Journal of Machine Learning Research 23 (2022) 1-96 Submitted 11/19; Revised 10/21; Published 7/22

Robust Distributed Accelerated Stochastic Gradient Methods for Multi-Agent Networks

Alireza Fallah* AFALLAH@MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, United States of America

Mert Gürbüzbalaban*♣ MG1366@RUTGERS.EDU
Department of Management Science and Information Systems
Rutgers Business School
Piscataway, NJ 08854, United States of America

Asuman Ozdaglar* ASUMAN@MIT.EDU
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, United States of America

Umut Şimşekli*♠ UMUT.SIMSEKLI@INRIA.FR
INRIA - Département d'Informatique de l'École Normale Supérieure
PSL Research University
Paris, France

Lingjiong Zhu* ZHU@MATH.FSU.EDU
Department of Mathematics
Florida State University
Tallahassee, FL 32306, United States of America

* *The authors are in alphabetical order.*
♣ *Corresponding author.*
♠ *This work was partially conducted while U.Ş. was a visiting faculty member at University of Oxford, Department of Statistics and affiliated with LTCI, Télécom Paris, Institut Polytechnique de Paris.*

Editor: Prateek Jain

Abstract

We study distributed stochastic gradient (D-SG) method and its accelerated variant (D-ASG) for solving decentralized strongly convex stochastic optimization problems where the objective function is distributed over several computational units, lying on a fixed but arbitrary connected communication graph, subject to local communication constraints where noisy estimates of the gradients are available. We develop a framework which allows to choose the stepsize and the momentum parameters of these algorithms in a way to optimize performance by systematically trading off the bias, variance and dependence to network effects. When gradients do not contain noise, we also prove that D-ASG can *achieve acceleration*, in the sense that it requires $\mathcal{O}(\sqrt{\kappa}\log(1/\varepsilon))$ gradient evaluations and $\mathcal{O}(\sqrt{\kappa}\log(1/\varepsilon))$

communications to converge to the same fixed point with the non-accelerated variant where κ is the condition number and ε is the target accuracy. For quadratic functions, we also provide finer performance bounds that are tight with respect to bias and variance terms. Finally, we study a multistage version of D-ASG with parameters carefully varied over stages to ensure exact convergence to the optimal solution. It achieves optimal and accelerated $\mathcal{O}(-k/\sqrt{\kappa})$ linear decay in the bias term as well as optimal $\mathcal{O}(\sigma^2/k)$ in the variance term. We illustrate through numerical experiments that our approach results in accelerated practical algorithms that are robust to gradient noise and that can outperform existing methods.

Keywords: Distributed Optimization, Accelerated Methods, Stochastic Optimization, Robustness, Multi-Agent Networks

1. Introduction

Advances in sensing and processing technologies, communication capabilities and smart devices have enabled deployment of systems where a massive amount of data is collected by many distributed autonomous units to make decisions. There are numerous such examples including a set of sensors collecting and processing information about a time-varying spatial field (e.g., to monitor temperature levels or chemical concentrations) (Blatt et al., 2007), a collection of mobile robots performing dynamic tasks spread over a region (Nedić et al., 2018), federated learning on edge devices (Konečnỳ et al., 2016; McMahan et al., 2017), on-device peer-to-peer learning (Koloskova et al., 2019) and distributed model training across a network or computers (Arjevani et al., 2020; Gürbüzbalaban et al., 2021; Scaman et al., 2018). In such systems, most of the information is often collected in a decentralized, distributed manner, and processing of information has to go hand-in-hand with its communication and sharing across these units over an undirected network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ defined by the set of (computational units) agents $\mathcal{V} = \{1, 2, \ldots, N\}$ connected by the edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. In such a setting, we consider the group of agents (i.e., the nodes) collaboratively solving the following optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x), \tag{1}$$

where each $f_i : \mathbb{R}^d \to \mathbb{R}$ is known by agent i only and therefore referred to as its *local objective function*. We assume each f_i is μ-strongly convex with L-Lipschitz gradients (hence f is also μ-strongly convex with L-Lipschitz gradient and we refer to $\kappa = L/\mu$ as its condition number). We also use x_* to denote the unique optimal solution of (1). In addition, we denote the local model of node i at iteration k by $x_i^{(k)} \in \mathbb{R}^d$.

We consider the setting where each agent i has access to noisy estimates $\tilde{\nabla} f_i(x)$ of the actual gradients satisfying the following assumption:

Assumption 1 *Recall that $x_i^{(k)}$ denotes the decision variable of node i at iteration k. We assume at iteration k, node i has access to $\tilde{\nabla} f_i\left(x_i^{(k)}, w_i^{(k)}\right)$ which is an estimate of $\nabla f_i\left(x_i^{(k)}\right)$ where $w_i^{(k)}$ is a random variable independent of $\left\{w_j^{(t)}\right\}_{j=1,\ldots,N,t=1,\ldots,k-1}$ and $\left\{w_j^{(k)}\right\}_{j \neq i}$. Moreover, we assume*

$$\mathbb{E}\left[\tilde{\nabla} f_i\left(x_i^{(k)}, w_i^{(k)}\right) \middle| x_i^{(k)}\right] = \nabla f_i\left(x_i^{(k)}\right), \quad \mathbb{E}\left[\left\|\tilde{\nabla} f_i\left(x_i^{(k)}, w_i^{(k)}\right) - \nabla f_i\left(x_i^{(k)}\right)\right\|^2 \middle| x_i^{(k)}\right] \leq \sigma^2.$$

To simplify the notation, we suppress the $w_i^{(k)}$ dependence, and denote $\tilde{\nabla} f_i\left(x_i^{(k)}, w_i^{(k)}\right)$ by $\tilde{\nabla} f_i\left(x_i^{(k)}\right)$.

This arises naturally in distributed learning problems where $f_i(x)$ represents the expected loss $\mathbb{E}_{\eta_i}[f_i(x, \eta_i)]$ where η_i are independent data points collected at node i (see e.g. Pu and Nedić (2018); Lan et al. (2020); Pu et al. (2021)). For this setting, $\tilde{\nabla} f_i(x)$ is an unbiased estimator of $\nabla f_i(x)$ which we assume satisfies the bounded variance assumption of Assumption 1. In Appendix E, we will discuss the unbounded variance assumption (Assumption 5) that extends Assumption 1, and show that all the main results in the paper can be extended.

Note that in our setting, a master node that can coordinate the computations is not available unlike the master/slave architecture studied in the literature (see e.g. Mishchenko et al. (2018); Agarwal and Duchi (2011); Hakimi et al. (2019); Lee et al. (2018); Meng et al. (2016); Jaggi et al. (2014); Xin and Khan (2020)). Furthermore, our setting covers an arbitrary network topology that is more general than particular network topologies such as the complete graph or ring graph.

Deterministic variants of problem (1) have been studied extensively in the literature. Much of the work builds on the Distributed Gradient (DG) method proposed in Nedic and Ozdaglar (2009) where each agent keeps local estimates of the optimal solution of (1) and updates by a combination of weighted average of neighbors' estimates and a gradient step (normalized by the stepsize α_k) of the local objective function. Nedic and Ozdaglar (2009) analyzed the case with convex and possibly nonsmooth local objective functions, constant stepsize $\alpha_k = \alpha > 0$, and agents linked over an undirected connected graph and showed that the ergodic average of local estimates of the agents converge at rate $\mathcal{O}(1/k)$ to an $\mathcal{O}(\alpha)$ neighborhood of the optimal solution of problem (1) (where k denotes the number of iterations). Yuan et al. (2016) considered this algorithm for the case that local functions are smooth, i.e., $\nabla f_i(x)$ are Lipschitz continuous, and when $f_i(x)$ are either convex, restricted strongly convex or strongly convex. For the convex case, they show the network-wide mean estimate converges at rate $\mathcal{O}(1/k)$ to an $\mathcal{O}(\alpha)$ neighborhood of the optimal solution, and for the strongly convex case, all local estimates converge at a linear rate $\mathcal{O}(\exp(-k/\Theta(\kappa)))$ to an $\mathcal{O}(\alpha)$ neighborhood of x_*.[1]

There have been many recent works on developing new distributed deterministic algorithms with faster convergence rate and exact convergence to the optimal solution x_*. We start by summarizing the literature in this area that are most relevant to this work. First, Shi et al. (2015) provides a novel algorithm which can be viewed as a primal-dual algorithm for the constrained reformulation of problem (1) (see Mokhtari and Ribeiro (2016) for this interpretation) that achieves *exact convergence* with linear rate to the optimal solution; however the linear convergence rate with the recommended stepsize is $\rho = 1 - \mathcal{O}(\frac{1}{\kappa^2})$

1. For two real-valued functions f and g, we say $f = \Theta(g)$ if there exist positive constants C_ℓ and C_u such that $C_\ell g(x) \leq f(x) \leq C_u g(x)$ for every x in the domain of f and g with $\|x\|$ being sufficiently large.

where κ is the condition number (see Table 1). This convergence guarantee will be slow for ill-conditioned problems when κ is large. Second, Qu and Li (2018) proposes to update the DG method such that agents also maintain, exchange, and combine estimates of gradients of the global objective function of (1). This update is based on a technique called "gradient tracking" (see e.g. Di Lorenzo and Scutari (2015, 2016)) which enables better control on the global gradient direction and yields a linear rate of convergence to the optimal solution (see Jakovetić (2019) for a unified analysis of these two methods). In a follow up paper, Qu and Li (2020) also considered an acceleration of their algorithm and achieved a linear convergence rate $\mathcal{O}(\exp(-k/\Theta(\kappa^{5/7})))$ to the optimal solution. To our best knowledge, whether an accelerated primal variant of the DG algorithm can achieve the non-distributed $\mathcal{O}(\exp(-k/\Theta(\sqrt{\kappa})))$ linear rate to a neighborhood of the optimum solution with $\sqrt{\kappa}$ dependence has been an open problem. Alternative distributed first-order methods besides DG have also been studied. In particular, if additional assumptions are made such as the explicit characterization of Fenchel dual of the local objective functions, referred to as the *dualable* setting as in Scaman et al. (2018); Uribe et al. (2021)), then it is known that the multi-step dual accelerated (MSDA) method of Scaman et al. (2018) achieves the $\mathcal{O}(\exp(-k/\Theta(\sqrt{\kappa})))$ linear rate to the optimum with $\sqrt{\kappa}$ dependence. For deterministic distributed optimization problems under smooth and strongly convex objectives, Dvinskikh and Gasnikov (2019) proposed the PSTM algorithm and provided accelerated convergence guarantees. Recently, Scaman et al. (2019) provided lower bounds which matches the upper bounds of Dvinskikh and Gasnikov (2019) up to logarithmic factors (see also Scaman et al. (2019) for a discussion of deterministic optimal algorithms under different assumptions (Lipschitz continuity, strong convexity, smoothness, and a combination of strong convexity and smoothness)).

This paper focuses on the Distributed Stochastic Gradient (D-SG) method (which is a stochastic version of the DG method) and its momentum enhanced variant, Distributed Accelerated Stochastic Gradient (D-ASG) method. These methods are relevant for solving distributed learning problems and are natural decentralized versions of the stochastic gradient and its variant based on Nesterov's momentum averaging (Nesterov, 2003; Can et al., 2019). In this paper, we focus on strongly convex and smooth objectives. Several works studied D-SG under these assumptions although D-ASG remains relatively understudied except the deterministic case (see e.g. Jakovetić et al. (2014); Xi et al. (2017); Li et al. (2020); Qu and Li (2016)). The performance of distributed algorithms such as D-SG and their deterministic versions depend on the connectivity of the underlying network structure as expected. In particular, when D-SG and D-ASG are run on undirected graphs, the propagation of information among neighbors is governed by a symmetric mixing matrix W which depend on the network structure and its eigenvalues affect the convergence rates. In particular; the largest eigenvalue of the matrix W is one, and the second largest (in modulus) of the eigenvalues of W, which we refer to as γ in this paper (formally defined in (8)), arises in the study of distributed algorithms such as D-SG. We summarize the existing convergence rate results for D-SG in Table 1.[2] Among these, Rabbat (2015) studied composite stochastic optimization problems and showed a $\mathcal{O}(\sigma^2/k)$ convergence rate for D-SG and its mirror descent variant. Koloskova et al. (2019) studied decentralized stochastic gradient

2. See also Shamir and Srebro (2014) for a different noise model than ours in the mini-batch setting, where each objective f_i can be expressed as a finite sum.

algorithms when the nodes compress (e.g. quantize or sparsify) their updates. Pu et al. (2021) provided an asymptotic network independent sublinear rate. In our approach, we use a dynamical system representation of these iterative algorithms (presented in Lessard et al. (2016) and further used in Hu and Lessard (2017); Aybat et al. (2020, 2019)) to provide rate estimates for convergence of the local agent iterates to a neighborhood of the optimal solution of problem (1). Our bounds are presented in terms of three components: (i) a *bias term* that shows the decay rate of the initialization error (i.e., distance of the initial estimates to the optimal solution) independent of gradient noise, (ii) a *variance term* that depends on the error level σ^2 of local objective functions' gradients, measuring the "robustness" of the algorithm to noise (in a sense that we will define precisely later), (iii) a *network effect* that highlights the dependence on the structure of the network. In this paper, in addition to the convergence analysis for D-SG and D-ASG, our purpose is to study the trade-offs and interplays between these three terms that affect the performance.

Contributions. We have three sets of contributions.

First, we study the convergence rate of DSG with constant stepsize which is used in many practical applications (Alghunaim and Sayed, 2020, 2018; Dieuleveut et al., 2020). Our bounds provide tighter guarantees on the bias term as well as novel guarantees on the variance term for this algorithm. For quadratic functions, we provide sharper estimates for the bias, variance, and network effect terms that are tight, as there exist simple quadratic functions that achieve these bounds.

Second, we consider D-ASG with constant stepsize. We show that the bias term decays linearly with rate $\mathcal{O}(-k/\sqrt{\kappa})$ to a neighborhood of the optimal solution, and thus, it achieves an accelerated rate. We also provide an explicit characterization for this neighborhood, in terms of noise and network structure parameters, with the variance term dominating for small enough stepsize. When the objectives f_i are all quadratic, we obtain non-asymptotic guarantees that are explicit in terms of their linear convergence rate and dependence to noise, generalizing available known guarantees for ASG to the distributed setting (Can et al., 2019).

For both algorithms, following earlier work on non-distributed versions of these algorithms (Aybat et al., 2020), we use our explicit characterization of bias, variance, and network effect terms to provide a computational framework that can choose algorithm parameters to trade-off these difference effects in a systematic manner. In the centralized setting, it has been observed and argued that accelerated algorithms are often more sensitive to noise than non-accelerated algorithms (see e.g. Flammarion and Bach (2015); d'Aspremont (2008); Aybat et al. (2019); Hardt (2014)), however to our knowledge this behavior has not been systematically studied in the context of decentralized algorithms. We study the asymptotic variance of the D-SG and D-ASG iterates as a measure of robustness to random gradient noise and provide explicit expressions for this quantity for quadratic objectives as well as upper bounds for strongly convex objectives. This allows us to compare D-SG and D-ASG in terms of their robustness to random noise properties. Our results (see the discussion after Theorem 7) show that indeed D-ASG can be less robust compared to D-SG depending on the choice of the momentum and stepsize parameters, shedding further light into the tuning of hyperparameters (stepsize and momentum) in the distributed setting.

Algorithm	Extra Assump.	Stepsize α_k	Convergence Rate
EXTRA Shi et al. (2015)	$\sigma = 0$	α as in Shi et al. (2015)	$\mathbb{E}\left\|x_i^{(k)} - x_*\right\|^2 \leq \mathcal{O}\left(\rho^k\right)$ where $\rho = 1 - \mathcal{O}(1/\kappa^2)$
D-SG Tsianos and Rabbat (2012)	Yes‡	$\mathcal{O}\left(\frac{1}{k}\right)$	$\mathbb{E}f\left(x_i^{(k)}\right) - f_* \leq \mathcal{O}\left(G^2 \frac{\log(\sqrt{N}k)}{(1-\sqrt{\gamma})k}\right)$
D-SG Rabbat (2015)	No	$\mathcal{O}\left(\frac{1}{k}\right)$	$\mathbb{E}f\left(x_i^{(k)}\right) - f_* \leq \mathcal{O}\left(\frac{\kappa^2 \|x_i^{(0)} - x_*\|^2}{k^2} + \frac{\sigma^2}{N\mu^2 k}\right) + o\left(\frac{1}{k}\right)$
D-SG Pu et al. (2021)†	No	$\mathcal{O}\left(\frac{1}{k}\right)$	$\mathbb{E}\left\|\bar{x}^{(k)} - x^*\right\|^2 \leq \mathcal{O}\left(\frac{\sigma^2}{\mu^2 Nk} + \frac{\kappa^2}{(1-\gamma)^2\mu^2 k^2} + \frac{\sigma^2\kappa^2}{(1-\gamma)\mu^2 k^2}\right)$
D-SG Koloskova et al. (2019)†	Yes‡	$\mathcal{O}\left(\frac{1}{k}\right)$	$\mathbb{E}f\left(x_{avg}^{(k)}\right) - f_* \leq \mathcal{O}\left(\frac{\sigma^2}{\mu Nk} + \frac{\kappa G^2}{\mu(1-\gamma)^4 k^2} + \frac{G^2}{\mu(1-\gamma)^6 k^3}\right)$
D-SG **Proposition 3 in this paper**	No	α	$\mathbb{E}\left\|\bar{x}^{(k)} - x_*\right\|^2$ $\leq \mathcal{O}\Big((1-\alpha\mu)^{2k}\left\|\bar{x}^{(0)} - x_*\right\|^2 + \frac{\alpha\sigma^2}{\mu N}$ $+\kappa^2\alpha^2\left(\frac{D^2}{(1-\gamma)^2 N} + \frac{\sigma^2}{1-\gamma^2}\right)\Big)$ $\mathbb{E}\left\|x_i^{(k)} - x_*\right\|^2$ $\leq \mathcal{O}\Big((1-\alpha\mu)^{2k}\left\|\bar{x}^{(0)} - x_*\right\|^2 + \frac{\alpha\sigma^2}{\mu N}$ $+\kappa^2\alpha^2\left(\frac{D^2}{(1-\gamma)^2 N} + \frac{\sigma^2}{1-\gamma^2}\right)$ $+\gamma^k\|x^{(0)}\|^2 + \frac{D^2\alpha^2}{(1-\gamma)^2} + \frac{\sigma^2 N\alpha^2}{1-\gamma^2}\Big)$
D-ASG **Proposition 10 in this paper**	No	α	$\mathbb{E}\left\|\bar{x}^{(k)} - x_*\right\|^2$ $\leq \mathcal{O}\Big(\left(1 - \frac{\sqrt{\alpha\mu}}{2}\right)^k \frac{\|x^{(0)} - x^*\|^2}{N} + \frac{\sigma^2\sqrt{\alpha}}{\mu\sqrt{\mu N}} + \frac{\kappa^2 C_0\alpha}{N(1-\gamma)^2}\Big)$ $\mathbb{E}\left\|x_i^{(k)} - x_*\right\|^2$ $\leq \mathcal{O}\Big(\left(1 - \frac{\sqrt{\alpha\mu}}{2}\right)^k \frac{\|x^{(0)} - x^*\|^2}{N}$ $+\frac{\sigma^2\sqrt{\alpha}}{\mu\sqrt{\mu N}} + \left(\frac{\kappa^2}{N} + 1\right)\frac{C_0\alpha}{(1-\gamma)^2}\Big)$
D-MASG **Corollary 18 in this paper**	No	$\mathcal{O}\left(\frac{1}{k}\right)$	$\mathbb{E}\left\|\bar{x}^{(k)} - x_*\right\|^2$ $\leq \mathcal{O}\Big(\exp\left(-\frac{k}{\Theta(\sqrt{\tilde{\kappa}})}\right)\frac{\left\|x^{(0)} - x^*\right\|^2}{N}$ $+\frac{\sigma^2}{N\mu\sqrt{\mu k}} + \frac{\kappa^2 C_0}{N(1-\gamma)^2 k^4}\Big)$ $\mathbb{E}\left\|x_i^{(k)} - x_*\right\|^2$ $\leq \mathcal{O}\Big(\exp\left(-\frac{k}{\Theta(\sqrt{\tilde{\kappa}})}\right)\frac{\left\|x^{(0)} - x^*\right\|^2}{N}$ $+\frac{\sigma^2}{N\mu\sqrt{\mu k}} + \left(\frac{\kappa^2}{N} + 1\right)\frac{C_0}{(1-\gamma)^2 k^4}\Big)$

Table 1: Summary for D-SG and D-ASG. $\bar{x}^{(k)}$ denotes the average of nodes' estimates at time k, i.e., $\bar{x}^{(k)} := \frac{1}{N}\sum_{i=1}^N x_i^{(k)}$, and $x_{avg}^{(k)}$ is a weighted average defined in Koloskova et al. (2019). Also, $\gamma \in (0,1)$ is the second largest modulus of the eigenvalues of the mixing matrix W (formally defined in (8)). In the table, μ denotes the strong convexity constant, L is the gradient Lipschitz constant and $\kappa = L/\mu$ is the condition number, whereas $\tilde{\kappa} := \frac{\kappa+1}{\lambda_N^W}$ is a scaled condition number (formally introduced in (38)), where λ_N^W is the smallest positive eigenvalue of W, D^2 is defined in (27) such that $D^2 = \mathcal{O}\left(L^2\mathbb{E}\|x^{(0)} - x^*\|^2 + \|\nabla F(x^*)\|^2\right)$ as $\alpha \to 0$, and C_0 is an explicitly computable constant such that $C_0 = \mathcal{O}(1)$ as $\alpha \to 0$.
†: The authors analyze a D-SG method with a slightly different update then ours.
‡: The authors make the extra assumption $\sup_{i,j}\mathbb{E}\left\|\nabla f_i\left(x_i^{(j)}\right)\right\|^2 \leq G^2$.

Finally, we study a multistage version of D-ASG, building on the non-distributed method in Aybat et al. (2019), whereby a distributed accelerated stochastic gradient method with constant stepsize and momentum parameter is used at every stage, with parameters carefully varied over stages to ensure exact convergence to the optimal solution x_*. Similar to Aybat et al. (2019), a momentum restart is used to enable stitching the improvement obtained over consecutive stages. We show that our proposed method achieves an accelerated $\mathcal{O}(-k/\sqrt{\kappa})$ linear decay in the bias term as well as a $\mathcal{O}(\sigma^2/k)$ term in the variance term and $\mathcal{O}((1-\gamma)^{-2}/k^4)$ in terms of network effect, where $1-\gamma$ is the spectral gap of the network, see (8) for a formal definition. We also show that the node averages also achieves $O(\frac{1}{Nk})$ for the variance term with a tight dependency to the number of nodes N. This dependency to k and $\sqrt{\kappa}$ is optimal in the context of centralized black-box stochastic optimization. This suggests that our analysis is tight in terms of its k and $\sqrt{\kappa}$ dependency, although the problems we consider is not black-box optimization but finite-sum problems. Such a dependency on k and $\sqrt{\kappa}$ was obtained previously for the PBSTM algorithm of Dvinskikh and Gasnikov (2019) which is optimal up to logarithmic terms. To the best of our knowledge, our analysis provides the best bounds for the D-ASG algorithm. Our results show that D-ASG without noise converges to a fixed point with the accelerated rate, i.e. the rate has a $\sqrt{\kappa}$ dependency to the condition number. A summary of all our convergence results is provided in Table 1. We also provide numerical experiments that show the efficiency of the D-ASG method in a number of decentralized optimization settings.

Other related work. There has been a growing recent interest in the dynamical system representation of distributed optimization algorithms to facilitate their analysis and design. In particular, Sundararajan et al. (2020) provides a framework to design a broad class of distributed algorithms for deterministic decentralized optimization for time-varying graphs. This framework provides worst-case certificates of linear convergence via semi-definite programming. Other related papers (Sundararajan et al., 2017, 2019) allow analysis and design of deterministic distributed optimization algorithms. However, these results and approaches are targeted for deterministic distributed algorithms and they do not directly apply to the stochastic algorithms we consider in this paper. Robustness of stochastic optimization algorithms to stepsize have also been considered in the literature. In particular, the accelerated gradient methods of Lan (2012, Theorem 2, Corollary 1) do enjoy various robustness properties to noise; in particular, for appropriate stepsize choices, if L is a Lipschitz constant of the gradient, σ^2 the noise, and D the diameter of the underlying domain, one may achieve rates roughly

$$\mathbb{E}\left[f\left(x^k\right)-f(x_*)\right] \leq \frac{LD^2}{k^2}+\frac{\sigma D}{\sqrt{k}}\left(\gamma^{-1}+\gamma\right),$$

where $\gamma>0$ is a particular stepsize multiplier choice. Thus, misspecifying γ does not force a massive degradation in convergence rates, which reflects the robustness considerations of Nemirovski et al. (2009). The work of Duchi et al. (2012b, Theorem 2.1.) also shows a similar robustness result to stepsize specification.

