + \frac{L^2}{n} \sum_{t=0}^{T} (\eta_t + 2\eta_t^2 L) \mathsf{E} k \mathbf{X}^t \quad \bar{\mathbf{X}}^t k^2 + LG^2 \sum_{t=0}^{T} \eta_t^2$$

$$\mathsf{E}[f(\bar{x}^0)] \quad f \; + \frac{2L\sigma^2}{n} \sum_{t=0}^{T} \eta_t^2 + 10\tilde{\rho} L^2 \sum_{t=0}^{T} (\eta_t^3 + 2\eta_t^4 L)(\sigma^2 + G^2) + LG^2 \sum_{t=0}^{T} \eta_t^2$$

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:

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

i.e.

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

where $\Delta_0 := \mathsf{E}[f(\bar{x}^0)] \quad f$. 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 $\sum_t \eta_t^2 \quad \eta^2 \log(T)$, $\sum_t \eta_t^3 \quad \eta^3 / \sqrt{T}$ and $\sum_t \eta_t^4 \quad \eta^4 / T$, therefore we get the following convergence result:

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

which results in the second line of the result. □

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

From the update of Algorithm 2 we have that:

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

where

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

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.5. *Suppose Assumption 3.1 holds. Algorithm 2 satisfies:*

$$f(\bar{x}^{t+1}) \quad f(\bar{x}^t) \qquad \eta_t k \, r\, f(\bar{x}^t) k + 2\eta_t k \bar{z}^t \qquad r\, f(\bar{x}^t) k + \frac{\eta_t}{n} \sum_{i=1}^{n} k z_i^t \qquad \bar{z}^t k + \frac{\eta_t^2 L}{2} \tag{B.5}$$

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

$$\begin{aligned}
f(\bar{x}^{t+1}) \quad f(\bar{x}^t) \quad & r\, f(\bar{x}^t)^{>}(\bar{x}^{t+1} \quad \bar{x}^t) + \frac{L}{2} k \bar{x}^{t+1} \quad \bar{x}^t k^2 \\
= \quad & \eta_t \, r\, f(\bar{x}^t)^{>} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{z_i^t}{k z_i^t k} \right) + \frac{\eta_t^2 L}{2} \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{z_i^t}{k z_i^t k} \right\|^2 \\
& \eta_t \, r\, f(\bar{x}^t)^{>} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{z_i^t}{k z_i^t k} \right) + \frac{\eta_t^2 L}{2} \\
= \quad & \eta_t (\, r\, f(\bar{x}^t) \quad \bar{z}^t)^{>} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{z_i^t}{k z_i^t k} \right) \quad \eta_t (\bar{z}^t)^{>} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{z_i^t}{k z_i^t k} \right) + \frac{\eta_t^2 L}{2} \\
= \quad & \eta_t (\, r\, f(\bar{x}^t) \quad \bar{z}^t)^{>} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{z_i^t}{k z_i^t k} \right) \quad \eta_t (\bar{z}^t)^{>} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{z_i^t}{k z_i^t k} \quad \frac{\bar{z}^t}{k \bar{z}^t k} \right) \quad \eta_t k \bar{z}^t k + \frac{\eta_t^2 L}{2} \\
& 2\eta_t k \, r\, f(\bar{x}^t) \quad \bar{z}^t k \quad \eta_t k \, r\, f(\bar{x}^t) k + \eta_t k \bar{z}^t k \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{z_i^t}{k z_i^t k} \quad \frac{\bar{z}^t}{k \bar{z}^t k} \right\| + \frac{\eta_t^2 L}{2}
\end{aligned}$$

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

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

which concludes the proof. □

We have the following dual convergence.

Lemma B.6. *We have*

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

where

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

Consequently, we get:

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

where we have the following conventions:

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

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

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

Therefore we get

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

where the second term is due to smoothness of each f_i, also the last term is by $\mathsf{E}[\langle\delta_3^{\tau_1}, \delta_3^{\tau_2}\rangle] = 0$ for any $\tau_1 \neq \tau_2$ (due to unbiased assumption) and

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

since all the cross inner-product terms vanish due to unbiased assumption. □

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

Lemma B.7. *For the update of Algorithm 2, we have:*

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

Proof. Since

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

where $\mathbf{A}^t := \mathbf{X}^t - \eta_t\hat{\mathbf{Z}}^t$. Taking $c = \frac{2\rho^2}{1-\rho^2} > 0$ 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

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

where we use

$$\|\frac{1}{n}\sum_{i=1}^n v^i - \frac{1}{n}\sum_{i=1}^n v^i\|^2 = \|\frac{1}{n}\sum_{i=1}^n v^i\|^2 - \|\frac{1}{n}\sum_{i=1}^n v^i\|^2 \le \frac{1}{n}\sum_{i=1}^n\|v^i\|^2$$