Notation. Let $\mathcal{S}_{\mu,L}(\mathbb{R}^d)$ denote the set of functions from $\mathbb{R}^d$ to $\mathbb{R}$ that are μ-strongly convex and L-smooth, that is, for every $x,y\in\mathbb{R}^d$,

$$\frac{L}{2}\|x-y\|^2 \geq f(x)-f(y)-\nabla f(y)^T(x-y) \geq \frac{\mu}{2}\|x-y\|^2,$$

where we have the condition number $\kappa = L/\mu$. Let $0_{a\times b}$ denote the zero matrix with a rows and b columns. Given a collection of square matrices $[A_i]_{i=1}^m$, the matrix $\mathbf{diag}([A_i]_{i=1}^m)$ denotes the block diagonal square matrix with i-th diagonal block equal to A_i. For two matrices $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$, we denote their Kronecker product by $A \otimes B$. For two functions g, h defined over positive integers, we say $f = \mathcal{O}(g)$ if there exists a constant C_u and a positive integer n_0 such that $f(n) \leq C_u g(n)$ for every positive integer $n \geq n_0$. We say $f = \tilde{\mathcal{O}}(g)$ if there exists a constant C_u and a positive integer n_0 such that $f(n) \leq C_u g(n)\log(n)$ for every positive integer $n \geq n_0$. We use the notation $\|A\|_2$ to denote the 2-norm (largest singular value) of a matrix A, whereas we use $\|A\|_F$ to denote the Frobenius norm of A. For two real-valued functions f and g, we say $f = \Theta(g)$ as $x \to 0$ if there exist positive constants C_ℓ and C_u such that $C_\ell g(x) \leq f(x) \leq C_u g(x)$ for every x in a neighborhood of 0 and lying in the domains of f and g.

2. Distributed Stochastic Gradient and Its Accelerated Variant

We will first study the distributed stochastic gradient (D-SG) method which is the stochastic version of the distributed gradient (DG) method introduced in Nedic and Ozdaglar (2009), and then focus on its accelerated variant.

Consider an undirected network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ that is connected by edges $\mathcal{E} \subseteq \mathcal{V}\times\mathcal{V}$, where $\mathcal{V} = \{1, \ldots, N\}$ denotes the set of vertices. We associate this network with an $N \times N$ symmetric, doubly stochastic weight matrix W. We have $W_{ij} = W_{ji} > 0$ if $(i, j) \in \mathcal{E}$ and $i \neq j$, and $W_{ij} = W_{ji} = 0$ if $(i, j) \notin \mathcal{E}$ and $i \neq j$, and finally $W_{ii} = 1 - \sum_{j\neq i} W_{ij} > 0$ for every $1 \leq i \leq N$.[3] It is known that the eigenvalues of a doubly stochastic matrix W can be ordered in a descending manner satisfying:

$$1 = \lambda_1^W > \lambda_2^W \geq \cdots \geq \lambda_N^W > -1,$$

where the largest eigenvalue is 1 with an all-one eigenvector, i.e. $W\mathbf{1} = \mathbf{1}$, and the smallest eigenvalue is greater than -1. The eigenvalues of W can be used to study the properties of the network associated with the weight matrix W (see e.g. Chung (1997)). For example, if W represents the transition matrix of a Markov chain, then $1 - \max\{|\lambda_2^W|, |\lambda_N^W|\}$, known as the spectral gap, can be used to measure the mixing time of the Markov chain, i.e. how fast the Markov chain converges to its stationary distribution (see e.g. Levin et al. (2009)). Such a matrix W always exists (see e.g. Boyd et al. (2006)) if the graph is not bi-partite and there can be different choices of W (Shi et al., 2015). For bi-partite graphs, one can also construct such a matrix W by considering the transition matrix of a lazy random walk on the graph (see e.g. Chung (1997)).

Next, we make a few definitions for the sake of subsequent analysis. First define the average iterates

$$\bar{x}^{(k)} := \frac{1}{N}\sum_{i=1}^N x_i^{(k)} \in \mathbb{R}^d. \tag{2}$$

Next we define the column vector

$$x^{(k)} = \left[\left(x_1^{(k)}\right)^T, \left(x_2^{(k)}\right)^T, \ldots, \left(x_N^{(k)}\right)^T\right]^T \in \mathbb{R}^{Nd}, \tag{3}$$

3. We adopt the convention that the node is a neighbor of itself, i.e. $(i, i) \in \mathcal{E}$.

which concatenates the local decision variables into a single vector. We also define $x^* \in \mathbb{R}^{Nd}$ as

$$x^* = \begin{bmatrix} x_*^T & x_*^T & \cdots & x_*^T \end{bmatrix}^T, \tag{4}$$

which is the column vector of length Nd that concatenates N copies of the optimizer x_* to the problem (1).

In addition, we define $F : \mathbb{R}^{Nd} \to \mathbb{R}$ as

$$F(x) := F(x_1, \ldots, x_N) = \sum_{i=1}^{N} f_i(x_i),$$

where

$$\tilde{\nabla} F\left(x^{(k)}\right) = \left[\left(\tilde{\nabla} f_1\left(x_1^{(k)}\right)\right)^T, \left(\tilde{\nabla} f_2\left(x_2^{(k)}\right)\right)^T, \ldots, \left(\tilde{\nabla} f_N\left(x_N^{(k)}\right)\right)^T\right]^T,$$

which obeys

$$\mathbb{E}\left[\tilde{\nabla} F\left(x^{(k)}\right) \middle| x^{(k)}\right] = \nabla F\left(x^{(k)}\right), \quad \mathbb{E}\left[\left\|\tilde{\nabla} F\left(x^{(k)}\right) - \nabla F\left(x^{(k)}\right)\right\|^2 \middle| x^{(k)}\right] \leq \sigma^2 N, \tag{5}$$

due to Assumption 1. Furthermore, $F \in \mathcal{S}_{\mu,L}(\mathbb{R}^{Nd})$ is μ-strongly convex and L-smooth.

2.1 Distributed Stochastic Gradient (D-SG)

Recall that $x_i^{(k)}$ denotes the decision variable of node i at iteration k. The D-SG iterations update this variable by performing a stochastic gradient descent update with respect to the local cost function f_i together with a weighted averaging with the decision variables $x_j^{(k)}$ of node i's immediate neighbors $j \in \Omega_i := \{j : (i,j) \in \mathcal{E}\}$:

$$x_i^{(k+1)} = \sum_{j \in \Omega_i} W_{ij} x_j^{(k)} - \alpha \tilde{\nabla} f_i\left(x_i^{(k)}\right), \tag{6}$$

where $\alpha > 0$ is the stepsize. Note that we can express the D-SG iterations as

$$x^{(k+1)} = \mathcal{W} x^{(k)} - \alpha \tilde{\nabla} F\left(x^{(k)}\right), \tag{7}$$

where $\mathcal{W} := W \otimes I_d$.

Without noise, i.e., when $\tilde{\nabla} F(x^{(k)}) = \nabla F(x^{(k)})$, D-SG reduces to the DG algorithm. In this case, Yuan et al. (2016) show that DG algorithm is *inexact* in the sense that the iterates $x_i^{(k)}$ of the DG algorithm do not converge to the optimum x_* in general with constant stepsize, but instead converge linearly to a fixed point x_i^∞ that is in a neighborhood of the solution satisfying

$$\|x_i^\infty - x_*\| \leq C_1 \frac{\alpha}{1-\gamma} = \mathcal{O}\left(\frac{\alpha}{1-\gamma}\right), \quad \text{where} \quad \gamma := \max\left\{\left|\lambda_2^W\right|, \left|\lambda_N^W\right|\right\}, \tag{8}$$

for some constant C_1 with the explicit expression

$$C_1 := \sqrt{2L \sum_{i=1}^{N} \left(f_i(0) - f_i^*\right)} \cdot \left(1 + \frac{2(L+\mu)}{\mu}\right), \quad f_i^* := \min_{x \in \mathbb{R}^d} f_i(x), \tag{9}$$

provided that the stepsize α satisfies some conditions (Yuan et al., 2016) (see Lemma 20 in Appendix A for details).

Similar to (4), we define the column vector

$$x^\infty := \left[(x_1^\infty)^T, (x_2^\infty)^T, \cdots, (x_N^\infty)^T\right]^T \in \mathbb{R}^{Nd}, \tag{10}$$

which is a concatenation of the fixed point x_i^∞ of node i over all the nodes. It can be checked that the unique fixed point x^∞ to (7) in the noiseless setting is the solution to

$$(I_{Nd} - \mathcal{W})x^\infty + \alpha \nabla F(x^\infty) = 0. \tag{11}$$

This means that the sequence $\xi_k := x^{(k)} - x^\infty$ converges to zero with an appropriate choice of the stepsize. The performance of the algorithm can then be measured by the distance of x^∞ to $x^* \in \mathbb{R}^{Nd}$ given by (4).

2.2 Distributed Accelerated Stochastic Gradient (D-ASG)

Consider the following variant of D-SG:

$$\begin{aligned} x_i^{(k+1)} &= \sum_{j\in\Omega_i} W_{ij} y_j^{(k)} - \alpha \tilde{\nabla} f_i\left(y_i^{(k)}\right), \\ y_i^{(k)} &= (1+\beta)x_i^{(k)} - \beta x_i^{(k-1)}, \end{aligned} \tag{12}$$

where $\alpha > 0$ is the stepsize and $\beta \geq 0$ is called the *momentum parameter*. This algorithm has also been considered in the literature by Jakovetić et al. (2014) in the noiseless setting.

We define the average iterates $\bar{x}^{(k)}$ and the column vector $x^{(k)}$ as in (2) and (3), respectively. Also, similar to (3), we define the column vector

$$y^{(k)} = \left[\left(y_1^{(k)}\right)^T, \left(y_2^{(k)}\right)^T, \ldots, \left(y_N^{(k)}\right)^T\right]^T \in \mathbb{R}^{Nd}.$$

Then, we can re-write the D-ASG iterates (12) as:

$$\begin{aligned} x^{(k+1)} &= \mathcal{W} y^{(k)} - \alpha \tilde{\nabla} F\left(y^{(k)}\right), \\ y^{(k)} &= (1+\beta)x^{(k)} - \beta x^{(k-1)}, \end{aligned} \tag{13}$$

for $k \geq 0$ starting from the initial values $x_i^{(0)} \in \mathbb{R}$ and $x_i^{(-1)} \in \mathbb{R}$ for each node i. Here, $\alpha > 0$ is the stepsize and $\beta \geq 0$ is the momentum parameter. Note that for $\beta = 0$, D-ASG reduces to the D-SG algorithm. When there is a single node, i.e. $N = 1$, D-ASG also reduces to the Nesterov's (non-distributed) accelerated stochastic gradient algorithm (ASG) (Nesterov, 2003). Note that this algorithm is also inexact in the sense that both $\{x^{(k)}\}$ and $\{y^{(k)}\}$ will also converge to the same point $x^\infty = y^\infty$ in the noiseless setting where x^∞ is the fixed point of the distributed gradient (DG) algorithm defined by (11).

2.3 Convergence Rates and Robustness to Gradient Noise

Consider both D-SG and D-ASG algorithms, subject to gradient noise satisfying Assumption 1. For this scenario, the noise is *persistent*, i.e., it does not decay over time, and it is possible that the limit of $x^{(k)}$ as $k \to \infty$ may not exist (even in the non-distributed setting), see Can et al. (2019); therefore, one natural way,[4] of defining *robustness* of an algorithm to gradient noise, is to consider the worst-case limiting variance along all possible subsequences, i.e.

$$J_\infty := \frac{1}{\sigma^2 N} \limsup_{k\to\infty} \operatorname{Var}\left(x^{(k)}\right). \tag{14}$$

In the special case, when F is a quadratic function and the gradient noise is i.i.d. with an isotropic Gaussian distribution, the quantity J_∞ is equal to the square of the H_2 norm of the linear dynamical system corresponding to the D-ASG iterations (13) (see e.g. Zhou et al. (1996); Aybat et al. (2020)). H_2 norm is well-studied in the robust control theory as a robustness metric and has been considered in the distributed algorithms literature previously as a measure of robustness to white noise (see e.g. Pirani et al. (2018); Sarkar et al. (2018); Chapman (2015)). Indeed, we observe from (14) that J_∞ is equal to the ratio of the output variance and the input noise variance $\sigma^2 N$ (which is the variance of noise at the worst case), therefore it can be interpreted as a signal-to-noise ratio (SNR) measure, quantifying how robust the underlying algorithm is to white noise. We also note that the same definition was recently applied to optimization to develop noise-robust non-distributed algorithms (Aybat et al., 2020). Our definition (14) of robustness is motivated by such connections to the robust control and optimization literature.

In the next sections, we will provide bounds on the robustness level J_∞ and the expected distance to both the fixed point and the optimum for the D-SG and D-ASG algorithms. In particular, in the non-distributed setting, it is known that ASG can be less robust to noise compared to gradient descent (Hardt, 2014; Aybat et al., 2020); we will later obtain bounds in Section 2.3.3 for the robustness of D-ASG and D-SG which suggests a similar behavior in the distributed setting when the stepsize is small enough.

For analysis purposes, we consider the penalized objective function $F_{\mathcal{W},\alpha}(x) : \mathbb{R}^{Nd} \to \mathbb{R}$ defined as

$$F_{\mathcal{W},\alpha}(x) := \frac{1}{2\alpha} x^T (I_{Nd} - \mathcal{W}) x + F(x), \quad \alpha > 0. \tag{15}$$

Similar penalized objectives have also been considered in the past to analyze deterministic algorithms (see e.g. (Yuan et al., 2016, Section 2), Mansoori and Wei (2017)). It can be seen that its gradient (with respect to x) is $\nabla F_{\mathcal{W},\alpha}(x) = \frac{1}{\alpha}(I_{Nd} - \mathcal{W})x + \nabla F(x)$. Since $0_{Nd} \preceq I_{Nd} - \mathcal{W} \preceq (1 - \lambda_N^W) I_{Nd}$, we have also

$$F_{\mathcal{W},\alpha} \in \mathcal{S}_{\mu,L_\alpha}\left(\mathbb{R}^{Nd}\right) \quad \text{with} \quad L_\alpha := \frac{1 - \lambda_N^W}{\alpha} + L. \tag{16}$$

Furthermore, the unique minimizer z^* of $F_{\mathcal{W},\alpha}$ satisfies the first-order conditions

$$\nabla F_{\mathcal{W},\alpha}(z^*) = (I_{Nd} - \mathcal{W}) z^* + \alpha \nabla F(z^*) = 0.$$

4. There are other possible ways to define a robustness measure, see e.g. Aybat et al. (2020).

Then, it follows from (11) that $z^* = x^\infty$, i.e. the minimizer of $F_{\mathcal{W},\alpha}$ coincides with the limit point x^∞. In fact, we can re-write the D-SG iterations (7) as

$$x^{(k+1)} = x^{(k)} - \alpha \tilde{\nabla} F_{\mathcal{W},\alpha}\left(x^{(k)}\right), \tag{17}$$

which is equivalent to running a non-distributed stochastic gradient algorithm for minimizing an alternative objective $F_{\mathcal{W},\alpha}$ in dimension Nd. We can also re-write the D-ASG iterations (13) as

$$\begin{aligned} x^{(k+1)} &= y^{(k)} - \alpha \tilde{\nabla} F_{\mathcal{W},\alpha}\left(y^{(k)}\right), \\ y^{(k)} &= (1+\beta)x^{(k)} - \beta x^{(k-1)}. \end{aligned} \tag{18}$$

These iterations are identical to the iterations of the (non-distributed) ASG. In other words, D-ASG applied to solve the problem (1) in dimension d is equivalent to running a non-distributed ASG algorithm for minimizing an alternative objective $F_{\mathcal{W},\alpha}$ in dimension Nd.

This connection allows us to analyze both D-SG and D-ASG with existing techniques developed for non-distributed algorithms in Aybat et al. (2020, 2019) that builds on dynamical system representation of optimization algorithms.

2.3.1 Dynamical system representation

We first reformulate D-SG (17) and D-ASG update rules (18) as a discrete-time dynamical system:

$$\xi_{k+1} = A\xi_k + B\tilde{\nabla} F_{\mathcal{W},\alpha}(C\xi_k), \tag{19}$$

where ξ_k is the state, and A, B, C are system matrices that are appropriately chosen. For example, we can represent the D-SG iterates with the choice of

$$\xi_k := x^{(k)} - x^\infty, \quad A := I_{Nd}, \quad B := -\alpha I_{Nd}, \quad C := I_{Nd}. \tag{20}$$

Similarly, we can represent the D-ASG iterations as the dynamical system (19) with

$$\xi_k := \left[\left(x^{(k)} - x^\infty\right)^T, \left(x^{(k-1)} - x^\infty\right)^T\right]^T, \tag{21}$$

and $A = \tilde{A}_{\text{dasg}} \otimes I_{Nd}, B := \tilde{B}_{\text{dasg}} \otimes I_{Nd}, C := \tilde{C}_{\text{dasg}} \otimes I_{Nd}$ where

$$\tilde{A}_{\text{dasg}} = \begin{bmatrix} 1+\beta & -\beta \\ 1 & 0 \end{bmatrix}, \quad \tilde{B}_{\text{dasg}} = \begin{bmatrix} -\alpha \\ 0 \end{bmatrix}, \quad \tilde{C}_{\text{dasg}} = \begin{bmatrix} 1+\beta & -\beta \end{bmatrix}. \tag{22}$$

(see also Lessard et al. (2016) for such a dynamical system representation in the deterministic case). For studying the dynamical system (22), we introduce the following Lyapunov function

$$V_{P,\alpha,c}(\xi) := \xi^T P \xi + c\left[F_{\mathcal{W},\alpha}(T\xi + x^\infty) - F_{\mathcal{W},\alpha}(x^\infty)\right], \tag{23}$$

where $c \geq 0$ is a scalar, P is a positive semi-definite matrix and $T = I_{Nd}$ for D-SG and $T = [1\ 0] \otimes I_{Nd}$ for D-ASG. Since x^∞ is the minimum of $F_{\mathcal{W},\alpha}$, we observe that $V_{P,\alpha,c}(\xi)$

has non-negative values. In particular, $V_{P,\alpha,c}(0) = 0$. In the special case when $c = 0$, we obtain

$$V_P(\xi) := V_{P,\alpha,0}(\xi) = \xi^T P \xi.$$

In the next section, we obtain convergence results for D-SG and D-ASG for constant stepsize and momentum which also implies guarantees on the robustness measure J_∞. The analysis is based on studying the Lyapunov function (23) for different choices of the matrix P and the scalar c. In particular, for D-SG we can choose P to be the identity matrix and $c = 0$, however for D-ASG, the choice of P is less trivial and depends on the choice of the stepsize α and β in general. Here, our choice of the Lyapunov function (23) is motivated by Fazlyab et al. (2018) which studied this Lyapunov function to analyze accelerated gradient methods in the centralized deterministic setting.

2.3.2 Analysis of Distributed Stochastic Gradient

We next provide a performance bound for D-SG in Theorem 1. It shows that the expected distance square to the fixed point $\mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right]$ can be bounded as a sum of two terms: *i*) A *bias term* that depends on the initialization and decays with a linear rate $\rho^2(\alpha)$ where $\rho(\alpha) = \max\left\{\left|1-\alpha\mu\right|, \left|\lambda_N^W - \alpha L\right|\right\}$. (*ii*) A *variance term* that scales linearly with the noise level σ^2 providing a bound on the asymptotic variance $\limsup_{k\to\infty} \mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right]$ and hence the robustness level J_∞. When there is no noise (when $\sigma = 0$), the variance term is zero, and we obtain a linear convergence rate for the (deterministic) DG algorithm with rate $\rho^2(\alpha)$. This improves the previously best known convergence rate ρ_δ^2 for DG obtained in Yuan et al. (2016), where $\rho_\delta^2 := 1 - \frac{\alpha\mu L}{\mu+L} + \alpha\delta - \alpha^2\delta\frac{\mu L}{\mu+L}$, which can get arbitrarily close to $1 - \frac{\alpha\mu L}{\mu+L}$, see Theorem 7 in Yuan et al. (2016). We also note that the convergence rate and robustness we provide in Theorem 1 is tight for D-SG in the sense that they are attained for some quadratic choices of the objective (see Remark 32 in Appendix C).

For proving Theorem 1, we exploit the above-mentioned fact that running D-SG on the objective F is equivalent to running (non-distributed SG) on the modified objective $F_{\mathcal{W},\alpha}$ and we build on the existing results for non-distributed stochastic gradient (Aybat et al., 2020, Prop. 4.3); the proof is given in Appendix B.

Theorem 1 *Consider running D-SG method with stepsize $\alpha \in \left(0, \frac{1+\lambda_N^W}{L}\right)$. Then, for every $k \geq 0$,*

$$\mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right] \leq \rho(\alpha)^{2k}\left\|x^{(0)} - x^\infty\right\|^2 + \frac{1-\rho(\alpha)^{2k}}{1-\rho(\alpha)^2}\sigma^2\alpha^2 N, \tag{24}$$

where $\rho(\alpha) = \max\left\{\left|1-\alpha\mu\right|, \left|\lambda_N^W - \alpha L\right|\right\} \in [0,1)$. As a result, the robustness of the D-SG method satisfies

$$J_\infty(\alpha) \leq \frac{\alpha^2}{1-\rho(\alpha)^2}.$$

We recall that the penalized objective $F_{\mathcal{W},\alpha}$ depends on the network and the stepsize. The fixed point x^∞ is the minimum of the penalized objective $F_{\mathcal{W},\alpha}$. In general, the difference $\|x^{(\infty)} - x^*\|$ is not zero and it depends on the network structure and the stepsize α.

We call this term the "network effect"; it can be controlled by the the inequality (8). The following corollary is obtained by a direct application of the inequality (8) to Theorem 1.

Corollary 2 *Consider running D-SG method with stepsize* $\alpha \in \left(0, \frac{1+\lambda_N^W}{\mu+L}\right)$. *Then, for every* $k \geq 0$,

$$\mathbb{E}\left[\left\|x^{(k)} - x^{\infty}\right\|^2\right] \leq (1-\alpha\mu)^{2k}\left\|x^{(0)} - x^{\infty}\right\|^2 + \alpha\sigma^2 N \frac{1-(1-\alpha\mu)^{2k}}{\mu(2-\alpha\mu)}, \tag{25}$$

which implies that the robustness of the D-SG method satisfies

$$J_{\infty}(\alpha) \leq \frac{\alpha}{\mu(2-\alpha\mu)}.$$

In addition, if $\alpha \leq \frac{1}{L+\mu}$, *we have*

$$\mathbb{E}\left[\left\|x^{(k)} - x^{*}\right\|^2\right] \leq 2(1-\alpha\mu)^{2k}\left\|x^{(0)} - x^{\infty}\right\|^2 + 2\alpha\sigma^2 N \frac{1-(1-\alpha\mu)^{2k}}{\mu(2-\alpha\mu)} + \frac{2\alpha^2 C_1^2 N}{(1-\gamma)^2}, \tag{26}$$

where γ, C_1 *are given in* (8)-(9).

Next, we provide the performance bound on the distance between the average of iterates $\bar{x}^{(k)}$ and the minimizer x_*. Here, we can show that the asymptotic variance of the averaged iterates $\bar{x}^{(k)}$ with constant stepsize is $\mathcal{O}(\sigma^2/N)$; this is because averaging the iterates also averages the noise over the nodes.

Proposition 3 *Assume* $0 < \alpha \leq \frac{2}{L+\mu}$, $\alpha < \frac{1+\lambda_N^W}{L}$ *and* $\mu\alpha(1+\lambda_N^W - \alpha L) < 1$. *Then, for any* k, *we have*

$$\begin{aligned}\mathbb{E}\left\|\bar{x}^{(k)} - x_*\right\|^2 \leq\ & 8\left(\frac{\alpha}{\mu(1-\frac{\alpha L}{2})} + \frac{(1+\alpha L)^2}{\mu^2(1-\frac{\alpha L}{2})^2}\right)\left(\frac{L^2D^2\alpha^2}{N(1-\gamma)^2} + \frac{L^2\sigma^2\alpha^2}{(1-\gamma^2)}\right) \\ &+ 8\frac{\gamma^{2k} - \left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)^k}{\gamma^2 - 1 + \alpha\mu\left(1-\frac{\alpha L}{2}\right)}\frac{L^2\gamma^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 \\ &+ 2(1-\alpha\mu)^{2k}\|x_0 - x_*\|^2 + \frac{1-(1-\alpha\mu)^{2k}}{\mu(1-\frac{\alpha\mu}{2})}\frac{\alpha\sigma^2}{N},\end{aligned}$$

and for every $i = 1, 2, \ldots, N$ *and any* k,

$$\begin{aligned}\mathbb{E}\left\|x_i^{(k)} - x_*\right\|^2 \leq\ & 8\gamma^{2k}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{8D^2\alpha^2}{(1-\gamma)^2} + \frac{8\sigma^2 N\alpha^2}{(1-\gamma^2)} \\ &+ 16\left(\frac{\alpha}{\mu(1-\frac{\alpha L}{2})} + \frac{(1+\alpha L)^2}{\mu^2(1-\frac{\alpha L}{2})^2}\right)\left(\frac{L^2D^2\alpha^2}{N(1-\gamma)^2} + \frac{L^2\sigma^2\alpha^2}{(1-\gamma^2)}\right) \\ &+ 16\frac{\gamma^{2k} - \left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)^k}{\gamma^2 - 1 + \alpha\mu\left(1-\frac{\alpha L}{2}\right)}\frac{L^2\gamma^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 \\ &+ 4(1-\alpha\mu)^{2k}\|x_0 - x_*\|^2 + 2\frac{1-(1-\alpha\mu)^{2k}}{\mu(1-\frac{\alpha\mu}{2})}\frac{\alpha\sigma^2}{N},\end{aligned}$$

where

$$D^2 := 4L^2\mathbb{E}\left\|x^{(0)} - x^*\right\|^2 + 8L^2\frac{C_1^2\alpha^2 N}{(1-\gamma)^2} + \frac{2L^2\alpha\sigma^2 N}{\mu(1+\lambda_N^W - \alpha L)} + 4\left\|\nabla F\left(x^*\right)\right\|^2, \tag{27}$$

where γ, C_1 are given in (8)-(9).

Remark 4 (Convergence rate of the averaged D-SG iterates) *Note that given iteration budget $K > 0$, if we take $\alpha = \frac{\log(K)}{\mu K}$ in the setting of Proposition 3, then we have $(1-\alpha\mu)^{2K} = \mathcal{O}(1/K^2)$ and we obtain $\mathbb{E}\left\|\bar{x}^{(K)} - x_*\right\|^2 = \tilde{\mathcal{O}}(\frac{1}{NK} + \frac{1}{K^2})$ where $\tilde{\mathcal{O}}(\cdot)$ hides a logarithmic factor in K.*

2.3.3 Analysis of Distributed Accelerated Stochastic Gradient

Throughout this section, we state the results under the following assumption.

Assumption 2 *We assume all eigenvalues of W are positive, i.e., we assume that $\lambda_N^W > 0$.*

We note that Assumption 2 is not restrictive in the sense that even if the weight matrix W does not satisfy this assumption, we can still apply the results in our paper by considering the modified weight matrix $W_\tau := \frac{\tau}{\tau+1}I + \frac{1}{\tau+1}W$ for $\tau > 1$ instead of W. Because, we have $\lambda_N^{W_\tau} > \frac{\tau-1}{\tau+1} > 0$ for $\tau > 1$ and therefore W_τ satisfies Assumption 2. We will elaborate this point further after Corollary 8 in Remark 9.

The following result extends Aybat et al. (2020) from non-distributed ASG to D-ASG.

Theorem 5 *Assume there exist $\rho \in (0,1)$ and a positive semi-definite 2×2 matrix $\tilde{P}$ such that*

$$\rho^2\tilde{X}_1 + (1-\rho^2)\tilde{X}_2 \succeq \begin{bmatrix} \tilde{A}_{dasg}^T\tilde{P}\tilde{A}_{dasg} - \rho^2\tilde{P} & \tilde{A}_{dasg}^T\tilde{P}\tilde{B}_{dasg} \\ \tilde{B}_{dasg}^T\tilde{P}\tilde{A}_{dasg} & \tilde{B}_{dasg}^T\tilde{P}\tilde{B}_{dasg} \end{bmatrix}, \tag{28}$$

where $\tilde{A}_{dasg}$, $\tilde{B}_{dasg}$ and $\tilde{C}_{dasg}$ are defined in (22) *and*

$$\tilde{X}_1 := \begin{bmatrix} \frac{\beta^2\mu}{2} & \frac{-\beta^2\mu}{2} & \frac{-\beta}{2} \\ \frac{-\beta^2\mu}{2} & \frac{\beta^2\mu}{2} & \frac{\beta}{2} \\ \frac{-\beta}{2} & \frac{\beta}{2} & \frac{\alpha(1+\lambda_N^W - L\alpha)}{2} \end{bmatrix}, \quad \tilde{X}_2 := \begin{bmatrix} \frac{(1+\beta)^2\mu}{2} & \frac{-\beta(1+\beta)\mu}{2} & \frac{-(1+\beta)}{2} \\ \frac{-\beta(1+\beta)\mu}{2} & \frac{\beta^2\mu}{2} & \frac{\beta}{2} \\ \frac{-(1+\beta)}{2} & \frac{\beta}{2} & \frac{\alpha(1+\lambda_N^W - L\alpha)}{2} \end{bmatrix}.$$

Let $P = \tilde{P}\otimes I_{Nd}$. Then, for every $k \geq 0$,

$$\mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right] \leq \rho^{2k}\frac{2V_{P,\alpha,1}(\xi_0)}{\mu} + \frac{1}{1-\rho^2}\frac{2\alpha^2\sigma^2 N}{\mu}\left(\tilde{P}_{11} + \frac{1-\lambda_N^W + \alpha L}{2\alpha}\right). \tag{29}$$

Therefore, the robustness of D-ASG iterations defined in (14) *satisfies*

$$J_\infty \leq \frac{2\alpha^2}{\mu(1-\rho^2)}\left(\tilde{P}_{11} + \frac{1-\lambda_N^W + \alpha L}{2\alpha}\right).$$

With the additional assumption $\alpha \leq \frac{1}{L+\mu}$, we have the following corollary.

Corollary 6 *Under the assumptions in Theorem 5, if in addition, $\alpha \leq \frac{1}{L+\mu}$, then we have*

$$\mathbb{E}\left[\left\|x^{(k)} - x^*\right\|^2\right] \leq 4\rho^{2k}\frac{V_{P,\alpha,1}(\xi_0)}{\mu} + \frac{1}{1-\rho^2}\frac{4\alpha^2\sigma^2 N}{\mu}\left(\tilde{P}_{11} + \frac{1-\lambda_N^W+\alpha L}{2\alpha}\right) + \frac{2\alpha^2 C_1^2 N}{(1-\gamma)^2},$$

where γ, C_1 are given in (8)-(9).

The results in Theorem 5 are stated in terms of a 2×2 matrix $\tilde{P}$ which solves the 3×3 matrix inequality (28). For any fixed α, β and ρ; this is a linear matrix inequality (LMI). Therefore, we can compute $\tilde{P}$ numerically by varying α, β and ρ on a grid and then solving the resulting LMIs with a software such as CVX (Grant et al., 2008) (see also Lessard et al. (2016) for a similar approach). However, in the next result, we obtain some explicit performance bounds in the special case when $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$; this choice of β is motivated by the fact that it is a common choice in the non-distributed and noiseless setting.[5] The proof is deferred to Appendix B; it is based on the fact that when $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$, $\rho = 1-\sqrt{\alpha\mu}$ and $\alpha \in \left(0, \frac{\lambda_N^W}{L}\right]$; $\tilde{P} = \tilde{S}_\alpha$ is an explicit solution to the matrix inequality (28) where

$$\tilde{S}_\alpha := \begin{bmatrix} \frac{1}{2\alpha} & -\frac{1-\sqrt{\alpha\mu}}{2\alpha} \\ -\frac{1-\sqrt{\alpha\mu}}{2\alpha} & -\left(\frac{1-\sqrt{\alpha\mu}}{2\alpha}\right)^2 \end{bmatrix} = vv^T \quad \text{where} \quad v := \begin{bmatrix} \frac{1}{\sqrt{2\alpha}} \\ \sqrt{\frac{\mu}{2}} - \sqrt{\frac{1}{2\alpha}} \end{bmatrix}.$$

Then, plugging in $\tilde{P} = \tilde{S}_\alpha$ in Theorem 5 and in the bound (5), we obtain performance guarantees in terms of the Lyapunov function $V_{S_\alpha,\alpha,1}$. To simplify the notation in this case, with slight abuse of notation, we let

$$V_{S,\alpha}(\xi) := V_{S_\alpha,\alpha,1}(\xi) = \xi^T S_\alpha \xi + F_{\mathcal{W},\alpha}(T\xi + x^\infty) - F_{\mathcal{W},\alpha}(x^\infty). \tag{30}$$

We have the following explicit performance bounds on the convergence and the robustness of D-ASG.

Theorem 7 *Consider running D-ASG method with $\alpha \in \left(0, \frac{\lambda_N^W}{L}\right]$ and $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$. Then, for any $k \geq 0$, we have*

$$\mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right] \leq 2\left(1-\sqrt{\alpha\mu}\right)^k \frac{V_{S,\alpha}\left(\xi_0\right)}{\mu} + \frac{\sigma^2 N\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(2-\lambda_N^W+\alpha L\right). \tag{31}$$

Therefore, the robustness measure (defined in (14)*) satisfies*

$$J_\infty(\alpha) \leq \frac{\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(2-\lambda_N^W+\alpha L\right).$$

With the additional assumption $\alpha \leq \frac{1}{L+\mu}$, we have the following corollary.

5. Furthermore it can be shown that it gives the fastest rate for quadratic objectives in the non-distributed case when there is no noise (Aybat et al., 2019).

Corollary 8 *Under the assumptions in Theorem 7, if in addition, $\alpha \leq \frac{1}{L+\mu}$, then we have*

$$\mathbb{E}\left[\left\|x^{(k)} - x^*\right\|^2\right] \leq 4\left(1-\sqrt{\alpha\mu}\right)^k \frac{V_{S,\alpha}\left(\xi_0\right)}{\mu} + \frac{2\sigma^2 N\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(2-\lambda_N^W+\alpha L\right) + \frac{2C_1^2 N\alpha^2}{(1-\gamma)^2}, \quad (32)$$

where γ, C_1 are given in (8)-(9).

Remark 9 (Dependency to the spectral gap) *We can observe from Corollary 8 that among the three error terms in our performance bounds for D-ASG, only the last error term is about the network effect which depends on the spectral gap and this last error term is linear in $1/(1-\gamma)^2$, where $1-\gamma$ is the spectral gap of the matrix W. We discussed earlier that Assumption 2 is not restrictive because even if the matrix W does not satisfy Assumption 2, one can consider the modified weight matrix $W_1 := \frac{1}{2}I + \frac{1}{2}W$ which will satisfy Assumption 2. When we use the modified weight matrix W_1, the spectral gap may get smaller (i.e. spectral gap of W_1 can be smaller than that of W) and consequently the error term $1/(1-\gamma)^2$ due to network effects in Corollary 8 may get (worse) larger. However, the network error term can only get larger by a constant factor of 4. To explain this point further, assume that W does not satisfy Assumption 2. In this case the smallest eigenvalue λ_N^W of the mixing matrix W can be negative. We have two cases: (I) $|\lambda_N^W| > |\lambda_2^W|$; (ii) $|\lambda_N^W| \leq |\lambda_2^W|$. In case (I), i.e. when $|\lambda_N^W| > |\lambda_2^W|$, the spectral gap is determined by λ_N^W in the sense that we have the spectral gap $\Delta(W) := 1-|\lambda_N^W|$ and the spectral gap of the shifted matrix $\Delta(W_1) = \Delta(\frac{I+W}{2}) = \frac{1-|\lambda_2(W)|}{2}$ can be larger; for instance when λ_N^W is close enough to -1. If that is the case, then shifting the W matrix will result in an improved spectral gap and improved convergence guarantees. If on the other hand, λ_N^W is sufficiently far away from -1, then the spectral gap of the shifted matrix can be smaller, but by a factor of at most 2; in other words we would have $\Delta(W) = 1-|\lambda_N^W| \leq 2\Delta(W_1) = 2\Delta\left(\frac{I+W}{2}\right) = 1-|\lambda_2^W|$. In case (II), i.e. when $|\lambda_N^W| \leq |\lambda_2^W|$, we have the spectral gap $\Delta(W) = 1-|\lambda_2^W|$ whereas the spectral gap of the shifted matrix satisfies $\Delta(W_1) = \Delta(\frac{I+W}{2}) = \frac{1-|\lambda_2(W)|}{2} = \frac{\Delta(W)}{2}$. In this case, the spectral gap becomes worse, but only by a factor of 2. To summarize, shifting the W matrix to $W_2 = \frac{1}{2}I + \frac{1}{2}$ so that Assumption 2 can be satisfied might lead to improved convergence results in some cases, and in some cases it can make the convergence bounds looser; but this looseness in the spectral gap is at most by a constant factor of 2, and the last error term for D-ASG in Corollary 8 has the order of $O(\frac{\alpha^2}{1-\gamma^2}) = \mathcal{O}(\frac{\alpha^2}{\Delta^2(W)})$. Therefore, this term can become worse only by a constant factor of 4. This shows that Assumption 2 is not very restrictive in terms of iteration complexity results as it can hold for any graph topology and for a wide class of choices of W.*

With slight abuse of notation, we let

$$V_{\bar{S},\alpha}(\bar{\xi}) := \bar{\xi}^T \bar{S}_\alpha \bar{\xi} + f\left(\bar{T}\bar{\xi} + x_*\right) - f(x_*), \quad (33)$$

where

$$\bar{S}_\alpha := \tilde{S}_\alpha \otimes I_d,$$

and

$$\bar{T} := [1\ 0] \otimes I_d.$$

Using the Lyapunov function defined in (33), the next result establishes a performance bound for the node averages $\bar{x}^{(k)}$. We see that the variance term of our bound is proportional to $\frac{\sigma^2}{N}$ due to the averaging effect and is decreasing with N.

Proposition 10 *Consider the node averages $\bar{x}^{(k)}$ for the D-ASG algorithm with $0 < \alpha \leq \min\left\{\frac{1}{L+\mu}, \frac{\lambda_N^W}{L}\right\}$ and $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$ and the initialization $x^{(0)} = x^{(-1)} = 0$. For any k, we have*

$$\begin{aligned}
&\mathbb{E}\left\|\bar{x}^{(k)} - x_*\right\|^2 \\
&\leq \left(1 - \frac{\sqrt{\alpha\mu}}{2}\right)^k \frac{2V_{\bar{S},\alpha}\left(\bar{\xi}_0\right)}{\mu} + \frac{8}{\gamma^2\mu\sqrt{\mu}}\sqrt{\alpha}H_1H_3\frac{\gamma^{2k} - (1-\sqrt{\alpha\mu}/2)^k}{\gamma^2 - (1-\sqrt{\alpha\mu}/2)} + \frac{2}{\mu^2}\alpha H_1H_2 \\
&\qquad + \frac{2\sigma^2\sqrt{\alpha}}{\mu\sqrt{\mu}N}\left(1 + \frac{\sqrt{\alpha\mu}}{2}\right)(1+\alpha L),
\end{aligned}$$

and for every $i = 1, 2, \ldots, N$ and any k,

$$\begin{aligned}
&\mathbb{E}\left\|x_i^{(k)} - x_*\right\|^2 \\
&\leq \left(1 - \frac{\sqrt{\alpha\mu}}{2}\right)^k \frac{4V_{\bar{S},\alpha}\left(\bar{\xi}_0\right)}{\mu} + \frac{16}{\gamma^2\mu\sqrt{\mu}}\sqrt{\alpha}H_1H_3\frac{\gamma^{2k} - (1-\sqrt{\alpha\mu}/2)^k}{\gamma^2 - (1-\sqrt{\alpha\mu}/2)} + \frac{4}{\mu^2}\alpha H_1H_2 \\
&\quad + 16\gamma^{2k}\left(4\frac{V_{S,\alpha}\left(\xi_0\right)}{\mu} + \frac{2\sigma^2N\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(2 - \lambda_N^W + \alpha L\right) + \frac{2C_1^2N\alpha^2}{(1-\gamma)^2} + \|x^*\|^2\right) \\
&\qquad + \frac{8D_y^2\alpha^2}{(1-\gamma)^2} + \frac{8\sigma^2N\alpha^2}{(1-\gamma)^2} + \frac{16C_0\alpha}{(1-\gamma)^2} + \frac{4\sigma^2\sqrt{\alpha}}{\mu\sqrt{\mu}N}\left(1 + \frac{\sqrt{\alpha\mu}}{2}\right)(1+\alpha L),
\end{aligned}$$

where C_0 is a positive constant,[6] *and*

$$\begin{aligned}
D_y^2 := 4L^2\left((1+\beta)^2 + \beta^2\right)&\left(4\frac{V_{S,\alpha}\left(\xi_0\right)}{\mu} + \frac{2\sigma^2N\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(2 - \lambda_N^W + \alpha L\right) + \frac{2C_1^2N\alpha^2}{(1-\gamma)^2}\right) \\
&+ 2\left\|\nabla F\left(x^*\right)\right\|^2, \qquad (34)
\end{aligned}$$

and

$$\begin{aligned}
H_1 &:= 8\left(1 + \frac{\mu\alpha}{2} - \sqrt{\alpha\mu}\right) + \frac{2L^2\alpha}{\mu} + \left(L\alpha + 2 + \mu\alpha\right)\sqrt{\alpha\mu} - 2\mu\alpha, \\
H_2 &:= \frac{2}{N}L^2\left((1+\beta)^2 + \beta^2\right)\left(\frac{4D_y^2\alpha}{(1-\gamma)^2} + \frac{4\sigma^2N\alpha}{(1-\gamma)^2} + \frac{8C_0}{(1-\gamma)^2}\right), \\
H_3 &:= \frac{2}{N}L^2\left((1+\beta)^2 + \beta^2\right)\left(4\frac{V_{S,\alpha}\left(\xi_0\right)}{\mu} + \frac{2\sigma^2N\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(2 - \lambda_N^W + \alpha L\right) + \frac{2C_1^2N\alpha^2}{(1-\gamma)^2} + \|x^*\|^2\right).
\end{aligned}$$

6. An exact expression for the constant C_0 can be obtained from our proof technique. However, for the simplicity of the presentation, we did not specify the constant C_0 explicitly.

Remark 11 (Convergence rate of the averaged D-ASG iterates) *For a given iteration budget $K > 0$, if we take $\alpha = \frac{1}{\mu}\left(\frac{4\log(K)}{K}\right)^2$ in the setting of Proposition 10, then we have $(1 - \frac{\sqrt{\alpha\mu}}{2})^K = \mathcal{O}(1/K^2)$ as well as $\frac{\gamma^{2k} - (1-\sqrt{\alpha\mu}/2)^K}{\gamma^2 - (1-\sqrt{\alpha\mu}/2)} = \mathcal{O}(1/K^2)$. Consequently, we obtain $\mathbb{E}\left\|\bar{x}^{(K)} - x_*\right\|^2 = \tilde{\mathcal{O}}(\frac{1}{NK} + \frac{1}{K^2})$ where $\tilde{\mathcal{O}}(\cdot)$ hides a logarithmic factor in K.*

Constants in Theorem 7. λ_N^W and γ can typically be estimated with a distributed algorithm; for instance when $W = I - L$ (see e.g. Tran and Kibangou (2014)). For regularized problems of the form $f_i(x) = \tilde{f}_i(x) + \frac{\lambda}{2}\|x\|^2$ with $\tilde{f}_i$ convex, the parameter μ of strong convexity can be taken as the regularization parameter λ and therefore is known. The Lipschitz constant L can be estimated with a line search similar to Beck and Teboulle (2009); Schmidt et al. (2015). The constant C_1 depends on L, μ and σ explicitly.

We note that if we possess a lower bound $\underline{\mu}$ on the strong convexity parameter μ and an upper bound $\bar{L}$ on the strong convexity constant, our results in Theorem 1 and Theorem 7 will hold if replace $\underline{\mu}$ with μ and $\bar{L}$ with L. If a lower bound on the strong convexity constant cannot be estimated and if the strong convexity constant is instead over-estimated, it is known that this can lead to slower convergence, even for (centralized) SG and ASG. For example, if the strong convexity constant is overestimated by a factor of $c > 1$, i.e. the estimated constant is $\bar{\mu} = c\mu$ where μ is the actual strong convexity constant; convergence rate of (centralized) SG on some quadratic examples can be as slow as $\mathcal{O}(\frac{1}{k^{1/c}})$ (compared to the $\mathcal{O}(\frac{1}{k})$ rate that can be achieved if the strong convexity constant can be accurately estimated) (see e.g. Nemirovski et al. (2009)). Our bounds reflect a similar behavior. For example, for D-SG, with perfect knowledge of the strong convexity constant, for a given iteration budget K, we can choose the stepsize $\alpha = \frac{\log(K)}{2\mu K}$ and our Corollary 2 will lead to the bound $\mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right] = \tilde{\mathcal{O}}(1/K)$ where $\tilde{\mathcal{O}}(\cdot)$ hides some logarithmic factors in K. If we were to overestimate the strong convexity constant by a factor of c, the same stepsize choice will lead to a slower convergence rate of $\mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right] = \mathcal{O}(1/K^{1/c})$. Similar observations also hold for D-ASG. That being said, it is worth noting that, even for the deterministic and centralized case, Arjevani and Shamir (2016) have shown that for a wide class of algorithms including accelerated gradient methods, it is not possible to obtain accelerated rates, i.e. bounds of the form $L\|x_0 - x_*\|^2 \exp(\mathcal{O}(1)\frac{k}{\sqrt{\kappa}})$ after k iterations where $\kappa = L/\mu$ is the condition number, without having a good estimate (lower bound) of the strong convexity parameter. Therefore, it is somehow expected that to get the accelerated convergence rates, one needs to have some information about the problem constants such as μ and L.

We also note that in practice, for regularized problems such as L_2 regularized logistic regression or ridge regression, the regularizer $\frac{\lambda}{2}\|x\|^2$ provides a lower bound on μ directly (where we can simply take $\mu = \lambda$). If L and μ are known approximately, the stepsize can be set to $\alpha \in (0, (1 + \lambda_N^W)/L)$ for D-SG (Theorem 1) and the stepsize can be set to $\alpha \in (0, \lambda_N^W/L]$ and the momentum parameter can be set to $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$ for D-ASG (Theorem 7) as an initial guess and can be further tuned to the data set.

Robustness of D-SG vs D-ASG. We derived in Theorem 1 that for D-SG, for small stepsize α, the rate of convergence is $1 - \alpha\mu$ while $J_\infty(\alpha) \leq \frac{\alpha}{\mu(2-\alpha\mu)}$, and in Theorem 7

that for D-ASG, for small stepsize α and $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$, the rate of convergence is $1-\sqrt{\alpha\mu}$, while $J_\infty(\alpha) \leq \frac{\sqrt{\alpha}}{\mu\sqrt{\mu}}(2-\lambda_N^W+\alpha L) = \mathcal{O}\big(\frac{\sqrt{\alpha}}{\mu\sqrt{\mu}}\big)$. Hence, for a fixed α, D-ASG converges faster than D-SG, but is less robust and more sensitive to noise for the same stepsize that is small enough, and this suggests that there is a trade-off between convergence rate and robustness. Next, we discuss how one can trade between convergence rate and robustness in a more systematic manner.

Trading off convergence rate with the robustness and the network term. Equation (32) shows that large stepsize leads to faster rate $1-\sqrt{\alpha\mu}$, but the variance term (that is proportional to robustness J_∞) and the network term in our bounds get larger. Consider minimizing the sum of variance and network terms there, subject to a constraint on the rate:

$$\begin{aligned} &\min J_{tot}(\alpha) := \frac{2\sigma^2 N\alpha}{\mu\sqrt{\alpha\mu}}\left(2-\lambda_N^W+\alpha L\right) + \frac{2C_1^2 N\alpha^2}{(1-\gamma)^2}, \\ &\text{subject to} \quad 0\leq\alpha\leq\bar{\alpha}, \quad 1-\sqrt{\alpha\mu}\leq\rho_*(1+\delta), \end{aligned} \tag{35}$$

where $\bar{\alpha} := \min\left(\frac{\lambda_N^W}{L}, \frac{1}{L+\mu}\right)$ and $\rho_* := 1-\sqrt{\bar{\alpha}\mu}$ is the best rate we can certify with (32) and $\delta \in \left[0, \frac{1}{\rho_*}-1\right]$ is the percentage of the best achievable rate we would like to trade with robustness and network effects. The constraints specify an interval for the stepsize to lie in, and the objective J_{tot} can be optimized in this interval explicitly by calculating the first-order conditions. By letting $z := \sqrt{\alpha}$, it can be checked that the optimization problem (35) is equivalent to

$$\begin{aligned} &\min_{z\geq 0} G(z) := \frac{2\sigma^2 N}{\mu\sqrt{\mu}} z\left(2-\lambda_N^W+z^2 L\right) + \frac{2C_1^2}{(1-\gamma)^2} z^4, \\ &\text{subject to} \quad \sqrt{\bar{\alpha}} \geq z \geq \frac{1-\rho_*(1+\delta)}{\sqrt{\mu}}. \end{aligned}$$

We also have

$$G'(z) = \frac{2\sigma^2 N}{\mu\sqrt{\mu}}\left(2-\lambda_N^W\right) + \frac{6\sigma^2 N}{\mu\sqrt{\mu}} L z^2 + \frac{8C_1^2}{(1-\gamma)^2} z^3 > 0,$$

for any $z>0$ and hence $G(z)$ is strictly increasing. Therefore, the solution of the minimization problem is $z^* = \frac{1-\rho_*(1+\delta)}{\sqrt{\mu}}$, and the optimal stepsize is $\alpha^* = \frac{(1-\rho_*(1+\delta))^2}{\mu}$. This choice of stepsize will lead to the tightest performance bounds in our analysis for the same rate and provides some guidance about how the stepsize can be chosen.