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

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

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

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

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

where the first inequality uses Cauchy-Schwarz inequality, and the second utilizes Lipschitz smoothness of each f_i also note that $k\,r\,f_i(\bar{x}^{t+1}) \quad r\,f_i(\bar{x}^t)k \quad Lk\bar{z}^t k \quad L$. □

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

Lemma B.8. *For the update of Algorithm 2, if decreasing sequences such that* $0 \quad \alpha_{t+1} \quad \alpha_t \quad 1$ *and* $0 \quad \eta_{t+1} \quad \eta_t \quad 1$*, we have:*

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

where

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

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

Proof. The first line is by Lemma B.7 and B.3 by taking $a_\tau = \frac{1}{n} \mathsf{E}k\mathbf{X}^\tau \quad \bar{\mathbf{X}}^\tau k^2$, $b_\tau = \eta_\tau^2 (1+\rho^2)/(1 \quad \rho^2)$, $c_\tau = 1$ and $r = (1+\rho)/2$ in Lemma B.3 directly.

For the second line, by Lemma B.7 we get

$$\begin{aligned}
\sqrt{\frac{1}{n} \mathsf{E}k\mathbf{X}^{t+1} \quad \bar{\mathbf{X}}^{t+1} k^2} \quad & \sqrt{\frac{1+\rho}{2} \frac{1}{n} \mathsf{E}k\mathbf{X}^t \quad \bar{\mathbf{X}}^t k^2 + \eta_t^2 \frac{1+\rho^2}{1 \quad \rho^2} \frac{1}{n} \mathsf{E}k\hat{\mathbf{Z}}^t \quad \bar{\hat{\mathbf{Z}}}^t k^2} \\
& \sqrt{\frac{1+\rho}{2}} \sqrt{\frac{1}{n} \mathsf{E}k\mathbf{X}^t \quad \bar{\mathbf{X}}^t k^2} + \eta_t \sqrt{\frac{1+\rho^2}{1 \quad \rho^2}} \sqrt{\frac{1}{n} \mathsf{E}k\hat{\mathbf{Z}}^t \quad \bar{\hat{\mathbf{Z}}}^t k^2} \\
& \sqrt{\frac{1+\rho}{2}} \sqrt{\frac{1}{n} \mathsf{E}k\mathbf{X}^t \quad \bar{\mathbf{X}}^t k^2} + \eta_t \sqrt{\frac{1+\rho^2}{1 \quad \rho^2}}
\end{aligned}$$

where the second inequality is by $\sqrt{a+b} \quad \sqrt{a} + \sqrt{b}$ and third is by $k\hat{\mathbf{Z}}^t \quad \bar{\hat{\mathbf{Z}}}^t k^2 \quad n$. Now taking $a_\tau = \sqrt{\frac{1}{n} \mathsf{E}k\mathbf{X}^\tau \quad \bar{\mathbf{X}}^\tau k^2}$, $b_\tau = \eta_\tau \sqrt{(1+\rho^2)/(1 \quad \rho^2)}$, $c_\tau = \alpha_\tau$ 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.7 we get

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

by using Lemma B.7 for two times. Also since

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

where for the third line we again use Lemma B.7. Combining all above equations gives the second line of the theorem.

As for the forth line, note that from Lemma B.7 and $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$, we have:

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

Repeating the proof of the third line gives the fourth line. □

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

Theorem B.2. *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:*

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

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

Proof. By Lemma B.5, we have

$$\eta_t\mathbb{E}\|\nabla f(\bar{x}^t)\|\le\mathbb{E}[f(\bar{x}^t)-f(\bar{x}^{t+1})]+2\eta_t\mathbb{E}\|\bar{z}^t-\nabla f(\bar{x}^t)+\nabla f(\bar{x}^t)\|+\mathbb{E}[\frac{\eta_t}{n}\sum_{i=1}^{n}\|z_i^t-\bar{z}^t\|]+\frac{\eta_t^2L}{2}$$
$$\le\mathbb{E}[f(\bar{x}^t)-f(\bar{x}^{t+1})]+2\eta_t\mathbb{E}\|\bar{z}^t-\nabla f(\bar{x}^t)\|+\eta_t\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|z_i^t-\bar{z}^t\|^2}+\frac{\eta_t^2L}{2}$$