2.4 Quadratic Objectives

Our study so far has been focused on strongly convex objectives. In Appendix C, we analyze the special case of strongly convex *quadratic* objectives when f_i is quadratic at every node i. Note that, in this case $F(x)$ is also quadratic. We obtain tight results in terms of rate and robustness that improve upon current results. In particular, we obtain the same convergence rate $\rho(\alpha)^2 = (1-\alpha\mu)^2$ for D-SG method but better convergence rate

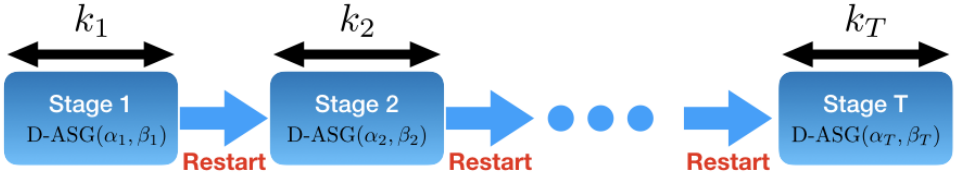


Figure 1: The scheme of the Distributed Multistage ASG (D-MASG) method

$\rho^2_{\text{dasg}} = (1 - \sqrt{\alpha\mu})^2$ for D-ASG method (instead of $\rho^2_{\text{dasg}} = 1 - \sqrt{\alpha\mu}$ for the strongly-convex setting). We also obtain explicit formulas for the robustness measure J_∞ for quadratic objectives for both D-SG and D-ASG (instead of upper bounds for the strongly-convex setting) under an additional assumption on the structure of the noise as well as explicit bounds on the asymptotic variance of the components of the node average vector $\bar{x}^{(k)}$.

3. An Exact Multistage Distributed Method

In the previous sections, we mainly focused on the D-SG and D-ASG methods with *constant* step size and momentum parameters. For these algorithms, we studied the problem of tuning their parameters so that the iterates converge to a neighborhood of x_* that depends on the stepsize α. In this section, however, our focus is to design a distributed *exact algorithm* that uses time-varying stepsize and momentum parameters and converges to the optimum x_* when the number of iterations grows.

We propose the Distributed Multistage ASG (D-MASG) method which is a distributed version of M-ASG proposed in Aybat et al. (2019). As illustrated in Figure 1, D-MASG consists of T stages where at each stage $t \in \{1, 2, \ldots, T\}$, we run D-ASG with parameters α_t and $\beta_t = \frac{1-\sqrt{\mu\alpha_t}}{1+\sqrt{\mu\alpha_t}}$ for n_t iterations where α_t and n_t will be chosen in a particular way. These stages are stitched together using a *momentum restart* technique which means that the first two iterates of every stage are equal to the last iterate of the previous stage. The details of D-MASG are provided in Algorithm 1 where the iterate $x^{t,m}$ denotes the m-th iterate of the t-th stage.

For any $t \leq T$, let L_t denote the total number of iterations up to the end of stage t, i.e,

$$L_t := \sum_{i=1}^{t} k_i, \tag{36}$$

with the convention that $L_0 := 0$. Let $x^{(k)}$ be the sequence that records all the inner and outer iterations of the D-MASG algorithm, obtained by concatenating the sequences $\{x^{(t,m)}\}_{m=1}^{k_t}$ for all stages t and inner iterates indexed by m. In other words, k is the counter for the total number of stochastic gradient evaluations and for $L_{t-1} < k \leq L_t$, we have

$$x^{(k)} = x^{(t,k-L_{t-1})}. \tag{37}$$

To characterize the convergence rate of D-MASG, we first analyze the evolution of iterates over one single stage. To simplify our presentation, we define the *scaled condition number*

Algorithm 1: Distributed Multistage Accelerated Stochastic Gradient Algorithm (D-MASG)

Input : Initial iterate $x^{(0)}$, The sequence $\{\alpha_i\}_{i=1}^T$ of stepsizes, The sequence $\{k_i\}_{i=1}^T$ of length of stages.

Set $x^{(0,k_0)} = x^{(0)}$;
for $t = 1;\ t \leq T;\ t = t+1$ **do**
  Set $x^{(t,-1)} = x^{(t,0)} = x^{(t-1,k_{t-1})}$;
  for $m = 0;\ m \leq k_t - 1;\ m = m+1$ **do**
    Set $\beta_t = \frac{1-\sqrt{\mu\alpha_t}}{1+\sqrt{\mu\alpha_t}}$;
    Set $y^{(t,m)} = (1+\beta_t)x^{(t,m)} - \beta_t x^{(t,m-1)}$;
    Set $x^{(t,m+1)} = \mathcal{W}y^{(t,m)} - \alpha_t \tilde{\nabla} f\left(y^{(t,m)}\right)$
  end
end

as

$$\tilde{\kappa} := \frac{L+\mu}{\mu\lambda_N^W} = \frac{\kappa+1}{\lambda_N^W}, \tag{38}$$

where we assume for the rest of this section that Assumption 2 holds, i.e. $\lambda_N^W > 0$.

Proposition 12 *Consider running D-ASG with initialization $x^{(-1)} = x^{(0)} = 0$ and parameters $\alpha \in (0, \bar{\alpha}]$ where $\bar{\alpha} = \min\left\{\frac{\lambda_N^W}{L}, \frac{1}{L+\mu}\right\}$ and $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$. Then, for any $k \geq 0$,*

$$\mathbb{E}\left[\left\|x^{(k)} - x^*\right\|^2\right] \leq 4\exp(-k\sqrt{\alpha\mu})\left\|x^{(0)} - x^*\right\|^2 + 6N\left(\frac{\sqrt{\alpha}}{\mu\sqrt{\mu}}\sigma^2 + \frac{C_1^2\alpha^2}{(1-\gamma)^2}\right),$$

where γ, C_1 are given in (8)-(9).

D-MASG with one stage is equivalent to running the D-ASG algorithm. Based on the previous result, we immediately obtain the following corollary which provides performance bounds for one-stage D-MASG.

Corollary 13 *Given k, consider running D-MASG for one stage with $k_1 = k$ and $\alpha_1 = \frac{\lambda_N^W}{L+\mu}\left(p\sqrt{\tilde{\kappa}}\log k/k\right)^2$ for some $p \geq 1$, where $\tilde{\kappa}$ is given in* (38). *Then, for any*

$$k \geq p\sqrt{\tilde{\kappa}}\max\left\{2\log\left(p\sqrt{\tilde{\kappa}}\right), e\right\},$$

we have

$$\mathbb{E}\left[\left\|x^{(k)} - x^*\right\|^2\right] \leq \frac{4}{k^p}\left\|x^{(0)} - x^*\right\|^2 + \frac{6Np\log k}{\mu^2 k}\left(\sigma^2 + \frac{C_1^2(p\log k)^3}{(1-\gamma)^2k^3}\right),$$

where γ, C_1 are given in (8)-(9).

In the next proposition, we propose a particular way to choose the stepsize α_t and the stage length k_t for every stage $t \in [1, T]$ and obtain performance guarantees for the distance to the optimum after T stages. In our proposed approach, the length of stages is geometrically increasing whereas the stepsize of each stage is chosen in a geometrically decaying manner. The length of the first stage k_1 can be an arbitrary positive integer and our performance bounds depends on how it is chosen.

Proposition 14 *Consider running D-MASG with the following parameters:*

$$k_1 \geq 1, \quad \alpha_1 = \frac{\lambda_N^W}{L+\mu}, \quad k_t = 2^t \left\lceil p\sqrt{\tilde{\kappa}} \log(2) \right\rceil, \quad \alpha_t = \frac{\lambda_N^W}{2^{2t}(L+\mu)},$$

with $p \geq 7$. Then, for any $t \geq 0$:

$$\mathbb{E}\left[\left\|x^{(L_{t+1})} - x^*\right\|^2\right] \leq \frac{4}{2^{(p-2)t}} \exp\left(-\frac{k_1}{\sqrt{\tilde{\kappa}}}\right) \left\|x^{(0)} - x^*\right\|^2 + \frac{12N\sigma^2}{2^t \mu^2 \sqrt{\tilde{\kappa}}} + \frac{12N}{2^{4t}} \left(\frac{C_1 \lambda_N^W}{L(1-\gamma)}\right)^2,$$

where γ, C_1 are given in (8)-(9) and $\tilde{\kappa}$ is given in (38).

The previous result gives performance bounds for last iterate of every stage. Using this result, we can also derive upper bounds for the error after k iterations as follows.

Proposition 15 *Consider running D-MASG with the parameters given in Proposition 14. Then, for any $k > k_1$:*

$$\begin{aligned}
&\mathbb{E}\left[\left\|x^{(k)} - x^*\right\|^2\right] \\
&\leq \mathcal{O}(1) \left(\left(\frac{6p\sqrt{\tilde{\kappa}}}{k-k_1}\right)^{p-2} \exp\left(-\frac{k_1}{\sqrt{\tilde{\kappa}}}\right) \left\|x^{(0)} - x^*\right\|^2 + \frac{Np\sigma^2}{\mu^2(k-k_1)} + \frac{Np^4 C_1^2 (1-\gamma)^{-2}}{\mu^2 (k-k_1)^4} \right),
\end{aligned}$$

where γ, C_1 are given in (8)-(9) and $\tilde{\kappa}$ is given in (38).

Note that Proposition 15 provides us with a degree of freedom in choosing k_1. In the following corollary we characterize two special cases. We omit the proof as it is a straightforward consequence of Proposition 15.

Corollary 16 *Consider running D-MASG with the parameters given in Proposition 14. In particular, by choosing $k_1 = \lceil (p-2)\log(6p\tilde{\kappa})\sqrt{\tilde{\kappa}} \rceil$, we have*

$$\mathbb{E}\left[\left\|x^{(k)} - x^*\right\|^2\right] \leq \mathcal{O}(1) \left(\frac{1}{k^{p-2}} \left\|x^{(0)} - x^*\right\|^2 + \frac{Np\sigma^2}{\mu^2 k} + \frac{Np^4 C_1^2 (1-\gamma)^{-2}}{\mu^2 k^4} \right),$$

for any $k \geq 2k_1$. Also, for a given number of iterations, k, by choosing $p = 7$ and $k_1 = \lceil \frac{k}{C} \rceil$ for some constant $C \geq 2$, we have

$$\mathbb{E}\left[\left\|x^{(k)} - x^*\right\|^2\right] \leq \mathcal{O}(1) \left(\exp\left(-\frac{k}{C\sqrt{\tilde{\kappa}}}\right) \left\|x^{(0)} - x^*\right\|^2 + \frac{N\sigma^2}{\mu^2 k} + \frac{NC_1^2 (1-\gamma)^{-2}}{\mu^2 k^4} \right),$$

for any $k \geq 2\sqrt{\tilde{\kappa}}$, where γ, C_1 are given in (8)-(9) and $\tilde{\kappa}$ is given in (38).

Note that our results also provide bounds on the number of iterations required to find an ϵ-solution, i.e. a point x^ϵ that satisfies $\mathbb{E}\left[\|x^\epsilon - x^*\|^2\right] \leq \epsilon$ for a given $\epsilon > 0$. This is obtained in the next corollary. We omit the proof as it follows directly from the previous corollary; by bounding bias, variance, and network effect terms, each by $\epsilon/3$.

Corollary 17 *Let $\epsilon > 0$ be an arbitrary positive number. Consider running D-MASG with the parameters given in Proposition 14. Assume choosing $p = 7$ and $k_1 = \lceil \sqrt{\tilde{\kappa}} \log\left(\frac{\Delta}{\epsilon}\right) \rceil$ where Δ is the optimality gap, an upper bound on the initial error, i.e., $\Delta \geq \left\|x^{(0)} - x^*\right\|^2$. Then, D-MASG leads to an ϵ-close solution x^ϵ after at most*

$$\mathcal{O}(1)\left(\sqrt{\tilde{\kappa}}\log\left(\frac{\Delta}{\epsilon}\right) + \frac{N\sigma^2}{\mu^2\epsilon} + \frac{N^{1/4}\sqrt{C_1(1-\gamma)^{-1}}}{\sqrt{\mu}\sqrt[4]{\epsilon}}\right) \tag{39}$$

iterations, where γ, C_1 are given in (8)-(9) and $\tilde{\kappa}$ is given in (38).

Previously, we obtained optimal convergence results for the average iterates and individual iterates (Proposition 10). Similar to Corollary 16, we have the following result. We omit the proof as it follows directly from the previous corollary; by bounding bias, variance, and network effect terms, each by $\epsilon/3$.

Corollary 18 *Consider running D-MASG with the parameters given in Proposition 14. In particular, by choosing $k_1 = \lceil (p-2)\log(6p\tilde{\kappa})\sqrt{\tilde{\kappa}} \rceil$, we have*

$$\mathbb{E}\left[\left\|\bar{x}^{(k)} - x_*\right\|^2\right] \leq \mathcal{O}(1)\left(\frac{1}{k^{p-2}}\frac{\left\|x^{(0)} - x^*\right\|^2}{N} + \frac{p\sigma^2}{N\mu\sqrt{\mu}k} + \frac{p^4 C_0 L^2 (1-\gamma)^{-2}}{N\mu^2 k^4}\right),$$

$$\mathbb{E}\left[\left\|x_i^{(k)} - x_*\right\|^2\right]$$

$$\leq \mathcal{O}(1)\left(\frac{1}{k^{p-2}}\frac{\left\|x^{(0)} - x^*\right\|^2}{N} + \frac{p\sigma^2}{N\mu\sqrt{\mu}k} + \left(\frac{L^2}{N\mu^2} + 1\right)\frac{p^4 C_0 (1-\gamma)^{-2}}{k^4}\right),$$

for any $k \geq 2k_1$ and $i = 1, 2, \ldots, N$. Also, for a given number of iterations, k, by choosing $p = 7$ and $k_1 = \lceil \frac{k}{C} \rceil$ for some constant $C \geq 2$, we have

$$\mathbb{E}\left[\left\|\bar{x}^{(k)} - x_*\right\|^2\right] \leq \mathcal{O}(1)\left(\exp\left(-\frac{k}{C\sqrt{\tilde{\kappa}}}\right)\frac{\left\|x^{(0)} - x^*\right\|^2}{N} + \frac{\sigma^2}{N\mu\sqrt{\mu}k} + \frac{C_0 L^2 (1-\gamma)^{-2}}{N\mu^2 k^4}\right),$$

$$\mathbb{E}\left[\left\|x_i^{(k)} - x_*\right\|^2\right]$$

$$\leq \mathcal{O}(1)\left(\exp\left(-\frac{k}{C\sqrt{\tilde{\kappa}}}\right)\frac{\left\|x^{(0)} - x^*\right\|^2}{N} + \frac{\sigma^2}{N\mu\sqrt{\mu}k} + \left(\frac{L^2}{N\mu^2} + 1\right)\frac{C_0 (1-\gamma)^{-2}}{k^4}\right),$$

for any $k \geq 2\sqrt{\tilde{\kappa}}$ and $i = 1, 2, \ldots, N$, where γ, C_1 are given in (8)-(9) and $\tilde{\kappa}$ is given in (38).

Similar to Corollary 17, we can also provide bounds on the number of iterations required to find an ϵ-solution for the average iterates and an individual iterate, i.e. a point $\bar{x}^\epsilon := \frac{1}{N}\sum_{i=1}^N x_i^\epsilon$ that satisfies $\mathbb{E}\left[\|\bar{x}^\epsilon - x_*\|^2\right] \leq \epsilon$ and $\mathbb{E}\left[\|x_i^\epsilon - x_*\|^2\right] \leq \epsilon$ for a given $\epsilon > 0$. We have the following result.

Corollary 19 *Let $\epsilon > 0$ be an arbitrary positive number. Consider running D-MASG with the parameters given in Proposition 14. Assume choosing $p = 7$ and $k_1 = \lceil \sqrt{\tilde{\kappa}} \log\left(\frac{\Delta}{N\epsilon}\right)\rceil$ where Δ is the optimality gap, an upper bound on the initial error, i.e., $\Delta \geq \left\|x^{(0)} - x^*\right\|^2$. Then, D-MASG leads to an ϵ-close solution $\bar{x}^\epsilon$ after at most*

$$\mathcal{O}(1)\left(\sqrt{\tilde{\kappa}}\log\left(\frac{\Delta}{N\epsilon}\right) + \frac{\sigma^2}{\mu\sqrt{\mu}N\epsilon} + \frac{C_0^{1/4}\sqrt{L(1-\gamma)^{-1}}}{N^{1/4}\sqrt{\mu}\sqrt[4]{\epsilon}}\right) \tag{40}$$

iterations, and for any $1 \leq i \leq k$, D-MASG leads to an ϵ-close solution x_i^ϵ after at most

$$\mathcal{O}(1)\left(\sqrt{\tilde{\kappa}}\log\left(\frac{\Delta}{N\epsilon}\right) + \frac{\sigma^2}{\mu\sqrt{\mu}N\epsilon} + \left(\frac{\sqrt{L}}{N^{1/4}\sqrt{\mu}} + 1\right)\frac{C_0^{1/4}\sqrt{(1-\gamma)^{-1}}}{\sqrt[4]{\epsilon}}\right) \tag{41}$$

iterations, where C_0 is an explicitly computable constant such that $C_0 = \mathcal{O}(1)$ as $\alpha \to 0$ and γ is given in (8) and $\tilde{\kappa}$ is given in (38).

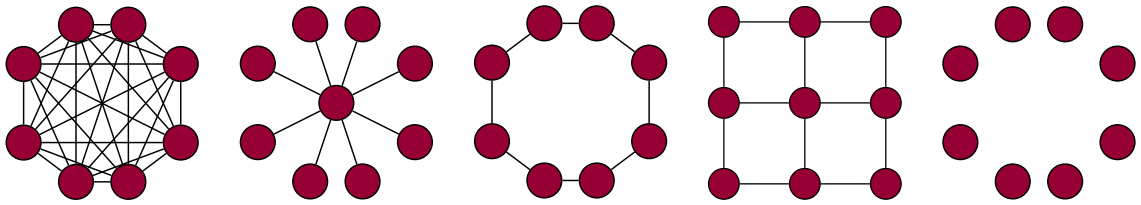
4. Numerical Results

In this section, we conduct several experiments to validate our theory and assess the performance of D-SG and D-ASG. We consider a (regularized) logistic regression problem which is a common formulation to solve binary classification tasks:

$$\min_{x\in\mathbb{R}^d}\left(\frac{1}{n}\sum_{i=1}^n \log\left(1+\exp\left(-y_i X_i^\top x\right)\right) + \lambda\|x\|_2^2\right), \tag{42}$$

where (X_i, y_i) denotes a data pair: $X_i \in \mathbb{R}^d$ is the feature vector and $y_i \in \{-1, 1\}$ denotes the label, and n denotes the number of data pairs.

In all our experiments, we assume that each computation node has access to a subset of all data points, and the noisy gradient in (6) and (13) basically becomes the *stochastic gradient* that is computed on a random sub-sample of the data. More precisely, we will assume that at each iteration, each computation node will draw a random sub-sample from the data points that it has access to, and compute the stochastic gradient by using this subsample. The size of the subsample will be determined by a single parameter $b \in (0, 1]$, which determines the ratio of the number of elements contained in the subsample to the total number of data points that are accessible to that node. For instance, if all the data points are evenly distributed to the nodes, i.e. each node has access to n/N number of distinct data points, the size of the data sub-sample that will be used for computing the stochastic gradients is determined as $(bn)/N$. If $b = 1$, the node will use all of its data points to compute the gradient, hence the variance of the gradient noise σ^2 will vanish. Similarly, a small $b \ll 1$ will result in a large σ^2.



(a) Fully-connected (b) Star (c) Circular (d) Grid (e) Disconnected

Figure 2: Illustration of the network architectures.

In the sequel, we first conduct experiments on a synthetic problem, which provides us a more sterilized environment where we have a direct control on the problem. Then, we conduct experiments on two binary classification data sets, where we implement the proposed algorithms and the competitors in C++ and run them on a real distributed environment. We will consider five different network architectures: (i) Connected: the network nodes can communicate with all the other nodes in the network, (ii) Star: the network nodes are only allowed to communicate with a central node (iii) Circular: the network notes are only allowed to communicate with their 'right' and 'left' neighbors, (iv) Grid: the network nodes are allowed to communicate with their upper, lower, left, and right neighbors, and finally (v) Disconnected: the network nodes are not allowed to communicate. These architectures are visualized in Figure 2. We note that we replicate each experiment 5 times and we report the average results. Finally, in our last experiments, we monitor the robustness of the proposed algorithms to potential inaccuracies in estimating the problem constants L and μ.

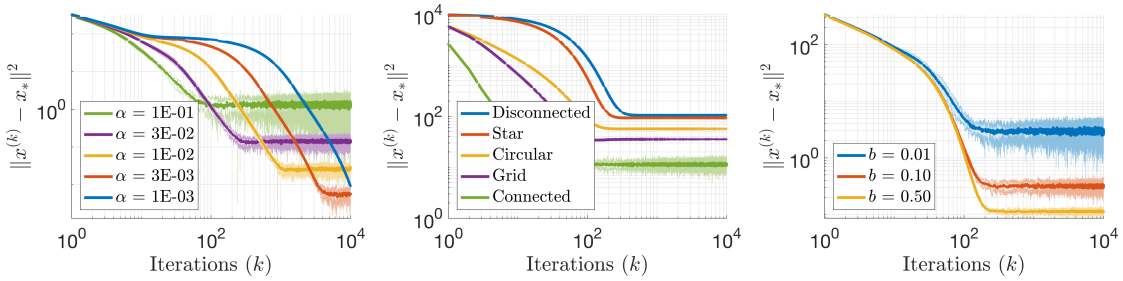
4.1 Synthetic Data Experiments

In this section, we present our experiments on a synthetic logistic regression problem, where our main goal is to validate Theorems 1 and 5 on the logistic regression task. In this set of experiments, we first generate synthetic data by simulating the following probabilistic model:

$$x_0 \sim \mathcal{N}(0, I), \qquad X_i \sim \mathcal{N}\left(0, \sigma_X^2 I\right), \qquad y_i | X_i, x_0 \sim \delta\left(y_i - \text{sign}\left(X_i^\top x_0\right)\right),$$

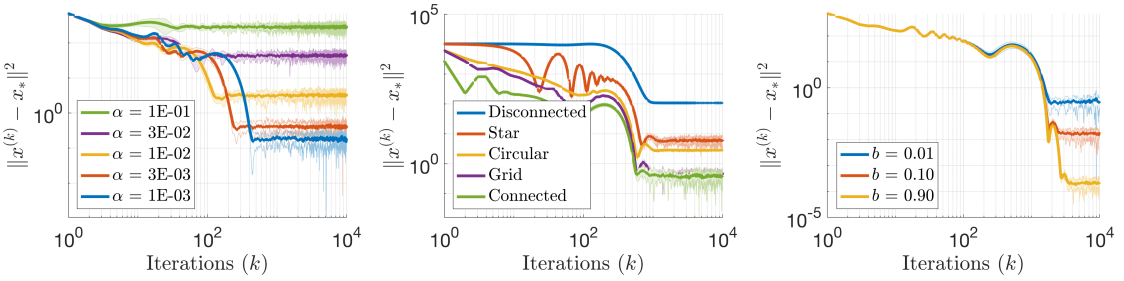
where x_0 denotes the data generating parameter and δ denotes the Dirac delta function to represent deterministic relations as a degenerate probability model. Once the set of pairs $(X_i, y_i)_{i=1}^n$ are generated, our goal becomes solving an ℓ_2-regularized logistic regression problem defined in (42). In this set of experiments, we simulate the distributed environment in MATLAB and we provide our implementation in the supplementary material. Unless stated otherwise, we first generate $n = 1000$ data points, set the dimension $d = 100$, data variance $\sigma_X^2 = 5$, $\lambda = 0.05$, the number of nodes $N = 10$, the batch proportion $b = 0.1$, $\alpha = 0.1$, and we consider the circular network architecture.

Figure 3 illustrates the results for D-SG. In Figure 3(a), we investigate the convergence behavior of D-SG for varying step-size α. The results clearly demonstrate the trade off between the convergence rate and the asymptotic variance: for larger α the algorithm



(a) Step-size (b) Network architecture (c) Minibatch size ($\alpha = 0.1$)

Figure 3: Synthetic data experiments on D-SG. The highlighted areas represent 3 standard deviations.



(a) Step-size (b) Network architecture (c) Minibatch size ($\alpha = 6 \times 10^{-5}$)

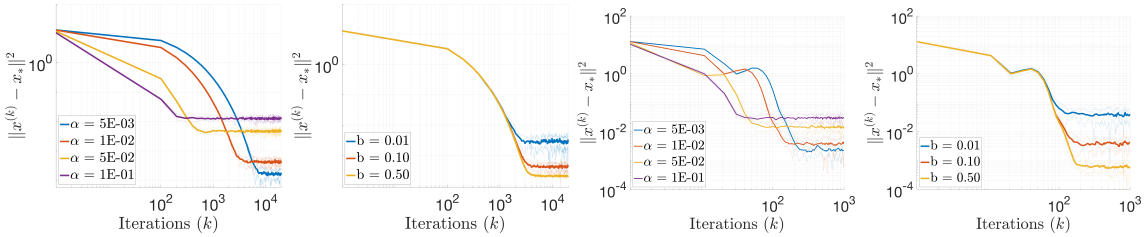
Figure 4: Synthetic data experiments on D-ASG. The highlighted areas represent 3 standard deviations.

attains a faster convergence rate but the resulting asymptotic variance becomes larger, as indicated by Theorem 1.

In the next experiment, we investigate the performance of D-SG for varying network architectures. In this setting we set $N = 1000$ in order to illustrate the differences more clearly. As illustrated in Figure 3(b), the results are intuitive: we observe that the disconnected graph non-surprisingly has the largest asymptotic variance. Furthermore, the performance improves as the graph becomes more connected: the performance of the (fully-connected) connected network is the best and degrades gradually as we go from the grid topology to the star topology.

In our third experiment, we investigate the effect of the noise variance σ^2 by altering the batch proportion b. As shown in Figure 3(c), decreasing the batch size results in an increased asymptotic variance. This behavior is also correctly captured by Theorem 1: decreasing b increases the noise variance σ^2 and hence the second term in (26) dominates for large number of iterations.

In our next set of experiments, we replicate the previous three experiments by replacing D-SG with D-ASG. Figure 4 illustrates the results. We observe a similar outcome to the ones of the previous set of experiments. Figure 4(a) verifies that the step-size determines



(a) D-SG (b) D-ASG

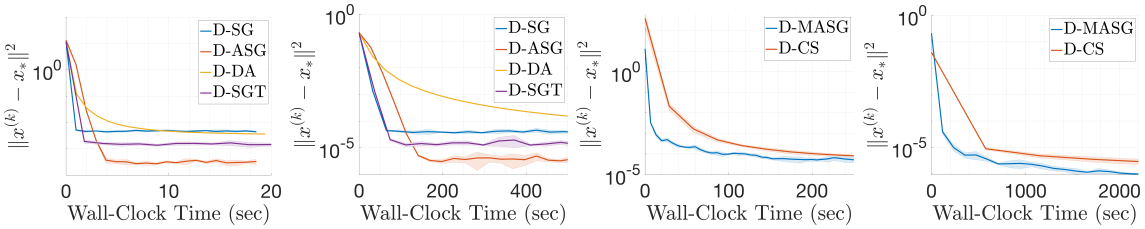
Figure 5: Evaluation of D-SG and D-ASG on MNIST and a real distributed environment with $N = 10$ interconnected computers. Thin lines represent 3 standard deviations.

the trade off between the convergence rate and the asymptotic variance as suggested by Theorem 5. Figure 4(b) illustrates the behavior of the algorithm under different network settings with $N = 1000$. We again observe that the disconnected network is performing worse than the other network architectures as expected; however, as opposed to Figure 3(b), there is no significant difference between the grid and the connected networks. This result suggests that the usage of the momentum in D-ASG compensates the additional difficulty introduced by the sparsely connected network architecture. In our last experiment, we investigate the behavior of D-ASG for varying gradient noise variance. As illustrated in Figure 4(c), the asymptotic error increases with the decreasing batch proportion b. More importantly, compared to D-SG, the increase in the asymptotic variance turns out to be significantly larger for D-ASG, which illustrates that D-ASG is less robust to the gradient noise. This observation also supports our theory (cf. the remark about robustness in Section 2.3.3).

4.2 Real Data Experiments

In this section, we consider a real-data setting, where we evaluate the algorithms on a real distributed environment. We consider the same logistic regression problem on two binary classification data sets and compare the performance of D-SG and D-ASG with their natural competitors, namely distributed dual averaging (D-DA) (Duchi et al., 2012a), distributed stochastic gradient tracking (D-SGT) (Pu and Nedić, 2021), and distributed communication sliding (D-CS) (Lan et al., 2020). Among these algorithms D-CS is an exact algorithm, similar to D-MASG. As data sets, we use the MNIST, and the Epsilon data sets. The MNIST data set contains 70K binary images (of size $d = 20 \times 20$) corresponding to 10 different digits.[7] To obtain a binary classification problem, we extract the images corresponding to the digits 0 and 8, where we end up with $n = 11774$ images in total. On the other hand, the Epsilon data set is one of the standard binary classification data sets,[8] and contains $n = 400$K samples with $d = 2000$.

7. `http://yann.lecun.com/exdb/mnist`.
8. `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`.



(a) MNIST (b) Epsilon (c) MNIST (d) Epsilon

Figure 6: (a)-(b) Comparison of D-SG and D-ASG with D-DA and D-SGT on the two data sets. (c)-(d) Comparison of D-MASG and D-CS on the two data sets. The highlighted areas represent 1 standard deviation.

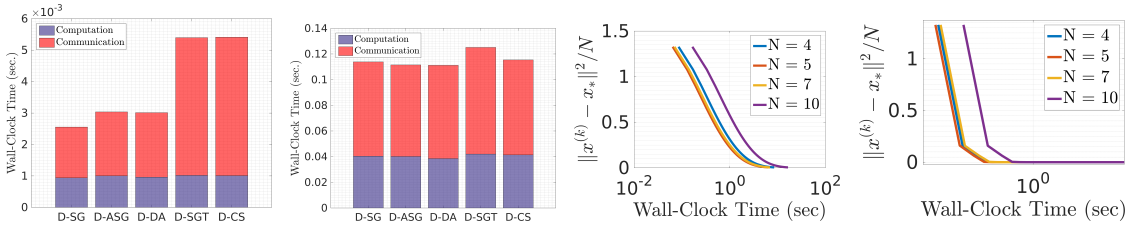
We have implemented all the algorithms in C++ by using a low-level message passing protocol for parallel processing, namely the OpenMPI library.[9] In order to have a realistic experimental environment, we have conducted these experiments on a cluster interconnected computers, each of which is equipped with different quality CPUs and memories. We set $b = 0.1$ unless stated otherwise.

In the first experiment, similar to the previous section, we monitor the behavior of D-SG and D-ASG with varying step-sizes and batch proportions in order to affirm that our theoretical results also hold in the real problem setting. Figure 5 illustrates the results. We observe that, even under the real data/distributed environment setting the algorithms exhibit the same behavior. The trade off between the convergence rate and the asymptotic variance is still present and D-ASG is significantly less robust to the stochastic gradient noise.

In our next experiment, we compare the performances of the inexact algorithms, namely D-SG and D-ASG with D-DA and D-SGT on the two data sets. The results are illustrated in Figure 6(a)-(b). In all settings, we observe that the performance of D-SG and D-DA are very similar, whereas the variance reduction step improves the performance of D-SGT over these two algorithms. The results show that D-ASG outperforms all these three algorithms and illustrate the acceleration brought by the use of momentum.

We then proceed to comparing the exact algorithms D-MASG and D-CS. We note that the D-CS algorithm has two levels of nested iterations: an outer iteration and an inner iteration. At each outer iteration the algorithm makes the nodes communicate two times, whereas the actual optimization is done in the inner iteration and the number of inner iterations can be varied depending on the communication cost: if the communication cost is high, the number of inner iterations should be high as well in order to make the communications less often. In order to make the wall-clock-time comparison between D-CS and D-MASG fairer, we set the number of inner iterations to 2, since D-MASG has only one round of communications at every iteration. We also note that the computational requirements of each inner iteration of D-CS are significantly higher than the one of D-MASG.

9. `https://www.open-mpi.org`.



(a) MNIST b=0.5 (b) Epsilon b=0.1 (c) MNIST D-SG (d) MNIST D-ASG

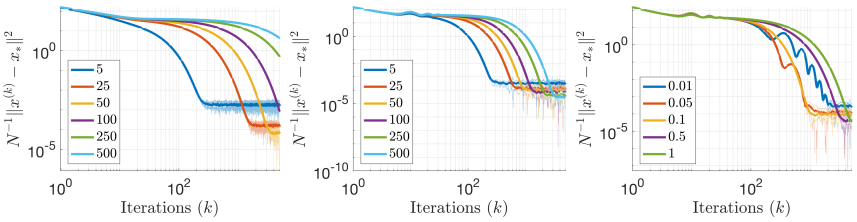
Figure 7: Investigation of the computational requirements.

We first investigated the performance of D-CS and D-MASG under the circular network setting. As opposed to the previous experiments, we did not observe a significant performance improvement over D-CS. We suspect that the Polyak-Ruppert-type averaging of D-CS is providing some acceleration to D-CS. However, when we evaluate the two algorithms under the connected network setting, we obtain improved results, which are visualized in Figure 6 (c)-(d). The results show that, on the MNIST data set D-MASG provides a slight improvement over D-CS, whereas on the Epsilon data set the difference between the computational costs of D-CS and D-MASG become more prominent, which yields a significant improvement over D-CS.

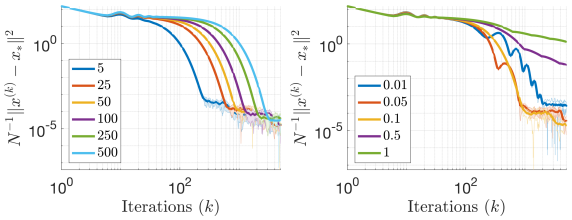
Next, we investigate the computational aspects of the aforementioned algorithms. In Figures 7(a) and 7(b) we measure the average times that the algorithms spend in terms of computation and communication per iteration. We observe that in both cases, the computation times of the algorithms is similar to each other. On the other hand, when the dimension of the problem is smaller (in the case of MNIST), the communication cost of D-SGT and D-CS dominates the overall complexity.[10] However, when the dimension of the problem increases (in the case of Epsilon), the computation time increases superlinearly with the increasing dimension, which results in a similar proportion of computation/communication for all the algorithms. Combined with the performance comparison results (e.g. Figure 6), this experiment suggests that D-ASG achieves a good balance between computational complexity and accuracy: while having similar computational complexity to D-SG and D-DA, it is able to provide better performance than D-SGT and D-CS, which have larger computational costs.

In our final experiment, we investigate the behavior of D-SG and D-ASG on the increasing number computation nodes N (while keeping all the other parameters unchanged). Figures 7(c) and 7(d) show the results. We observe that, the convergence behavior improves when we increase N from 4 to 5; however, further increasing N results in a degraded performance, since the overall computation time is dominated by the communication cost, a typical situation observed in synchronized distributed optimization (Kaya et al., 2019; Şimşekli et al., 2018).

10. In this experiment, the number of inner iterations of D-CS is set to 1.



(a) Varying L for D-SG (b) Varying L (left) and μ (right) for D-ASG



(c) Varying L (left) and μ (right) for D-MASG

Figure 8: Change in performance with respect to inaccurate estimates for L and μ. The highlighted areas represent 3 standard deviations.

4.3 Robustness to Hyperparameters

In our last set of experiments, we aim at investigating the performance of our algorithms in the case where the problem constants L and μ cannot be estimated accurately. Here, we re-consider the MNIST data set in the simulated distributed environment and run the three proposed algorithms for different estimates for L and μ. For D-SG we set the step-size $\alpha = (1 + \lambda_N^W)/L$, whereas for D-ASG we set $\alpha = \lambda_N^W/L$ and $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$. Finally, for D-MASG we set $\tilde{\kappa} = \frac{L+\mu}{\mu\lambda_N^W}$ as in (38) and use the setting reported in Proposition 14.

In this problem, we first compute an estimate for L and μ from the data, where we obtain $L \approx 50$ and $\mu \approx 0.1$. Then, we vary L from 5 to 500 by fixing $\mu = 0.1$, and we vary μ from 0.01 to 1 by fixing $L = 50$. Accordingly, we run the algorithms with hyperparameters that are computed with these values for L and μ. Figure 8 visualizes the results. In Figure 8(a), we observe that, when L is set close to 50, D-SG performs similarly, whereas for lower or higher values of L the performance degrades. On the other hand, in Figure 8(b), we observe that D-ASG is also robust to the values of L and μ: the performance of the algorithm does not significantly vary for varying L and μ. Finally, Figure 8(c) illustrates the performance of D-MASG. Here, in terms of varying L, we again observe a robust behavior, where the performance of the algorithm stays almost the same for different values of L. On the other hand, we also observe that the algorithm has a strong dependency on the estimate of μ, where an overestimation of the value of μ might significantly slow down the convergence.

We conclude that, when a reasonably good estimate for L and μ can be obtained, D-SG and D-MASG perform well. We also observe that on this data set the performance of D-ASG is robust when subject to changes in the parameters L and μ.

5. Conclusion

Stochastic gradient (SG) methods are workhorse algorithms in machine learning practice. There is an increasing need to run stochastic gradient methods in distributed environments, either because the data is inherently distributed (for instance when collected by autonomous units such as smart phones or sensors) and processing it in a non-distributed way is impractical for real-time decision making, or the data is non-distributed but due to its volume distributing the data to multiple computational units become unavoidable for scalability reasons. This motivates the study of the performance of SG methods on arbitrary networks where there the performance depends on the interplay between the bias, variance and network effects. In this paper, we focused on distributed stochastic gradient (D-SG) and its accelerated version (D-ASG) with constant and decaying stepsize. We provided a number of convergence results for D-SG and D-ASG that improve the existing convergence results. Our performance bounds captures the trade-offs in the bias, variance terms and the network effects and are illustrated by our numerical experiments. We also proposed a multi-stage variant of D-ASG with an optimal dependency to bias and variance terms. In this work, we considered synchronous algorithms which require nodes to update their local copies synchronously. As part of future work, it would be interesting to study momentum acceleration in the context of asynchronous stochastic gradient algorithms where the nodes can do updates without requiring synchronization between the nodes.

Acknowledgments

The authors are grateful to the Associate Editor and three anonymous referees for helpful suggestions and comments. Alireza Fallah acknowledges support from MathWorks Engineering Fellowship. Mert Gürbüzbalaban's research is supported in part by the grants Office of Naval Research Award Number N00014-21-1-2244, National Science Foundation (NSF) CCF-1814888, NSF DMS-2053485, NSF DMS-1723085. Umut Şimşekli's research is partly supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements davenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). Lingjiong Zhu is grateful to the partial support from a Simons Foundation Collaboration Grant and the grants NSF DMS-2053454 and NSF DMS-2208303 from the National Science Foundation.

Appendix A. Intermediate Results

Lemma 20 *(Yuan et al., 2016, Corollary 9) Recall the definition of*

$$x^{\infty}=\left[\left(x_1^{\infty}\right)^T,\left(x_2^{\infty}\right)^T,\cdots,\left(x_N^{\infty}\right)^T\right]^T,$$

which is the unique fixed point of

$$(I_{Nd}-\mathcal{W})x^{\infty}+\alpha\nabla F(x^{\infty})=0. \tag{43}$$

If $\alpha\leq\min\left\{\frac{1+\lambda_N^W}{L},\frac{1}{L+\mu}\right\}$, *then*

$$\|x_i^{\infty}-x_*\|\leq C_1\frac{\alpha}{1-\gamma}=\mathcal{O}\left(\frac{\alpha}{1-\gamma}\right),$$

where x_* *is the solution to the optimization problem* (1) *and recall the definitions of* C_1, f_i^*, *and* γ*:*

$$C_1=\sqrt{2L\sum_{i=1}^{N}\left(f_i\left(0\right)-f_i^*\right)}\cdot\left(1+\frac{2(L+\mu)}{\mu}\right),\quad f_i^*=\min_{x\in\mathbb{R}^d}f_i(x),\quad \gamma=\max\left\{\left|\lambda_2^W\right|,\left|\lambda_N^W\right|\right\}.$$

Proof According to Corollary 9 in Yuan et al. (2016),

$$\|x_i^{\infty}-x_*\|\leq\frac{c_4}{\sqrt{1-c_3^2}}+\frac{\alpha\hat{D}}{1-\gamma},$$

where

$$\hat{D}:=\sqrt{2L\sum_{i=1}^{N}\left(f_i\left(0\right)-f_i^*\right)},$$

and

$$c_4:=\alpha^{3/2}\sqrt{\alpha+\delta^{-1}}\frac{L\hat{D}}{1-\gamma},\qquad c_3:=\sqrt{1-\alpha c_2+\alpha\delta-\alpha^2\delta c_2},$$

where

$$\delta:=\frac{c_2}{2(1-\alpha c_2)},\qquad c_2:=\frac{\mu L}{\mu+L}.$$

Hence, we can compute that

$$\begin{aligned}\frac{c_4}{\sqrt{1-c_3^2}}&=\frac{\sqrt{2}\alpha\sqrt{\alpha+\delta^{-1}}L\hat{D}}{\sqrt{c_2}(1-\gamma)}=\frac{\sqrt{2}\alpha\sqrt{\frac{2(\mu+L)}{\mu L}-\alpha}}{\sqrt{\frac{\mu L}{\mu+L}}}\frac{L\hat{D}}{1-\gamma}\\&\leq\frac{\sqrt{2}\alpha\sqrt{\frac{2(\mu+L)}{\mu L}}}{\sqrt{\frac{\mu L}{\mu+L}}}\frac{L\hat{D}}{1-\gamma}=\frac{2(\mu+L)}{\mu}\frac{\alpha\hat{D}}{1-\gamma},\end{aligned}$$

where the first equality follows from the fact that

$$1 - c_3^2 = \alpha c_2 + \alpha^2 \delta c_2 - \alpha\delta = \alpha c_2 \left(1 + \alpha\delta - \frac{\delta}{c_2}\right) = \frac{\alpha c_2}{2}.$$

The proof is complete. ■

Lemma 21 *Recall the definitions of $x^{(k)}$ and $\bar{x}^{(k)}$ as*

$$\begin{aligned} x^{(k)} &= \left[\left(x_1^{(k)}\right)^T, \left(x_2^{(k)}\right)^T, \ldots, \left(x_N^{(k)}\right)^T\right]^T \in \mathbb{R}^{Nd}, \\ \bar{x}^{(k)} &= \frac{1}{N}\sum_{i=1}^N x_i^{(k)} \in \mathbb{R}^d. \end{aligned} \tag{44}$$

Then, for any $k \in \mathbb{N}$, we have

$$\mathbb{E}\left\|\bar{x}^{(k)} - x_*\right\|^2 \leq \frac{1}{N}\mathbb{E}\left\|x^{(k)} - x^*\right\|^2,$$

where x_* is the solution to the optimization problem (1) *and $x^* = \left[x_*^T, \ldots, x_*^T\right]^T$.*

Proof Note that the function $x \mapsto \|x - x_*\|^2$ is convex. Therefore, by Jensen's inequality,

$$\left\|\bar{x}^{(k)} - x^*\right\|^2 = \left\|\frac{1}{N}\sum_{i=1}^N x_i^{(k)} - x^*\right\|^2 \leq \frac{1}{N}\sum_{i}^N \left\|x_i^{(k)} - x_*\right\|^2 = \frac{1}{N}\left\|x^{(k)} - x^*\right\|^2.$$

By taking the expectations, we obtain the desired result. ■

Appendix B. Proofs of Main Results in Section 2

B.1 Proofs of Main Results in Section 2.3.2

Before we proceed to the proof of Theorem 1, let us first state the following result from Aybat et al. (2019) which is stated for Nesterov's accelerated stochastic gradient method but holds for stochastic gradient descent as well, as it is the special case of Nesterov's algorithm for $\beta = 0$.

Lemma 22 (Lemma B.1, Aybat et al. (2019)) *Let $P = p \otimes I_{Nd}$ for some $p \geq 0$ and recall the Lyapunov function $V_P(\xi) = \xi^\top P \xi$. Then we have*

$$\begin{aligned} &\mathbb{E}[V_P(\xi_{k+1})] - \rho^2 \mathbb{E}[V_P(\xi_k)] \\ &\quad \leq \mathbb{E}\left[\begin{bmatrix} \xi_k \\ \nabla F\left(x^{(k)}\right) \end{bmatrix}^\top \begin{bmatrix} A^\top P A - \rho^2 P & A^\top P B \\ B^\top P A & B^\top P B \end{bmatrix} \begin{bmatrix} \xi_k \\ \nabla F\left(x^{(k)}\right) \end{bmatrix}\right] + N\sigma^2\alpha^2 p. \end{aligned} \tag{45}$$

Now, we are ready to prove Theorem 1.

B.1.1 Proof of Theorem 1

Proof First note that $F_{\mathcal{W},\alpha}$ is μ-strongly convex and L_α-smooth where $L_\alpha = \frac{1-\lambda_N^W}{\alpha} + L$.

Next, note that, as it is shown in Lessard et al. (2016), for every $\alpha \in (0, 2/L_\alpha)$, which is equivalent to $\alpha \in (0, (1+\lambda_N^W)/L)$, there exists $p > 0$ such that the following matrix inequality holds with $\rho(\alpha) = \max\{|1-\alpha\mu|, |1-\alpha L_\alpha|\} = \max\{|1-\alpha\mu|, |\lambda_N^W - \alpha L|\}$:

$$\begin{bmatrix} 2\mu L_\alpha I_d & -(\mu+L_\alpha) I_d \\ -(\mu+L_\alpha) I_d & 2 I_d \end{bmatrix} \succeq \begin{bmatrix} A^\top P A - \rho(\alpha)^2 P & A^\top P B \\ B^\top P A & B^\top P B \end{bmatrix}.$$

As a consequence, and by using Lemma 22, we have

$$\begin{aligned} &\begin{bmatrix} \xi_k \\ \nabla F\left(x^{(k)}\right) \end{bmatrix}^\top \begin{bmatrix} 2\mu L_\alpha I_d & -(\mu+L_\alpha) I_d \\ -(\mu+L_\alpha) I_d & 2 I_d \end{bmatrix} \begin{bmatrix} \xi_k \\ \nabla F\left(x^{(k)}\right) \end{bmatrix} \\ &\geq \begin{bmatrix} \xi_k \\ \nabla F\left(x^{(k)}\right) \end{bmatrix}^\top \begin{bmatrix} A^\top P A - \rho(\alpha)^2 P & A^\top P B \\ B^\top P A & B^\top P B \end{bmatrix} \begin{bmatrix} \xi_k \\ \nabla F\left(x^{(k)}\right) \end{bmatrix} \\ &\geq \mathbb{E}[V_P(\xi_{k+1})] - \rho(\alpha)^2 \mathbb{E}[V_P(\xi_k)] - \sigma^2 \alpha^2 p. \end{aligned} \tag{46}$$

Finally, note that, by using Theorem 2.1.12 in Nesterov (2003), we obtain

$$\begin{bmatrix} \xi_k \\ \nabla F\left(x^{(k)}\right) \end{bmatrix}^\top \begin{bmatrix} 2\mu L_\alpha I_d & -(\mu+L_\alpha) I_d \\ -(\mu+L_\alpha) I_d & 2 I_d \end{bmatrix} \begin{bmatrix} \xi_k \\ \nabla F\left(x^{(k)}\right) \end{bmatrix} \leq 0.$$

Plugging this in (46) and dividing both sides by p, implies

$$\rho(\alpha)^2 \mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right] + N\sigma^2\alpha^2 \geq \mathbb{E}\left[\left\|x^{(k+1)} - x^\infty\right\|^2\right]. \tag{47}$$

Finally, by iterating over k, we obtain

$$\mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right] \leq \rho(\alpha)^{2k} \left\|x^{(0)} - x^\infty\right\|^2 + \alpha^2\sigma^2 N \frac{1-\rho(\alpha)^{2k}}{1-\rho(\alpha)^2}.$$

We also achieve the bound on robustness using the definition of $J_\infty(\alpha)$:

$$J_\infty(\alpha) = \frac{1}{\sigma^2 N} \limsup_{k\to\infty} \operatorname{Var}\left(x^{(k)} - x^\infty\right) \leq \frac{1}{\sigma^2 N} \limsup_{k\to\infty} \mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right].$$

The proof is complete. ■

B.1.2 Proof of Corollary 2

Proof Note that we have $\left\|x^{(k)} - x^*\right\|^2 \leq 2\left\|x^{(k)} - x^\infty\right\|^2 + 2\left\|x^\infty - x^*\right\|^2$, and, for the case that $\alpha < \frac{1}{L+\mu}$, we also have $\|x^\infty - x^*\| \leq \frac{\alpha C_1 \sqrt{N}}{(1-\gamma)}$ from (8), which yields that

$$\mathbb{E}\left[\left\|x^{(k)} - x^*\right\|^2\right] \leq 2\rho^{2k}\left\|x^{(0)} - x^\infty\right\|^2 + 2\alpha^2\sigma^2 N \frac{1-\rho^{2k}}{1-\rho^2} + 2\frac{\alpha^2 C_1^2 N}{(1-\gamma)^2}.$$

■

B.1.3 Proof of Proposition 3

We recall that the average at k-th iteration is given by $\bar{x}^{(k)} := \frac{1}{N}\sum_{i=1}^{N} x_i^{(k)}$. Since $\mathcal{W}$ is doubly stochastic, we get

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \alpha \frac{1}{N}\sum_{i=1}^{N} \nabla f_i\left(x_i^{(k)}\right) - \alpha \bar{\xi}^{(k+1)}, \tag{48}$$

where

$$\bar{\xi}^{(k+1)} := \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{\nabla} f_i\left(x_i^{(k)}\right) - \nabla f_i\left(x_i^{(k)}\right)\right),$$

satisfies

$$\mathbb{E}\left[\bar{\xi}^{(k+1)}\middle|\mathcal{F}_k\right] = 0, \qquad \mathbb{E}\left\|\bar{\xi}^{(k+1)}\right\|^2 \le \frac{\sigma^2}{N}. \tag{49}$$

We can deduce from (48) that

$$\bar{x}^{(k+1)} = \bar{x}^{(k)} - \alpha \nabla f\left(\bar{x}^{(k)}\right) + \alpha \mathcal{E}_{k+1} - \alpha \bar{\xi}^{(k+1)},$$

where

$$\mathcal{E}_{k+1} := \nabla f\left(\bar{x}^{(k)}\right) - \frac{1}{N}\sum_{i=1}^{N} \nabla f_i\left(x_i^{(k)}\right).$$

First, we will show that the error term $\mathcal{E}_{k+1}$ is small. The following result essentially follows from Lemma 7 in Gürbüzbalaban et al. (2021) and hence the proof is omitted here.