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

Now sum up from $t = 0$ to T 1 and using Lemma B.5 and B.8, we get

$$\sum_{t=0}^{T} \eta_t \mathbb{E}\|\nabla f(\bar{x}^t)\| \leq \Delta_0 + 2\sum_{t=0}^{T} \eta_t \mathbb{E}\|\bar{z}^t - \nabla f(\bar{x}^t)\| + \sum_{t=0}^{T} \eta_t \sqrt{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|z_i^t - \bar{z}^t\|^2} + \frac{L}{2}\sum_{t=0}^{T} \eta_t^2$$

$$\leq \Delta_0 + 2\sum_{t=0}^{T} \eta_t \left(L\sum_{\tau=1}^{t} \beta_{(\tau+1):t}\eta_\tau + L\tilde{\rho}\sum_{\tau=0}^{t} \beta_{(\tau+1):t}\alpha_\tau\eta_\tau + \sigma\sqrt{\frac{1}{n}\sum_{\tau=0}^{t} \beta_{(\tau+1):t}^2\alpha_\tau^2} \right)$$

$$+ \tilde{\rho}^2\sqrt{10\sigma^2 + 5L^2}\sum_{t=0}^{T} \eta_t\alpha_t + 2\sqrt{5}L\tilde{\rho}^3\sum_{t=0}^{T} \eta_t^2\alpha_t + \frac{L}{2}\sum_{t=0}^{T} \eta_t^2 \tag{B.8}$$

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

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

and

$$\sum_{t=0}^{T} \eta_t \sum_{\tau=1}^{t} \beta_{(\tau+1):t}\eta_\tau = \sum_{t=0}^{T} \frac{\eta}{T^{3/4}} \sum_{\tau=1}^{t} (1 - \frac{\alpha}{T^{1/2}})^{t-\tau} \frac{\eta}{T^{3/4}}$$

$$= \frac{\eta^2}{T^{3/2}} \sum_{t=0}^{T} \sum_{\tau=1}^{t} (1 - \frac{\alpha}{T^{1/2}})^{t-\tau}$$

$$= \frac{\eta^2}{T^{3/2}} \sum_{t=0}^{T} (1 - \frac{\alpha}{T^{1/2}})^t \frac{T^{1/2}}{\alpha}[(1 - \frac{\alpha}{T^{1/2}})^{t+1} - 1]$$

$$\frac{\eta^2}{\alpha T} \sum_{t=0}^{T} (1 - \frac{\alpha}{T^{1/2}}) \leq \frac{\eta^2}{\alpha}$$

Similarly

$$\sum_{t=0}^{T} \eta_t \sum_{\tau=0}^{t} \beta_{(\tau+1):t}\alpha_\tau\eta_\tau \leq \frac{\eta^2}{T^{1/2}}$$

$$\sum_{t=0}^{T} \eta_t \sqrt{\sum_{\tau=0}^{t} \beta_{(\tau+1):t}^2\alpha_\tau^2} \leq \eta\sqrt{\bar{\alpha}}$$

Now plugging everything back we get:

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

$$+ \tilde{\rho}^2\sqrt{10\sigma^2 + 5L^2}\frac{\alpha\eta}{T^{1/4}} + 2\sqrt{5}L\tilde{\rho}^3\frac{\alpha\eta^2}{T} + \frac{L}{2}\frac{\eta^2}{T^{1/2}}$$

i.e.

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

$$+ \frac{\tilde{\rho}^2\sqrt{10\sigma^2 + 5L^2}\alpha}{T^{1/2}} + \frac{2\sqrt{5}L\tilde{\rho}^3\alpha^2\eta}{T^{7/4}}$$

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 $k\bar{z}^t \quad r\, f(\bar{x}^t)k$.

It remains to bound the consensus error (third line), which follows directly from Lemma B.8 and (B.9). □

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 Hübler et al. (2023).

Lemma B.9. *Let $q \in (0,1)$, $p \quad 0$ and $t > 0$. Further let positive integers a, b s.t. $2 \quad a \quad b$, then we have that for any $\alpha > 0$,*

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

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

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

and in particular,

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

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

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

Proof. For the first equation we get

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

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

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

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

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

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

$$(q \quad p)\int_a^b {}_1 t^{q\ p\ 1}\exp\left(\frac{\alpha t^{1\ q}}{1\quad q}\right)dt \quad (q \quad p)b^{q\ 1}\int_a^b {}_1 t^{\ p}\exp\left(\frac{\alpha t^{1\ q}}{1\quad q}\right)dt$$

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 Hübler et al. (2023). We have

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

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. □

Lemma B.10. *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*

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

and

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

where $\beta_t = 1 \quad \alpha_t$.

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 Hübler et al. (2023). We prove the second line of this Lemma as a show case and refer to Lemma 11 in Hübler et al. (2023) for the

detail of the proof of the last two lines. For the second line, we have

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

where we use Lemma B.9 for the first inequality. □

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

Theorem B.3. *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 $\eta>0$, the update of Algorithm 2 satisfies:*

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

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

Proof. Same as the proof of Theorem B.2, we get

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

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.10 we get:

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

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

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. □