Lemma 23 (Lemma 7 in Gürbüzbalaban et al. (2021)) *Assume that $\alpha < \frac{1+\lambda_N^W}{L}$ and $\mu\alpha(1+\lambda_N^W - \alpha L) < 1$. For any k, we have*

$$\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 \le \frac{4L^2\gamma^{2k}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{4L^2D^2\alpha^2}{N(1-\gamma)^2} + \frac{4L^2\sigma^2\alpha^2}{(1-\gamma^2)},$$

where D^2 is defined in (27).

Let us define x_k as the iterates of the centralized algorithm:

$$x_{k+1} = x_k - \alpha \nabla f\left(x_k\right) - \alpha \bar{\xi}^{(k+1)},$$

with $x_0 = \bar{x}^{(0)}$. In the next lemma, we will show that the average of iterates $\bar{x}^{(k)}$ and the iterates of the centralized algorithm x_k are close to each other.

Lemma 24 *Assume that $\alpha < \frac{1+\lambda_N^W}{L}$ and $\mu\alpha(1+\lambda_N^W - \alpha L) < 1$. For any k, we have*

$$\begin{aligned}\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 \le \alpha\left(\frac{\alpha}{\mu(1-\frac{\alpha L}{2})} + \frac{(1+\alpha L)^2}{\mu^2(1-\frac{\alpha L}{2})^2}\right)\left(\frac{4L^2D^2\alpha}{N(1-\gamma)^2} + \frac{4L^2\sigma^2\alpha}{(1-\gamma^2)}\right)\\ + \frac{\gamma^{2k} - \left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)^k}{\gamma^2 - 1 + \alpha\mu\left(1-\frac{\alpha L}{2}\right)}\frac{4L^2\gamma^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2.\end{aligned} \tag{50}$$

Proof The proof of Lemma 24 will be provided in Appendix F. ■

Next, we quote the following result from Gürbüzbalaban et al. (2021) which provides an bound on the distance between $x_i^{(k)}$ and $\bar{x}^{(k)}$ for every $1 \leq i \leq N$.

Lemma 25 (Lemma 6 in Gürbüzbalaban et al. (2021)) *In the setting of Lemma 23, for any k and $i = 1, \ldots, N$, we have*

$$\mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 \leq \sum_{i=1}^{N} \mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 \leq 4\gamma^{2k}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{4D^2\alpha^2}{(1-\gamma)^2} + \frac{4\sigma^2 N\alpha^2}{(1-\gamma^2)},$$

where D is defined in (27).

Completing the proof of Proposition 3. Finally, we are ready to complete the proof of Proposition 3. First, we notice that

$$\mathbb{E}\left\|\bar{x}^{(k)} - x_*\right\|^2 \leq 2\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 + 2\mathbb{E}\left\|x_k - x_*\right\|^2,$$

and for every $i = 1, 2, \ldots, N$,

$$\begin{aligned}\mathbb{E}\left\|x_i^{(k)} - x_*\right\|^2 &\leq 2\mathbb{E}\left\|\bar{x}^{(k)} - x_*\right\|^2 + 2\mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 \\ &\leq 4\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 + 4\mathbb{E}\left\|x_k - x_*\right\|^2 + 2\mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2.\end{aligned}$$

By Proposition 4.3. in Aybat et al. (2020), we have for any $\alpha \leq \frac{2}{L+\mu}$,

$$\begin{aligned}\mathbb{E}\|x_k - x_*\|^2 &\leq (1-\alpha\mu)^{2k}\|x_0 - x_*\|^2 + \frac{1-(1-\alpha\mu)^{2k}}{1-(1-\alpha\mu)^2}\frac{\alpha^2\sigma^2}{N} \\ &= (1-\alpha\mu)^{2k}\|x_0 - x_*\|^2 + \frac{1}{2}\frac{1-(1-\alpha\mu)^{2k}}{\mu(1-\alpha\mu/2)}\frac{\alpha\sigma^2}{N}. \qquad (51)\end{aligned}$$

By (50), we have an upper bound for $\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2$, and by Lemma 25, we have an upper bound for $\mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2$, which completes the proof.

B.2 Proofs of Main Results in Section 2.3.3

Before we proceed to the proof of Theorem 5, let us state the following result from Aybat et al. (2019).

Lemma 26 (Lemma 2.2, Aybat et al. (2019)) *Consider to ASG iterates to minimize the function $F_{\mathcal{W},\alpha}$ in* (15). *Assume there exist $\rho \in (0,1)$ and a positive semi-definite 2×2 matrix $\tilde{P}$ such that*

$$\rho^2\tilde{X}_1 + (1-\rho^2)\tilde{X}_2 \succeq \begin{bmatrix} \tilde{A}_{dasg}^T\tilde{P}\tilde{A}_{dasg} - \rho^2\tilde{P} & \tilde{A}_{dasg}^T\tilde{P}\tilde{B}_{dasg} \\ \tilde{B}_{dasg}^T\tilde{P}\tilde{A}_{dasg} & \tilde{B}_{dasg}^T\tilde{P}\tilde{B}_{dasg} \end{bmatrix},$$

where

$$\tilde{X}_1 := \begin{bmatrix} \frac{\beta^2\mu}{2} & \frac{-\beta^2\mu}{2} & \frac{-\beta}{2} \\ \frac{-\beta^2\mu}{2} & \frac{\beta^2\mu}{2} & \frac{\beta}{2} \\ \frac{-\beta}{2} & \frac{\beta}{2} & \frac{\alpha(2-L_\alpha\alpha)}{2} \end{bmatrix}, \quad \tilde{X}_2 := \begin{bmatrix} \frac{(1+\beta)^2\mu}{2} & \frac{-\beta(1+\beta)\mu}{2} & \frac{-(1+\beta)}{2} \\ \frac{-\beta(1+\beta)\mu}{2} & \frac{\beta^2\mu}{2} & \frac{\beta}{2} \\ \frac{-(1+\beta)}{2} & \frac{\beta}{2} & \frac{\alpha(2-L_\alpha\alpha)}{2} \end{bmatrix}.$$

Let $P = \tilde{P} \otimes I_{Nd}$. Then, for every $k \geq 0$,

$$\mathbb{E}[V_{P,\alpha,1}(\xi_k)] \leq \rho^2 \mathbb{E}[V_{P,\alpha,1}(\xi_{k-1})] + \alpha^2\sigma^2 N \left(\tilde{P}_{11} + \frac{L_\alpha}{2}\right).$$

Now, we are ready to prove Theorem 5.

B.2.1 Proof of Theorem 5

Proof First recall that $F_{\mathcal{W},\alpha}$ is μ-strongly convex and L_α-smooth where $L_\alpha = \frac{1-\lambda_N^W}{\alpha} + L$. Next, Lemma 26 implies that

$$\mathbb{E}\left[V_{P,\alpha,1}(\xi_k)\right] \leq \rho^{2k} V_{P,\alpha,1}(\xi_0) + \frac{1}{1-\rho^2}\alpha^2\sigma^2 N \left(\tilde{P}_{11} + \frac{L_\alpha}{2}\right).$$

By the μ-strong convexity of $F_{\mathcal{W},\alpha}$ and the fact that $\nabla F_{\mathcal{W},\alpha}(x^\infty) = 0$, we have

$$\left\|x^{(k)} - x^\infty\right\|^2 \leq \frac{2}{\mu}\left[F_{\mathcal{W},\alpha}\left(x^{(k)}\right) - F_{\mathcal{W},\alpha}\left(x^\infty\right)\right] \leq 2\frac{V_{P,\alpha,1}(\xi_k)}{\mu}.$$

Also, by the definition of $J_\infty(\alpha)$,

$$J_\infty(\alpha) = \frac{1}{\sigma^2 N}\limsup_{k\to\infty} \mathrm{Var}\left(x^{(k)} - x^\infty\right) \leq \frac{1}{\sigma^2 N}\limsup_{k\to\infty} \mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right].$$

The proof is complete. ■

B.2.2 Proof of Corollary 6

Proof If $\alpha \leq \frac{1}{L+\mu}$, then we have

$$\left\|x^{(k)} - x^*\right\|^2 \leq 2\left\|x^{(k)} - x^\infty\right\|^2 + 2\left\|x^\infty - x^*\right\|^2,$$

and $\|x^\infty - x^*\| \leq \frac{\alpha C_1\sqrt{N}}{(1-\gamma)}$ from (8). Also, by the proof of Lemma 21,

$$\left\|\bar{x}^{(k)} - x_*\right\|^2 \leq \frac{1}{N}\left\|x^{(k)} - x^*\right\|^2.$$

The proof is complete. ■

B.2.3 Proof of Theorem 7

Proof D-ASG reduces to the iterations (18) which are equivalent to applying non-distributed ASG to minimize the function $F_{\mathcal{W},\alpha} \in S_{\mu,L_\alpha}(\mathbb{R}^{Nd})$. Therefore, applying (Aybat et al., 2020, Proposition 4.6) and (Aybat et al., 2020, Corollary 4.9) from the literature for non-distributed ASG, we obtain

$$\mathbb{E}\left[V_{S,\alpha}\left(\xi_{k+1}\right)\right] \leq \left(1-\sqrt{\alpha\mu}\right)\mathbb{E}V_{S,\alpha}\left(\xi_k\right) + \frac{\sigma^2 N\alpha}{2}\left(1+\alpha L_\alpha\right), \tag{52}$$

which yields

$$\mathbb{E}\left[V_{S,\alpha}\left(\xi_k\right)\right] \leq \left(1-\sqrt{\alpha\mu}\right)^k V_{S,\alpha}\left(\xi_0\right) + \frac{\sigma^2 N\alpha}{2\sqrt{\alpha\mu}}\left(1+\alpha L_\alpha\right),$$

provided that $\alpha \in (0, \frac{1}{L_\alpha}]$ where $L_\alpha = \frac{1-\lambda_N^W}{\alpha} + L$ is the smoothness constant of $F_{\mathcal{W},\alpha}$. It can be checked that if $\alpha \in (0, \frac{\lambda_N^W}{L}]$ then, the condition $\alpha \in (0, \frac{1}{L_\alpha}]$ is satisfied. Plugging the value of L_α into (52) proves

$$\mathbb{E}\left[V_{S,\alpha}\left(\xi_k\right)\right] \leq \left(1-\sqrt{\alpha\mu}\right)^k V_{S,\alpha}\left(\xi_0\right) + \frac{\sigma^2 N\sqrt{\alpha}}{2\sqrt{\mu}}\left(2-\lambda_N^W+\alpha L\right), \tag{53}$$

for any $k \geq 0$.

By the μ-strong convexity of $F_{\mathcal{W},\alpha}$ and the fact that $\nabla F_{\mathcal{W},\alpha}(x^\infty) = 0$, we have

$$\left\|x^{(k)} - x^\infty\right\|^2 \leq \frac{2}{\mu}\left[F_{\mathcal{W},\alpha}\left(x^{(k)}\right) - F_{\mathcal{W},\alpha}\left(x^\infty\right)\right].$$

Therefore, (53) implies (31). Finally, by the definition of $J_\infty(\alpha)$,

$$J_\infty(\alpha) = \frac{1}{\sigma^2 N}\limsup_{k\to\infty}\operatorname{Var}\left(x^{(k)} - x^\infty\right) \leq \frac{1}{\sigma^2 N}\limsup_{k\to\infty}\mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right].$$

The proof is complete. ■

B.2.4 Proof of Corollary 8

Proof If $\alpha \leq \frac{1}{L+\mu}$, then we obtain (32) by applying (8) and moreover, we obtain (31) by applying Lemma 21. ■

B.2.5 Proof of Proposition 10

We recall that the averages at k-th iteration are given by $\bar{x}^{(k)} := \frac{1}{N}\sum_{i=1}^N x_i^{(k)}$ and $\bar{y}^{(k)} := \frac{1}{N}\sum_{i=1}^N y_i^{(k)}$. Since $\mathcal{W}$ is doubly stochastic, we get

$$\begin{aligned}
\bar{x}^{(k+1)} &= \bar{y}^{(k)} - \alpha\frac{1}{N}\sum_{i=1}^N \nabla f_i\left(y_i^{(k)}\right) - \alpha\bar{\xi}^{(k+1)}, \\
\bar{y}^{(k)} &= (1+\beta)\bar{x}^{(k)} - \beta\bar{x}^{(k-1)},
\end{aligned} \tag{54}$$

where

$$\bar{\xi}^{(k+1)} := \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{\nabla} f_i\left(y_i^{(k)}\right) - \nabla f_i\left(y_i^{(k)}\right)\right),$$

satisfies

$$\mathbb{E}\left[\bar{\xi}^{(k+1)}\middle|\mathcal{F}_k\right] = 0, \qquad \mathbb{E}\left\|\bar{\xi}^{(k+1)}\right\|^2 \leq \frac{\sigma^2}{N}. \tag{55}$$

We will establish several lemmas for completing the proof of Proposition 10. We start with stating and proving the following lemma which provides a bound on the L_2 distance between the local variables $x_i^{(k)}$ and the node averages $\bar{x}^{(k)}$.

Lemma 27 *Assume the conditions in Proposition 10 hold. For any k, we have*

$$\sum_{i=1}^{N}\mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 \leq 8\gamma^{2k}\left(4\frac{V_{S,\alpha}\left(\xi_0\right)}{\mu} + \frac{2\sigma^2 N\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(2-\lambda_N^W+\alpha L\right) + \frac{2C_1^2 N\alpha^2}{(1-\gamma)^2} + \|x^*\|^2\right) + \frac{4D_y^2\alpha^2}{(1-\gamma)^2} + \frac{4\sigma^2 N\alpha^2}{(1-\gamma)^2} + \frac{8C_0\alpha}{(1-\gamma)^2},$$

where C_0 is defined in Lemma 29 and D_y is defined in (34).

Proof The proof of Lemma 27 will be provided in Appendix F. ■

We can deduce from (54) that

$$\begin{aligned}\bar{x}^{(k+1)} &= \bar{x}^{(k)} - \alpha\nabla f\left(\bar{y}^{(k)}\right) + \alpha\mathcal{E}_{k+1} - \alpha\bar{\xi}^{(k+1)},\\ \bar{y}^{(k)} &= (1+\beta)\bar{x}^{(k)} - \beta\bar{x}^{(k-1)},\end{aligned}$$

where

$$\mathcal{E}_{k+1} := \nabla f\left(\bar{y}^{(k)}\right) - \frac{1}{N}\sum_{i=1}^{N}\nabla f_i\left(y_i^{(k)}\right). \tag{56}$$

Notice that we are abusing the notation here; and $\mathcal{E}_{k+1}$ is used to denote the error term for D-SG as well. Next, we will show that the error term $\mathcal{E}_{k+1}$ is small.

Lemma 28 *Assume the conditions in Proposition 10 hold. For any k, we have*

$$\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 \leq \frac{2}{N}L^2\left((1+\beta)^2+\beta^2\right)\left[4D_y^2\alpha^2\frac{1}{(1-\gamma)^2} + \frac{4\sigma^2 N\alpha^2}{(1-\gamma)^2} + \frac{8C_0}{(1-\gamma)^2}\alpha + 8\gamma^{2(k-1)}\left(4\frac{V_{S,\alpha}\left(\xi_0\right)}{\mu} + \frac{2\sigma^2 N\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(2-\lambda_N^W+\alpha L\right) + \frac{2C_1^2 N\alpha^2}{(1-\gamma)^2} + \|x^*\|^2\right)\right],$$

where C_0 is defined in Lemma 29 and $\mathcal{E}_k$ is defined in (56).

Proof The proof of Lemma 28 will be provided in Appendix F. ■

We recall from (33) that

$$V_{\bar{S},\alpha}(\bar{\xi}) = \bar{\xi}^T \bar{S}_\alpha \bar{\xi} + f\left(\bar{T}\bar{\xi} + x_*\right) - f(x_*), \tag{57}$$

where $\bar{S}_\alpha = \tilde{S}_\alpha \otimes I_d$. We can represent the average of the D-ASG iterations as the dynamical system

$$\bar{\xi}_{k+1} = A\bar{\xi}_k + B\tilde{\nabla} f\left(C\bar{\xi}_k\right) + D_{k+1}, \tag{58}$$

where ξ_k is the state, and A, B, C are system matrices that are appropriately chosen such that

$$\xi_k := \left[\left(\bar{x}^{(k)} - x_*\right)^T, \left(\bar{x}^{(k-1)} - x_*\right)^T\right]^T, \tag{59}$$

and $A = \tilde{A}_{\text{dasg}} \otimes I_d, B := \tilde{B}_{\text{dasg}} \otimes I_d, C := \tilde{C}_{\text{dasg}} \otimes I_d$ where

$$\tilde{A}_{\text{dasg}} := \begin{bmatrix} 1+\beta & -\beta \\ 1 & 0 \end{bmatrix}, \quad \tilde{B}_{\text{dasg}} := \begin{bmatrix} -\alpha \\ 0 \end{bmatrix}, \quad \tilde{C}_{\text{dasg}} := \begin{bmatrix} 1+\beta & -\beta \end{bmatrix}, \tag{60}$$

and

$$D_k := \begin{bmatrix} \alpha\mathcal{E}_k \\ 0 \end{bmatrix}, \tag{61}$$

where $\mathcal{E}_k$ is defined in (56). We will make use of the Lyapunov function (57) to establish the convergence of the averaged iterates in the remaining part of the proof. In the next result, we will obtain a helper lemma that shows that the difference between the consecutive iterates $x^{(k)} - x^{(k-1)}$ is bounded in L_2 with a bound that is proportional to the stepsize α.

Lemma 29 *Consider running D-ASG method with $\alpha \in \left(0, \frac{\lambda_N^W}{L}\right]$, $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$ and initialization $x^{(0)} = x^{(-1)} = 0$. Then, we have*

$$\sup_{k\geq 0} \mathbb{E}\left\|x^{(k)} - x^{(k-1)}\right\|^2 \leq \frac{2C_0}{\beta^2}\alpha, \tag{62}$$

for a positive constant C_0 that can be made explicit. Furthermore, C_0 is such that $C_0 = \mathcal{O}(1)$ as $\alpha \to 0$.

Proof The proof of Lemma 29 will be provided in Appendix F. ■

Lemma 30 *Assume the conditions in Proposition 10 hold. For any $\epsilon > 0$, there exists $C_\epsilon > 0$ such that*

$$\mathbb{E}\left[V_{\bar{S},\alpha}\left(\bar{\xi}_{k+1}\right)\right] \leq (1+\epsilon)\mathbb{E}\left[V_{\bar{S},\alpha}\left(\bar{\xi}_{k+1} - D_{k+1}\right)\right] + C_\epsilon \mathbb{E}\|D_{k+1}\|^2, \tag{63}$$

where

$$C_\epsilon := \frac{1}{2\epsilon}\max\left(4\left(\frac{1}{\alpha} + \frac{\mu}{2} - \frac{\sqrt{\mu}}{\sqrt{\alpha}}\right), \frac{L^2}{\mu}\right) + \frac{L}{2} + \frac{1}{\alpha} + \frac{\mu}{2} - \frac{\sqrt{\mu}}{\sqrt{\alpha}}, \tag{64}$$

$\bar{\xi}_{k+1}$ is defined by (58) *and D_{k+1} is defined by* (61).

Proof The proof of Lemma 30 will be provided in Appendix F. ■

Completing the proof of Proposition 10. D-ASG reduces to the iterations which are equivalent to applying non-distributed ASG to minimize the function $f \in S_{\mu,L_\alpha}(\mathbb{R}^d)$. Therefore, applying (Aybat et al., 2020, Proposition 4.6) and (Aybat et al., 2020, Corollary 4.9) from the literature for non-distributed ASG, we obtain

$$\mathbb{E}\left[V_{\bar{S},\alpha}\left(\bar{\xi}_{k+1}-D_{k+1}\right)\right] \leq (1-\sqrt{\alpha\mu})\,\mathbb{E}V_{\bar{S},\alpha}\left(\bar{\xi}_k\right)+\frac{\sigma^2\alpha}{2N}\left(1+\alpha L\right), \tag{65}$$

which yields

$$\begin{aligned}\mathbb{E}\left[V_{\bar{S},\alpha}\left(\bar{\xi}_{k+1}\right)\right] &\leq (1+\epsilon)\mathbb{E}\left[V_{\bar{S},\alpha}\left(\bar{\xi}_{k+1}-D_{k+1}\right)\right]+C_\epsilon\mathbb{E}\|D_{k+1}\|^2\\ &\leq (1+\epsilon)\left((1-\sqrt{\alpha\mu})\,\mathbb{E}V_{\bar{S},\alpha}\left(\bar{\xi}_k\right)+\frac{\sigma^2\alpha}{2N}\left(1+\alpha L\right)\right)+C_\epsilon\mathbb{E}\|D_{k+1}\|^2.\end{aligned}$$

Let us take $\epsilon=\frac{1}{2}\sqrt{\alpha\mu}$, then we get

$$\begin{aligned}\mathbb{E}\left[V_{\bar{S},\alpha}\left(\bar{\xi}_{k+1}\right)\right] &\leq \left(1+\frac{\sqrt{\alpha\mu}}{2}\right)\left((1-\sqrt{\alpha\mu})\,\mathbb{E}V_{\bar{S},\alpha}\left(\bar{\xi}_k\right)+\frac{\sigma^2\alpha}{2N}\left(1+\alpha L\right)\right)+C_{\frac{\sqrt{\alpha\mu}}{2}}\mathbb{E}\|D_{k+1}\|^2\\ &\leq \left(1-\frac{\sqrt{\alpha\mu}}{2}\right)\mathbb{E}V_{\bar{S},\alpha}\left(\bar{\xi}_k\right)+\left(1+\frac{\sqrt{\alpha\mu}}{2}\right)\frac{\sigma^2\alpha}{2N}\left(1+\alpha L\right)+C_{\frac{\sqrt{\alpha\mu}}{2}}\mathbb{E}\|D_{k+1}\|^2,\end{aligned}$$

and by Lemma 30, we have

$$\begin{aligned}C_{\frac{\sqrt{\alpha\mu}}{2}} &= \frac{1}{\sqrt{\alpha\mu}}\max\left(4\left(\frac{1}{\alpha}+\frac{\mu}{2}-\frac{\sqrt{\mu}}{\sqrt{\alpha}}\right),\frac{L^2}{\mu}\right)+\frac{L}{2}+\frac{1}{\alpha}+\frac{\mu}{2}-\frac{\sqrt{\mu}}{\sqrt{\alpha}}\\ &\leq \frac{1}{\sqrt{\alpha\mu}}\left(4\left(\frac{1}{\alpha}+\frac{\mu}{2}-\frac{\sqrt{\mu}}{\sqrt{\alpha}}\right)+\frac{L^2}{\mu}\right)+\frac{L}{2}+\frac{1}{\alpha}+\frac{\mu}{2}-\frac{\sqrt{\mu}}{\sqrt{\alpha}}\\ &= \frac{1}{2\alpha\sqrt{\alpha\mu}}H_1,\end{aligned}$$

where

$$H_1 = 8\left(1+\frac{\mu\alpha}{2}-\sqrt{\alpha\mu}\right)+\frac{2L^2\alpha}{\mu}+\left(L\alpha+2+\mu\alpha\right)\sqrt{\alpha\mu}-2\mu\alpha. \tag{66}$$

By Lemma 28, we have

$$\mathbb{E}\|D_{k+1}\|^2 \leq \alpha^2\left[\alpha H_2+8\gamma^{2(k-1)}H_3\right],$$

where D_k is defined by (61),

$$\begin{aligned}H_2 &= \frac{2}{N}L^2\left((1+\beta)^2+\beta^2\right)\left(4D_y^2\alpha\frac{1}{(1-\gamma)^2}+\frac{4\sigma^2N\alpha}{(1-\gamma)^2}+\frac{8C_0}{(1-\gamma)^2}\right),\\ H_3 &= \frac{2}{N}L^2\left((1+\beta)^2+\beta^2\right)\left(4\frac{V_{S,\alpha}\left(\xi_0\right)}{\mu}+\frac{2\sigma^2N\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(2-\lambda_N^W+\alpha L\right)+\frac{2C_1^2N\alpha^2}{(1-\gamma)^2}+\|x^*\|^2\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[V_{\bar{S},\alpha}\left(\bar{\xi}_{k+1}\right)\right] &\leq \left(1-\frac{\sqrt{\alpha\mu}}{2}\right)\mathbb{E}V_{\bar{S},\alpha}\left(\bar{\xi}_k\right)+\left(1+\frac{\sqrt{\alpha\mu}}{2}\right)\frac{\sigma^2\alpha}{2N}\left(1+\alpha L\right)\\ &\quad+\frac{1}{2\alpha\sqrt{\alpha\mu}}H_1\alpha^2\left[\alpha H_2+8\gamma^{2(k-1)}H_3\right]\\ &=\left(1-\frac{\sqrt{\alpha\mu}}{2}\right)\mathbb{E}V_{\bar{S},\alpha}\left(\bar{\xi}_k\right)+\left(1+\frac{\sqrt{\alpha\mu}}{2}\right)\frac{\sigma^2\alpha}{2N}\left(1+\alpha L\right)\\ &\quad+\frac{1}{2\sqrt{\mu}}\alpha\sqrt{\alpha}H_1H_2+\frac{4}{\gamma^2\sqrt{\mu}}\sqrt{\alpha}H_1H_3\gamma^{2k},\end{aligned}$$

which implies that

$$\begin{aligned}\mathbb{E}\left[V_{\bar{S},\alpha}\left(\bar{\xi}_k\right)\right] &\leq \left(1-\frac{\sqrt{\alpha\mu}}{2}\right)^k V_{\bar{S},\alpha}\left(\bar{\xi}_0\right)+\frac{4}{\gamma^2\sqrt{\mu}}\sqrt{\alpha}H_1H_3\sum_{i=0}^{k-1}\left(1-\frac{\sqrt{\alpha\mu}}{2}\right)^i\left(\gamma^2\right)^{k-1-i}\\ &\quad+\sum_{i=0}^{k-1}\left(1-\frac{\sqrt{\alpha\mu}}{2}\right)^i\left(\left(1+\frac{\sqrt{\alpha\mu}}{2}\right)\frac{\sigma^2\alpha}{2N}\left(1+\alpha L\right)+\frac{1}{2\sqrt{\mu}}\alpha\sqrt{\alpha}H_1H_2\right)\\ &\leq\left(1-\frac{\sqrt{\alpha\mu}}{2}\right)^k V_{\bar{S},\alpha}\left(\bar{\xi}_0\right)+\frac{4}{\gamma^2\sqrt{\mu}}\sqrt{\alpha}H_1H_3\frac{\gamma^{2k}-(1-\sqrt{\alpha\mu}/2)^k}{\gamma^2-(1-\sqrt{\alpha\mu}/2)}\\ &\quad+\frac{2}{\sqrt{\alpha\mu}}\left(\left(1+\frac{\sqrt{\alpha\mu}}{2}\right)\frac{\sigma^2\alpha}{2N}\left(1+\alpha L\right)+\frac{1}{2\sqrt{\mu}}\alpha\sqrt{\alpha}H_1H_2\right).\end{aligned}$$

By the μ-strong convexity of f and the fact that $\nabla f(x_*)=0$, we have

$$\left\|\bar{x}^{(k)}-x_*\right\|^2\leq\frac{2}{\mu}\left[f\left(\bar{x}^{(k)}\right)-f\left(x_*\right)\right]\leq\frac{2}{\mu}V_{\bar{S},\alpha}\left(\bar{\xi}_k\right),$$

which implies that

$$\begin{aligned}&\mathbb{E}\left\|\bar{x}^{(k)}-x_*\right\|^2\\ &\leq\left(1-\frac{\sqrt{\alpha\mu}}{2}\right)^k\frac{2V_{\bar{S},\alpha}\left(\bar{\xi}_0\right)}{\mu}+\frac{4}{\mu\sqrt{\mu}}\left(\left(1+\frac{\sqrt{\alpha\mu}}{2}\right)\frac{\sigma^2\sqrt{\alpha}}{2N}\left(1+\alpha L\right)+\frac{1}{2\sqrt{\mu}}\alpha H_1H_2\right)\\ &\quad+\frac{8}{\gamma^2\mu\sqrt{\mu}}\sqrt{\alpha}H_1H_3\frac{\gamma^{2k}-(1-\sqrt{\alpha\mu}/2)^k}{\gamma^2-(1-\sqrt{\alpha\mu}/2)}.\end{aligned}$$

Finally, for every $1\leq i\leq N$,

$$\mathbb{E}\left\|x_i^{(k)}-x_*\right\|^2\leq 2\mathbb{E}\left\|x_i^{(k)}-\bar{x}^{(k)}\right\|^2+2\mathbb{E}\left\|\bar{x}^{(k)}-x_*\right\|^2,$$

and by applying Lemma 27, the proof is complete.

Appendix C. Quadratic Objectives

In this section, we analyze the special case when f_i is quadratic at every node i under the same Assumption 1 with the main text. We assume

$$f_i(x) = \frac{1}{2}x^T Q_i x - p_i^T x + r_i,$$

where Q_i is an $d \times d$ symmetric positive definite matrix, $p_i \in \mathbb{R}^d$ and $r_i \in \mathbb{R}$ for $i = 1, 2, \ldots, N$. In this special case, the optimum to the (1) is explicitly given by

$$x^* = \left(\sum_{i=1}^N Q_i\right)^{-1} \sum_{i=1}^N p_i.$$

Furthermore, the function F defined as $F(x) := F(x_1, \ldots, x_N) := \frac{1}{N}\sum_{i=1}^N f_i(x_i)$ is also a quadratic function of the form

$$F(x) = \frac{1}{2}x^T Q x - p^T x + r, \tag{67}$$

where $Q = \mathbf{diag}(\{Q_i\}_{i=1}^N)$ is an $Nd \times Nd$ symmetric positive definite matrix:

$$Q := \begin{bmatrix} Q_1 & 0_d & \ldots & 0_d \\ 0_d & Q_2 & \ddots & 0_d \\ \vdots & \ddots & \ddots & 0_d \\ 0_d & \ldots & 0_d & Q_N \end{bmatrix}, \tag{68}$$

and

$$p := \begin{bmatrix} p_1^T & p_2^T \ldots & p_N^T \end{bmatrix}^T \in \mathbb{R}^{Nd} \tag{69}$$

is a column vector and $r = \sum_{i=1}^N r^i \in \mathbb{R}$ is a scalar. Moreover, the gradient of F is given by $\nabla F(x) = Qx - p$.

Throughout this section, and to simplify the derivations for quadratic functions, we focus on the case of additive noise. More formally, we consider the following noise assumption for this section:

Assumption 3 *At iteration k, node i has access to $\tilde{\nabla} f_i\left(x_i^{(k)}, w_i^{(k)}\right)$ which is an estimate of $\nabla f_i\left(x_i^{(k)}\right)$ and satisfies the conditions given in Assumption 1. In addition, we assume this randomness is in the form of additive noise, i.e., $\tilde{\nabla} f_i\left(x_i^{(k)}, w_i^{(k)}\right) = \nabla f_i\left(x_i^{(k)}\right) + w_i^{(k)}$.*

Also, similar to (3), we define the vector $w^{(k)}$ as:

$$w^{(k)} = \left[\left(w_1^{(k)}\right)^T, \left(w_2^{(k)}\right)^T, \ldots, \left(w_N^{(k)}\right)^T\right]^T \in \mathbb{R}^{Nd}. \tag{70}$$

C.1 Distributed Stochastic Gradient (D-SG)

The network-wide D-SG update (19) reduces to a linear recursion

$$x^{(k+1)} = (W \otimes I_d)\, x^{(k)} - \alpha \left[Q x^{(k)} + p \right] - \alpha w^{(k+1)},$$

where Q and p are defined in (68) and (69). Then, the network-wide update (19) reduces to

$$\xi_{k+1} = A_Q \xi_k - \alpha w^{(k+1)},$$

where $\xi_k = x^{(k)} - x^\infty$ and

$$A_Q = \mathcal{W} - \alpha Q.$$

By the assumption that f_i's are μ-strongly convex with L-Lipschitz gradients, we have $\mu I_{Nd} \preceq Q \preceq L I_{Nd}$. Since the stepsize $\alpha > 0$, it is easy to see that

$$\left(\lambda_N^W - \alpha L\right) I_{Nd} \preceq A_Q \preceq (1 - \alpha\mu) I_{Nd}. \tag{71}$$

The next result is on the *spectral radius* of A_Q which is defined as the maximum of the Euclidean norm of the eigenvalues of A_Q.

Proposition 31 *For any stepsize $\alpha > 0$,*

$$\rho(A_Q) = \|\mathcal{W} - \alpha Q\| = \max\left\{ |1 - \alpha\mu|\,, \left|\lambda_N^W - \alpha L\right| \right\}. \tag{72}$$

where ρ denotes the spectral radius of A_Q. In particular, if $\alpha \in \left(0, \frac{1+\lambda_N^W}{L+\mu}\right]$, then

$$\rho(A_Q) = 1 - \alpha\mu \in [0, 1).$$

Proof The equality (72) follows directly from (71). The second part, note that we have $1 - \alpha\mu > \lambda_N^W - \alpha L$ as $\mu \leq L$ and $\lambda_N^W < 1$. Furthermore, for $\alpha > 0$ small enough, it is easy to see from (72) that $\rho(A_Q) = 1 - \alpha\mu = |1 - \alpha\mu|$. The proof follows after checking that $1 - \alpha\mu = |1 - \alpha\mu| \geq |\lambda_N^W - \alpha L|$ for $\alpha \in \left[0, \frac{1+\lambda_N^W}{L+\mu}\right]$. ■

Remark 32 *In the noiseless case (when $\sigma = 0$), we have*

$$\|\xi_k\| = \|A_Q\|^k \|\xi_0\|,$$

provided that $x^{(0)}$ is chosen as an eigenvector corresponding to a largest singular value of the A_Q matrix. Therefore, by Proposition 31, this gives

$$\|\xi_k\| = \rho(\alpha)^k \|\xi_0\|,$$

where $\rho(\alpha)$ is as in Theorem 1. This shows that the analysis of Theorem 1 is tight in the sense that the convergence rate it provides for strongly convex objectives are attained for quadratics for particular choices of the initialization ξ_0 when $\sigma = 0$.

A consequence of Theorem 1 for strongly convex objectives is that for $\rho(\alpha) = \rho(A_Q) < 1$, the robustness measure, or equivalently the variance of the iterates in the limit for the quadratic objectives, satisfies the bound

$$J_\infty(\alpha) = \frac{1}{\sigma^2 N} \lim_{k\to\infty} \mathrm{Var}\left(\xi_k\right) = \frac{1}{\sigma^2 N} \limsup_{k\to\infty} \mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right]$$
$$\leq \frac{1}{1-\rho(\alpha)^2}\alpha^2 = \frac{\alpha^2}{1-\max\left\{|1-\alpha\mu|, |\lambda_N^W - \alpha L|\right\}^2}.$$

If we assume more structure on the noise, we can get tighter bounds. Consider the following assumption which says that the noise has a fixed covariance structure; this assumption is clearly stronger than Assumption 3.

Assumption 4 *The noise $w_i^{(k)}$ are independent, identically distributed (i.i.d.) for every i and k with zero mean and covariance matrix $\Sigma_{w_i} := \mathbb{E}\left[w_i^{(k)}\left(w_i^{(k)}\right)^T\right] = \frac{\sigma^2}{d} I_d$.*

The next theorem shows that we can get a tighter explicit representation of the variance of the iterates in terms of the eigenvalues of the iteration matrix A_Q.

Theorem 33 *Under Assumption 3 and Assumption 4, if $\alpha \in \left(0, \frac{1+\lambda_N^W}{\mu+L}\right]$, the D-SG iterates given by (20) satisfy*

$$\lim_{k\to\infty} Var\left(\xi_k\right) = \alpha^2 \frac{\sigma^2}{d} \sum_{i=1}^{Nd} \frac{1}{1-\mu_i^2}, \tag{73}$$

where μ_i are eigenvalues of $A_Q = \mathcal{W} - \alpha Q$, and hence the robustness measure is given by

$$J_\infty(\alpha) = \alpha^2 \frac{1}{Nd} \sum_{i=1}^{Nd} \frac{1}{1-\mu_i^2}.$$

Proof Note that the matrix A_Q is symmetric with real eigenvalues. Furthermore, by Proposition 31, we have $|\mu_i| \leq \rho(A_Q) < 1$ for every i. Therefore, the quantity on the right-hand side of (73) is well-defined. Define the covariance matrix

$$\Sigma_k = \mathbb{E}\left[\xi_k \xi_k^T\right].$$

We have the recursion

$$\Sigma_{k+1} = A_Q \Sigma_k A_Q^T + \alpha^2 \Sigma_w, \tag{74}$$

where $\Sigma_w = \mathbf{diag}([\Sigma_{w_i}]_{i=1}^N)$ is the covariance matrix of the noise, which is equal to $\frac{\sigma^2}{d} I_{Nd}$ by Assumption 4.

Let $W = VDV^T$ be an eigenvalue decomposition of W. Assume without loss of generality, that diagonal of D contains the eigenvalues in decreasing order, i.e. $D_{ii} = \lambda_i^W$. In this case, j-th column of V, say v_j is an eigenvector corresponding to λ_j^W. Note that the eigenvalues of $\mathcal{W} = W \otimes I_d$ are λ_j^W each with multiplicity d and we can choose the

corresponding eigenvectors as $v_j \otimes e_i$ for $j = 1, 2, \ldots, N$ and $i = 1, 2, \ldots, d$ where e_i is the standard basis. In other words, we can write

$$\mathcal{W} = \mathcal{V}\mathcal{D}\mathcal{V}^T, \quad \text{where} \quad \mathcal{D} = \begin{bmatrix} \lambda_1^W I_d & 0_d & \ldots & 0_d \\ 0_d & \lambda_2^W I_d & \ddots & 0_d \\ \vdots & \ddots & \ddots & \vdots \\ 0_d & \ldots & 0_d & \lambda_N^W I_d \end{bmatrix},$$

for some $\mathcal{V}$. We will write the D-SG iterations (7) with respect to this basis. Let

$$\hat{Q} := \mathcal{V}^T Q \mathcal{V}, \quad \hat{\xi}_k := \mathcal{V}^T \xi_k \mathcal{V}, \quad \hat{\Sigma}_k := \mathcal{V}^T \Sigma_k \mathcal{V}.$$

For $\Sigma_w = (\sigma^2/d) I_{Nd}$, we can write (74) as

$$\hat{\Sigma}_{k+1} = A_{\hat{Q}} \hat{\Sigma}_k A_{\hat{Q}}^T + \alpha^2 (\sigma^2/d) I_{Nd}, \tag{75}$$

where

$$A_{\hat{Q}} := \mathcal{D} - \alpha \hat{Q}.$$

We obtain

$$\lim_{k\to\infty} \hat{\Sigma}_k = \alpha^2(\sigma^2/d) \sum_{k=0}^{\infty} A_{\hat{Q}}^{2k} = \alpha^2(\sigma^2/d) \left(I - A_{\hat{Q}}^2\right)^{-1},$$

where $\hat{\mu}_i$ are the eigenvalues of $A_{\hat{Q}}$. Therefore,

$$\lim_{k\to\infty} \mathrm{Var}\left(\xi_k\right) = \lim_{k\to\infty} \mathrm{trace}(\Sigma_k) = \lim_{k\to\infty} \mathrm{trace}\left(\hat{\Sigma}_k\right) = \alpha^2(\sigma^2/d) \sum_{i=1}^{Nd} \frac{1}{1-\mu_i^2},$$

where μ_i are the eigenvalues of $A_{\hat{Q}}$ or equivalently of A_Q. This completes the proof. ■

Proposition 34 *Assume that Assumption 3 and Assumption 4 hold. For any $j = 1, \ldots, d$ and any $k \in \mathbb{N}$,*

$$\lim_{k\to\infty} Var\left(\bar{x}^{(k)}(j)\right) \leq \frac{\sigma^2}{Nd} \max_{i=1,2,\ldots,Nd} \frac{\alpha^2}{1-\mu_i^2},$$

where $\bar{x}^{(k)}(j)$ denotes the j-th entry of the node average $\bar{x}^{(k)}$ and μ_i are the eigenvalues of $\mathcal{W} - \alpha Q$.

Proof D-SG can be viewed a special case of D-ASG when the momentum parameter $\beta = 0$. The conclusion follows from the more general result for D-ASG in Proposition 39. ■

Next, for D-SG iterates, we provide bounds on $\mathbb{E}\left[\left\|x^{(k)} - x^{\infty}\right\|^2\right]$ and $\mathbb{E}\left[\left\|x^{(k)} - x^{*}\right\|^2\right]$.

Theorem 35 *Consider the D-SG iterates under Assumption 3 and Assumption 4. For every $k \in \mathbb{N}$,*

$$\mathbb{E}\left[\left\|x^{(k)} - x^{\infty}\right\|^2\right] \leq \rho_{dsg}^{2k}\left(\left\|\xi_0\xi_0^T\right\| + \frac{\alpha^2\sigma^2 N}{1-\rho_{dsg}^2}\right) + \alpha^2\frac{\sigma^2}{d}\sum_{i=1}^{Nd}\frac{1}{1-\mu_i^2}, \tag{76}$$

where $\rho_{dsg} := \max_{1\leq i\leq Nd}|\mu_i|$, where μ_i are eigenvalues of A_Q.

In particular, if $\alpha \in \left(0, \frac{1+\lambda_N^W}{L+\mu}\right]$, then

$$\mathbb{E}\left[\left\|x^{(k)} - x^{\infty}\right\|^2\right] \leq (1-\alpha\mu)^{2k}\left(\left\|\xi_0\xi_0^T\right\| + \frac{\alpha^2\sigma^2 N}{1-(1-\alpha\mu)^2}\right) + \alpha^2\frac{\sigma^2}{d}\sum_{i=1}^{Nd}\frac{1}{1-\mu_i^2}.$$

In addition, if we have $\alpha \in \left(0, \frac{1}{L+\mu}\right]$, then

$$\begin{aligned}&\mathbb{E}\left[\left\|x^{(k)} - x^{*}\right\|^2\right]\\ &\leq (1-\alpha\mu)^{2k}\left(2\left\|\xi_0\xi_0^T\right\| + \frac{2\alpha^2\sigma^2 N}{1-(1-\alpha\mu)^2}\right) + \frac{2\alpha^2\sigma^2}{d}\sum_{i=1}^{Nd}\frac{1}{1-\mu_i^2} + \frac{2\alpha^2 C_1^2 N}{(1-\gamma)^2}.\end{aligned} \tag{77}$$

Proof We recall that with $\xi_k = x^{(k)} - x^{\infty}$,

$$\xi_{k+1} = A_Q\xi_k - \alpha w^{(k+1)},$$

and therefore, we get:

$$\mathbb{E}\left[\xi_k\xi_k^T\right] = A_Q\mathbb{E}\left[\xi_{k-1}\xi_{k-1}^T\right](A_Q)^T + \alpha^2\frac{\sigma^2}{d}I_{Nd}, \tag{78}$$

Therefore,

$$X := \mathbb{E}\left[\xi_\infty\xi_\infty^T\right]$$

satisfies the discrete Lyapunov equation:

$$X = A_Q X (A_Q)^T + \alpha^2\frac{\sigma^2}{d}I_{Nd}.$$

By Theorem 33, we have

$$\operatorname{trace}(X) = \alpha^2\frac{\sigma^2}{d}\sum_{i=1}^{Nd}\frac{1}{1-\mu_i^2}.$$

Next by iterating equation (78) over k, we immediately obtain

$$\mathbb{E}\left[\xi_k\xi_k^T\right] = (A_Q)^k\,\xi_0\xi_0^T\left((A_Q)^T\right)^k + \sum_{j=0}^{k-1}(A_Q)^j\,\alpha^2\frac{\sigma^2}{d}I_{Nd}\left((A_Q)^T\right)^j,$$

so that

$$\mathbb{E}\left[\xi_k\xi_k^T\right] = \mathbb{E}\left[\xi_\infty\xi_\infty^T\right] + (A_Q)^k\,\xi_0\xi_0^T\left((A_Q)^T\right)^k - \sum_{j=k}^{\infty}(A_Q)^j\,\alpha^2\frac{\sigma^2}{d}I_{Nd}\left((A_Q)^T\right)^j,$$

which implies that

$$\begin{aligned}\operatorname{trace}\left(\mathbb{E}\left[\xi_k\xi_k^T\right]\right) &= \operatorname{trace}\left(\mathbb{E}\left[\xi_\infty\xi_\infty^T\right]\right) + (A_Q)^k\,\xi_0\xi_0^T\left((A_Q)^T\right)^k \\ &\quad - \sum_{j=k}^{\infty}(A_Q)^j\,\alpha^2\frac{\sigma^2}{d}I_{Nd}\left((A_Q)^T\right)^j \\ &\leq \operatorname{trace}(X) + \left\|(A_Q)^k\right\|^2\left\|\xi_0\xi_0^T\right\| + \sum_{j=k}^{\infty}\left\|(A_Q)^j\right\|^2\alpha^2\sigma^2 N \\ &\leq \operatorname{trace}(X) + \rho_{\text{dsg}}^{2k}\left\|\xi_0\xi_0^T\right\| + \alpha^2\sigma^2 N\frac{\rho_{\text{dsg}}^{2k}}{1-\rho_{\text{dsg}}^2},\end{aligned}$$

where we used the estimate:

$$\left\|A_Q^k\right\| \leq \|V\|^2\left(\max_{1\leq i\leq Nd}|\mu_i|\right)^k = \left(\max_{1\leq i\leq Nd}|\mu_i|\right)^k = \rho_{\text{dsg}}^k,$$

where we used the fact that $A_Q = \mathcal{W} - \alpha Q$ is symmetric with the decomposition $A_Q = V\mathbf{diag}\left([\mu_i]_{i=1}^{Nd}\right)V^T$, where μ_i are the eigenvalues of A_Q and the fact that $\|V\| = 1$ since V is orthogonal. Note that $\xi_k = x^{(k)} - x^\infty$, and this proves (76).

Finally, when $\alpha \in \left(0, \frac{1+\lambda_N^W}{L+\mu}\right]$, by Proposition 31, we get

$$\rho_{\text{dsg}} = \max_{1\leq i\leq Nd}|\mu_i| = 1 - \alpha\mu.$$

Moreover,

$$\left\|x^{(k)} - x^*\right\|^2 \leq 2\left\|x^{(k)} - x^\infty\right\|^2 + 2\left\|x^\infty - x^*\right\|^2,$$

and together with (8), it proves (77). The proof is complete. ■

C.2 Distributed Accelerated Stochastic Gradient (D-ASG)

First, let us recall that the network-wide update for D-ASG is given by

$$x^{(k+1)} = \mathcal{W}y^{(k)} - \alpha\left[\nabla F\left(y^{(k)}\right) + w^{(k+1)}\right], \tag{79}$$

$$y^{(k)} = (1+\beta)x^{(k)} - \beta x^{(k-1)}, \tag{80}$$

where $F : \mathbb{R}^{Nd} \to \mathbb{R}$, is defined as $F(y) := F(y_1, \ldots, y_N) = \sum_{i=1}^N f_i(y_i)$, and the noise $w^{(k+1)}$ satisfies (5).

In the quadratic case, i.e. F is quadratic and defined in (67), we can re-write the D-ASG iterates (79)-(80) as

$$\xi_{k+1} = A_{\text{dasg},Q}\xi_k + B_{\text{dasg}}w^{(k+1)}, \tag{81}$$

where

$$\xi_k := \left[\left(x^{(k)} - x^\infty\right)^T, \left(x^{(k-1)} - x^\infty\right)^T\right]^T,$$

and

$$A_{\text{dasg},Q} := \begin{bmatrix} (1+\beta)(\mathcal{W}-\alpha Q) & -\beta(\mathcal{W}-\alpha Q) \\ I_{Nd} & 0_{Nd} \end{bmatrix}, \tag{82}$$

and B_{dasg} is defined in Section 2.3.1 and Q is given in (68). Next, we obtain the spectral radius of $A_{\text{dasg},Q}$, that is the maximum of the Euclidean norm of the eigenvalues of $A_{\text{dasg},Q}$.

Proposition 36 *Let μ_i, $1 \le i \le Nd$, be the eigenvalues of $\mathcal{W}-\alpha Q$ listed in non-increasing order. We have*

$$\rho(A_{dasg,Q}) = \max_{1\le i\le Nd} \left\{ \left| \frac{(1+\beta)\mu_i \pm \sqrt{(1+\beta)^2\mu_i^2 - 4\beta\mu_i}}{2} \right| \right\}.$$

Proof Consider the eigenvalue decomposition

$$\mathcal{W}-\alpha Q = R\mathbf{diag}\left([\mu_i]_{i=1}^{Nd}\right) R^T,$$

where R is real orthogonal and the eigenvalues μ_i are listed in non-increasing order. Next, we introduce the matrix

$$U = \mathbf{diag}(R, R), \tag{83}$$

and the permutation matrix P_π associated with the permutation π over $\{1, 2, \ldots, 2Nd\}$ that satisfies

$$\pi(i) = \begin{cases} 2i-1 & \text{if } 1 \le i \le Nd, \\ 2(i-Nd) & \text{if } Nd+1 \le i \le 2Nd. \end{cases} \tag{84}$$

By definition, $P_\pi^{-1} = P_\pi^T = P_{\pi^{-1}}$. Then, we can write

$$UA_{\text{dasg},Q}U^T = \begin{bmatrix} (1+\beta)\mathbf{diag}\left([\mu_i]_{i=1}^{Nd}\right) & -\beta\mathbf{diag}\left([\mu_i]_{i=1}^{Nd}\right) \\ I_{Nd} & 0_{Nd} \end{bmatrix} \tag{85}$$

$$= P_\pi^T \mathbf{diag}\left([\tilde{T}_i]_{i=1}^{Nd}\right) P_\pi, \tag{86}$$

and

$$\tilde{T}_i := \begin{bmatrix} (1+\beta)\mu_i & -\beta\mu_i \\ 1 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}, \qquad 1 \le i \le Nd. \tag{87}$$

Therefore, the eigenvalues of $A_{\text{dasg},Q}$ coincide with the eigenvalues of $\tilde{T}_i$ which can be computed explicitly as $\frac{(1+\beta)\mu_i \pm \sqrt{(1+\beta)^2\mu_i^2 - 4\beta\mu_i}}{2}$. This completes the proof. ■

Remark 37 *In the noiseless case (when $\sigma = 0$ and $w^k = 0$), we have*

$$\|\xi_k\| = \left\| A_{dasg,Q}^k \right\| \|\xi_0\|,$$

provided that $x^{(0)}$ is chosen as an eigenvector corresponding to a largest singular value of the $A_{dasg,Q}$ matrix. By Gelfand's formula, we have

$$\rho(A_{dasg,Q}) = \lim_{k\to\infty} \left(\|\xi_k\|/\|\xi_0\|\right)^{1/k}.$$

Therefore, Proposition 36 gives an explicit characterization of the asymptotic convergence rate.

For $\rho = \rho(A_{\text{dasg},Q}) < 1$, it is clear from the iterations (81) that the second moments $\mathbb{E}\left[\left\|x^{(k)} - x^{\infty}\right\|^2\right]$ will stay bounded over k. In fact, a consequence of Theorem 5 for strongly convex objectives is that, the variance of the iterates satisfies

$$\limsup_{k\to\infty} \mathbb{E}\left[\left\|x^{(k)} - x^{\infty}\right\|^2\right] \leq \frac{1}{1-\rho^2}\alpha^2 \frac{2\sigma^2 N}{\mu}\left(\tilde{P}_{11} + \frac{1-\lambda_N^W + \alpha L}{2\alpha}\right),$$

and hence the robustness measure satisfies

$$J_{\infty}(\alpha) \leq \frac{1}{1-\rho^2}\alpha^2 \frac{2}{\mu}\left(\tilde{P}_{11} + \frac{1-\lambda_N^W + \alpha L}{2\alpha}\right).$$

The next theorem shows that we can get a tighter explicit representation of the variance of the iterates in terms of the eigenvalues of the iteration matrix $A_{\text{dasg},Q}$.

Theorem 38 *Assume that Assumption 3 and Assumption 4 hold. Let μ_i be the eigenvalues of $\mathcal{W} - \alpha Q$. Then we have*

$$\lim_{k\to\infty} Var\left(x^{(k)} - x^{\infty}\right) = \frac{\sigma^2}{d}\sum_{i=1}^{Nd} \frac{\alpha^2(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))}, \tag{88}$$

and hence the robustness measure is given by

$$J_{\infty}(\alpha) = \frac{1}{Nd}\sum_{i=1}^{Nd} \frac{\alpha^2(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))}.$$

Proof Similar to the D-SG case, the equilibrium covariance matrix $X = \lim_{k\to\infty}\mathbb{E}[\xi_k\xi_k^T]$ of the D-ASG iterates satisfies the corresponding discrete Lyapunov equation

$$A_{dasg,Q} X A_{dasg,Q}^T + \frac{\sigma^2}{d} BB^T = 0.$$

where $A_{dasg,Q}$ is as in (82). The proof will be based on constructing a solution to this equation by block diagonalizing the matrix $A_{dasg,Q}$ with a change of variable technique. More specifically, if we introduce the matrix $Y = P_\pi(UXU^T)P_\pi^T$, where U is an orthogonal matrix defined by (83) and P_π is the permutation matrix defined in (84). It follows from (86) that Y satisfies the discrete Lyapunov equation:

$$\mathbf{diag}\left([\tilde{T}_i]_{i=1}^{Nd}\right) Y \left[\mathbf{diag}\left([\tilde{T}_i]_{i=1}^{Nd}\right)\right]^T - Y + \frac{\sigma^2}{d} P_\pi U^T BB^T U P_\pi^T = 0,$$

where T_i is defined by (87). Furthermore, $\text{trace}(Y) = \text{trace}(X)$ since U is orthogonal. Similar as in the proof of Proposition 3.7. Aybat et al. (2020), we can solve for Y which takes the block diagonal matrix form:

$$Y = \begin{bmatrix} Y_1 & 0_{Nd} & \cdots & 0_{Nd} \\ 0_{Nd} & Y_2 & \cdots & 0_{Nd} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{Nd} & 0_{Nd} & \cdots & Y_{Nd} \end{bmatrix}, \tag{89}$$

where Y_i satisfies the equation

$$\begin{bmatrix} (1+\beta)\mu_i & -\beta\mu_i \\ 1 & 0 \end{bmatrix} Y_i \begin{bmatrix} (1+\beta)\mu_i & 1 \\ -\beta\mu_i & 0 \end{bmatrix} - Y_i + \frac{\sigma^2}{d} \begin{bmatrix} \alpha^2 & 0 \\ 0 & 0 \end{bmatrix} = 0.$$

We can explicitly solve for Y_i and get

$$Y_i = \frac{\sigma^2}{d} \begin{bmatrix} \frac{\alpha^2(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))} & \frac{\alpha^2(1+\beta)\mu_i}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))} \\ \frac{\alpha^2(1+\beta)\mu_i}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))} & \frac{\alpha^2(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))} \end{bmatrix}. \tag{90}$$

Since $\xi_k = \left[\left(x^{(k)} - x^\infty\right)^T, \left(x^{(k-1)} - x^\infty\right)^T\right]^T$, we have

$$\begin{aligned} \lim_{k\to\infty} \mathrm{Var}\left(x^{(k)} - x^\infty\right) &= \lim_{k\to\infty} \frac{1}{2}\mathrm{Var}(\xi_k) = \frac{1}{2}\mathrm{trace}(X) = \frac{1}{2}\mathrm{trace}(Y) \\ &= \frac{\sigma^2}{d} \sum_{i=1}^{Nd} \frac{\alpha^2(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))}, \end{aligned}$$

which completes the proof. ■

Proposition 39 *Assume that Assumption 3 and Assumption 4 hold. For any $j = 1, \ldots, d$ and any $k \in \mathbb{N}$,*

$$\lim_{k\to\infty} Var\left(\bar{x}^{(k)}(j)\right) \leq \frac{\sigma^2}{Nd} \max_{i=1,2,\ldots,Nd} \frac{\alpha^2(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))},$$

where $\bar{x}^{(k)}(j)$ denotes the j-th entry of the node average $\bar{x}^{(k)}$ and μ_i are the eigenvalues of $\mathcal{W} - \alpha Q$.

Proof It follows from the proof of Proposition 38 that the covariance matrix has the form

$$\mathbb{E}\left[\xi_\infty {\xi_\infty}^T\right] = ZYZ^T, \tag{91}$$

where Y is as in (89), $Z = UP_\pi$ is orthogonal, where ξ_∞ is a random vector whose distribution coincides with the distribution of ξ_k in the limit as $k \to \infty$. For a random vector q with mean zero, let $\mathrm{Cov}(q)$ denote the covariance matrix of q, i.e. $\mathrm{Cov}(q) = \mathbb{E}[qq^T]$. It follows that

$$\mathrm{Cov}(Z^T\xi_\infty) = Z^T\mathbb{E}\left[\xi_\infty {\xi_\infty}^T\right] Z = Y,$$

where

$$Z^T\xi_\infty = \lim_{k\to\infty} \begin{bmatrix} r_1^T x^{(k)} \\ r_1^T x^{(k-1)} \\ r_2^T x^{(k)} \\ r_2^T x^{(k-1)} \\ \vdots \\ r_{Nd}^T x^{(k)} \\ r_{Nd}^T x^{(k-1)} \end{bmatrix},$$

where r_i are the columns of R in the eigenvalue decomposition $\mathcal{W}-\alpha Q = R\mathbf{diag}\left([\mu_i]_{i=1}^{Nd}\right)R^T$. In other words, r_i are the eigenvectors corresponding to the eigenvalues μ_i. Using the block diagonal structure of Y with blocks Y_i, this shows that

$$\begin{aligned}&\lim_{k\to\infty}\mathrm{Cov}\left(\begin{bmatrix} r_i^T x^{(k)} \\ r_i^T x^{(k-1)}\end{bmatrix}\right) = Y_i \in \mathbb{R}^2,\\ &\lim_{k\to\infty}\mathbb{E}\left(\left(r_i^T x^{(k)}\right)\left(r_j^T x^{(k)}\right)\right) = Y_{2i-1,2j-1} = 0 \quad \text{for} \quad i\neq j,\end{aligned} \tag{92}$$

where Y_i is given by (90) and the matrix Y is given by (89). Therefore,

$$\lim_{k\to\infty}\mathrm{Var}\left(v_i^T x^{(k)}\right) = \begin{bmatrix}1 & 0\end{bmatrix} Y_i \begin{bmatrix}1\\0\end{bmatrix} = \frac{\sigma^2}{d}\frac{\alpha^2(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))}, \tag{93}$$

where Var denotes the variance and we used (90). The eigenvectors v_i are not explicitly available, but we know they are orthogonal forming a basis; therefore for any unit vector $u\in\mathbb{R}^{Nd}$, we can express it in a unique way as linear combinations of the basis vectors v_i, i.e.

$$u = \sum_{i=1}^{Nd} m_i v_i, \qquad m_i = \langle u, v_i\rangle,$$

for some scalars m_i that are not all zero. Since u has unit norm in $\mathbb{R}^{Nd}$, we have also

$$\|u\|^2 = 1 = \sum_{i=1}^{Nd} m_i^2.$$

Consequently,

$$\begin{aligned}&\lim_{k\to\infty}\mathrm{Var}\left(u^T x^{(k)}\right)\\ &= \lim_{k\to\infty}\mathrm{Var}\left(\left(\sum_{i=1}^{Nd} m_i r_i^T\right)x^{(k)}\right)\\ &= \sum_{i=1}^{Nd}\alpha_i^2\lim_{k\to\infty}\mathrm{Var}\left(r_i^T x^{(k)}\right) + 2\sum_{1\le i<j\le Nd} m_i m_j \lim_{k\to\infty}\mathbb{E}\left[\left(r_i^T x^{(k)}\right)\left(r_j^T x^{(k)}\right)\right]\\ &= \sum_{i=1}^{Nd} m_i^2\lim_{k\to\infty}\mathrm{Var}\left(r_i^T x^{(k)}\right)\\ &= \sum_{i=1}^{Nd}(\langle u, v_i\rangle)^2\frac{\sigma^2}{d}\frac{\alpha^2(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))},\end{aligned} \tag{94}$$

where we used (92). This formula expresses the asymptotic variance along any unit direction u. However, we can also obtain an upper bound from (94),

$$\lim_{k\to\infty} \mathrm{Var}\left(u^T x^{(k)}\right) \leq \max_{1\leq i\leq Nd} \lim_{k\to\infty} \mathrm{Var}\left(r_i^T x^{(k)}\right) \sum_{i=1}^{Nd} m_i^2 \tag{95}$$

$$= \max_{1\leq i\leq Nd} \lim_{k\to\infty} \mathrm{Var}\left(r_i^T x^{(k)}\right) \tag{96}$$

$$= \frac{\sigma^2}{d} \max_{i=1,2,\ldots,Nd} \frac{\alpha^2(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))}, \tag{97}$$

for any unit vector $u \in \mathbb{R}^{Nd}$ where we used (93). Furthermore, this bound does not grow with N as the eigenvalues μ_i are bounded satisfying $\lambda_N^W - \alpha L \leq \mu_i \leq 1 - \alpha\mu$. If we choose the vector u such that its entries are $u_j = 1/\sqrt{N}$ if $j \in \{1, d+1, 2d+1, \ldots, (N-1)d+1\}$ else 0. Then, u is a unit vector satisfying

$$u^T x^{(k)} = \sqrt{N}\bar{x}^{(k)}(1),$$

where $\bar{x}^{(k)}(1)$ denotes the first entry of the node average $\bar{x}^{(k)}$. Therefore, we get

$$\lim_{k\to\infty} \mathrm{Var}\left(u^T x^{(k)}\right) = N \lim_{k\to\infty} \mathrm{Var}\left(\bar{x}^{(k)}(1)\right)$$
$$\leq \frac{\sigma^2}{d} \max_{i=1,2,\ldots,Nd} \frac{\alpha^2(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))}.$$

Consequently,

$$\lim_{k\to\infty} \mathrm{Var}\left(\bar{x}^{(k)}(1)\right) \leq \frac{\sigma^2}{Nd} \max_{i=1,2,\ldots,Nd} \frac{\alpha^2(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))}.$$

Similarly, choosing u appropriately, we can obtain

$$\lim_{k\to\infty} \mathrm{Var}\left(\bar{x}^{(k)}(j)\right) \leq \frac{\sigma^2}{Nd} \max_{i=1,2,\ldots,Nd} \frac{\alpha^2(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))},$$

for any $j = 1, \ldots, d$, which completes the proof. ■

Next, for D-ASG iterates, we provide bounds on $\mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right]$ and $\mathbb{E}\left[\left\|x^{(k)} - x^*\right\|^2\right]$.

Theorem 40 *Assume that Assumption 3 and Assumption 4 hold. Consider the D-ASG iterates. For every $k \in \mathbb{N}$,*

$$\mathbb{E}\left[\left\|x^{(k)} - x^\infty\right\|^2\right] \leq (C_k)^2 \rho_{dasg}^{2k} \left(\left\|\xi_0 \xi_0^T\right\| + \frac{\alpha^2\sigma^2 N}{1-\rho_{dasg}^2}\right)$$
$$+ \frac{\alpha^2\sigma^2}{d} \sum_{i=1}^{Nd} \frac{(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))}. \tag{98}$$

In addition, if $\alpha \leq \frac{1}{L+\mu}$, *we have:*

$$\mathbb{E}\left[\left\|x^{(k)}-x^*\right\|^2\right] \leq (C_k)^2\rho_{dasg}^{2k}\left(2\left\|\xi_0\xi_0^T\right\|+\frac{2\alpha^2\sigma^2N}{1-\rho_{dasg}^2}\right)+2\frac{\alpha^2C_1^2N}{(1-\gamma)^2} + \frac{\alpha^2\sigma^2}{d}\sum_{i=1}^{Nd}\frac{(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))}, \quad (99)$$

where C_k, ρ_{dasg} *are defined in Lemma 41 and* μ_i *are the eigenvalues of* $\mathcal{W}-\alpha Q$.

In particular, when $\beta=\frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$, $\lambda_N^W>0$ *and* $\alpha\in\left(0,\min\left\{\frac{1}{L+\mu},\frac{\lambda_N^W}{L}\right\}\right]$, *we have*

$$\mathbb{E}\left[\left\|x^{(k)}-x^\infty\right\|^2\right] \leq (C_k)^2(1-\sqrt{\alpha\mu})^{2k}\left(\left\|\xi_0\xi_0^T\right\|+\frac{\alpha^2\sigma^2N}{1-(1-\sqrt{\alpha\mu})^2}\right) + \frac{\alpha^2\sigma^2}{d}\sum_{i=1}^{Nd}\frac{(1+\sqrt{\alpha\mu})(1+\sqrt{\alpha\mu}+(1-\sqrt{\alpha\mu})\mu_i)}{(1-\mu_i)(1+\sqrt{\alpha\mu}-(1-\sqrt{\alpha\mu})\mu_i)(4-(1-\mu_i)(3-\sqrt{\alpha\mu}))}, \quad (100)$$

$$\mathbb{E}\left[\left\|x^{(k)}-x^*\right\|^2\right] \leq (C_k)^2(1-\sqrt{\alpha\mu})^{2k}\left(2\left\|\xi_0\xi_0^T\right\|+\frac{2\alpha^2\sigma^2N}{1-(1-\sqrt{\alpha\mu})^2}\right)+2\frac{\alpha^2C_1^2N}{(1-\gamma)^2} + \frac{\alpha^2\sigma^2}{d}\sum_{i=1}^{Nd}\frac{(1+\sqrt{\alpha\mu})(1+\sqrt{\alpha\mu}+(1-\sqrt{\alpha\mu})\mu_i)}{(1-\mu_i)(1+\sqrt{\alpha\mu}-(1-\sqrt{\alpha\mu})\mu_i)(4-(1-\mu_i)(3-\sqrt{\alpha\mu}))}, \quad (101)$$

where μ_i *are the eigenvalues of* $\mathcal{W}-\alpha Q$ *and*

$$C_k=\max\left\{2k-1,\max_{i:0<\mu_i<1-\alpha\mu}\frac{1+\sqrt{\alpha\mu}+(1-\sqrt{\alpha\mu})\mu_i}{2\sqrt{\mu_i(1-\alpha\mu-\mu_i)}}\right\}.$$

Before we proceed to the proof of Theorem 40, let us first derive the following lemma providing an upper bound on the norm of $A_{\text{dasg},Q}^k$ for every $k\in\mathbb{N}$, which will be used later.

Lemma 41 *For any* $k\in\mathbb{N}$,

$$\left\|A_{dasg,Q}^k\right\|\leq C_k\rho_{dasg}^k,$$

where

$$C_k:=\max\left\{2k-1,\max_{i:\gamma_{i,+}\neq\gamma_{i,-}}\frac{1+\max\{|\gamma_{i,+}|,|\gamma_{i,-}|\}^2}{|\gamma_{i,+}-\gamma_{i,-}|}\right\},$$
$$\rho_{dasg}:=\max_{1\leq i\leq Nd}\max\{|\gamma_{i,+}|,|\gamma_{i,-}|\},$$

where $\gamma_{i,\pm}:=\frac{(1+\beta)\mu_i\pm\sqrt{(1+\beta)^2\mu_i^2-4\beta\mu_i}}{2}$, *and* μ_i *are the eigenvalues of* $A_Q=\mathcal{W}-\alpha Q$.

In particular, when $\beta=\frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$, $\lambda_N^W>0$ *and* $\alpha\in\left(0,\frac{\lambda_N^W}{L}\right]$, *we have* $\rho_{dasg}=1-\sqrt{\alpha\mu}$, *and*

$$C_k=\max\left\{2k-1,\max_{i:0<\mu_i<1-\alpha\mu}\frac{1+\sqrt{\alpha\mu}+(1-\sqrt{\alpha\mu})\mu_i}{2\sqrt{\mu_i(1-\alpha\mu-\mu_i)}}\right\}.$$

Proof The proof of Lemma 41 will be provided in Appendix F. ■

Now, we are ready to prove Theorem 40.

C.2.1 Proof of Theorem 40

Proof We recall that

$$\xi_{k+1} = A_{\text{dasg},Q}\xi_k + B_{\text{dasg}}w^{(k+1)},$$

and therefore, we get:

$$\mathbb{E}\left[\xi_k\xi_k^T\right] = A_{\text{dasg},Q}\mathbb{E}\left[\xi_{k-1}\xi_{k-1}^T\right](A_{\text{dasg},Q})^T + \begin{pmatrix} \alpha^2\frac{\sigma^2}{d}I_{Nd} & 0_{Nd} \\ 0_{Nd} & 0_{Nd} \end{pmatrix}, \tag{102}$$

Therefore,

$$X_{\text{dasg}} := \mathbb{E}\left[\xi_\infty\xi_\infty^T\right]$$

satisfies the discrete Lyapunov equation:

$$X_{\text{dasg}} = A_{\text{dasg},Q}X_{\text{dasg}}(A_{\text{dasg},Q})^T + \begin{pmatrix} \alpha^2\frac{\sigma^2}{d}I_{Nd} & 0_{Nd} \\ 0_{Nd} & 0_{Nd} \end{pmatrix}.$$

By Theorem 38 we have

$$\text{trace}(X_{\text{dasg}}) = \frac{\sigma^2}{d}\sum_{i=1}^{Nd}\frac{\alpha^2(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))}.$$

Next by iterating equation (102) over k, we immediately obtain

$$\begin{aligned}\mathbb{E}\left[\xi_k\xi_k^T\right] = (A_{\text{dasg},Q})^k\,\xi_0\xi_0^T\left((A_{\text{dasg},Q})^T\right)^k \\ + \sum_{j=0}^{k-1}(A_{\text{dasg},Q})^j\begin{pmatrix} \alpha^2\frac{\sigma^2}{d}I_{Nd} & 0_{Nd} \\ 0_{Nd} & 0_{Nd} \end{pmatrix}\left((A_{\text{dasg},Q})^T\right)^j,\end{aligned}$$

so that

$$\begin{aligned}\mathbb{E}\left[\xi_k\xi_k^T\right] = \mathbb{E}\left[\xi_\infty\xi_\infty^T\right] + (A_{\text{dasg},Q})^k\,\xi_0\xi_0^T\left((A_{\text{dasg},Q})^T\right)^k \\ - \sum_{j=k}^{\infty}(A_{\text{dasg},Q})^j\begin{pmatrix} \alpha^2\frac{\sigma^2}{d}I_{Nd} & 0_{Nd} \\ 0_{Nd} & 0_{Nd} \end{pmatrix}\left((A_{\text{dasg},Q})^T\right)^j,\end{aligned}$$

which implies that

$$\begin{aligned}\text{trace}\left(\mathbb{E}\left[\xi_k\xi_k^T\right]\right) &= \text{trace}\left(\mathbb{E}\left[\xi_\infty\xi_\infty^T\right]\right) + (A_{\text{dasg},Q})^k\,\xi_0\xi_0^T\left((A_{\text{dasg},Q})^T\right)^k \\ &\quad - \sum_{j=k}^{\infty}(A_{\text{dasg},Q})^j\begin{pmatrix} \alpha^2\frac{\sigma^2}{d}I_{Nd} & 0_{Nd} \\ 0_{Nd} & 0_{Nd} \end{pmatrix}\left((A_{\text{dasg},Q})^T\right)^j \\ &\le \text{trace}(X_{\text{dasg}}) + \left\|(A_{\text{dasg},Q})^k\right\|^2\|\xi_0\xi_0^T\| + \sum_{j=k}^{\infty}\left\|(A_{\text{dasg},Q})^j\right\|^2\alpha^2\sigma^2N \\ &\le \text{trace}(X_{\text{dasg}}) + (C_k)^2(\rho_{\text{dasg}})^{2k}\|\xi_0\xi_0^T\| + \alpha^2\sigma^2N(C_k)^2\frac{(\rho_{\text{dasg}})^{2k}}{1-(\rho_{\text{dasg}})^2},\end{aligned}$$

where we used the estimate from the proof of Lemma 41.

Note that $\xi_k = x^{(k)} - x^\infty$, this proves (98). Moreover,

$$\left\|x^{(k)} - x^*\right\|^2 \leq 2\left\|x^{(k)} - x^\infty\right\|^2 + 2\left\|x^\infty - x^*\right\|^2,$$

and together with (8), it proves (99).

Finally, when $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$, $\lambda_N^W > 0$ and $\alpha \in (0, \frac{\lambda_N^W}{L}]$, we have $\rho_{\text{dasg}} = 1 - \sqrt{\alpha\mu}$, and

$$\left\|A_{\text{dasg},Q}^k\right\| \leq C_k \cdot (1 - \sqrt{\alpha\mu})^k,$$

where

$$C_k = \max\left\{2k-1, \max_{i:0<\mu_i<1-\alpha\mu} \frac{1+\sqrt{\alpha\mu}+(1-\sqrt{\alpha\mu})\mu_i}{2\sqrt{\mu_i(1-\alpha\mu-\mu_i)}}\right\},$$

and by Theorem 38 with $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$ we have

$$\begin{aligned}\text{trace}(X_{\text{dasg}}) &= \frac{\sigma^2}{d}\sum_{i=1}^{Nd} \frac{\alpha^2(1+\beta\mu_i)}{(1-\mu_i)(1-\beta\mu_i)(2+2\beta-(1-\mu_i)(1+2\beta))} \\ &= \frac{\alpha^2\sigma^2}{d}\sum_{i=1}^{Nd} \frac{(1+\sqrt{\alpha\mu})(1+\sqrt{\alpha\mu}+(1-\sqrt{\alpha\mu})\mu_i)}{(1-\mu_i)(1+\sqrt{\alpha\mu}-(1-\sqrt{\alpha\mu})\mu_i)(4-(1-\mu_i)(3-\sqrt{\alpha\mu}))}.\end{aligned}$$

The proof is complete. ■

Appendix D. Proofs of Main Results in Section 3

D.0.1 Proof of Proposition 12

Proof Recall that, from Theorem 7, we have

$$\mathbb{E}\left[\left\|x^{(k)} - x^*\right\|^2\right] \leq 4\left(1-\sqrt{\alpha\mu}\right)^k \frac{V_{S,\alpha}\left(\xi_0\right)}{\mu} + \frac{2\sigma^2 N\alpha}{\mu\sqrt{\alpha\mu}}\left(2 - \lambda_N^W + \alpha L\right) + \frac{2C_1^2 N\alpha^2}{(1-\gamma)^2}. \quad (103)$$

Next, note that, $\xi_0 = \left[\left(x^{(0)} - x^\infty\right)^\top, \left(x^{(0)} - x^\infty\right)^\top\right]^\top$, and therefore,

$$\begin{aligned}V_{S,\alpha}\left(\xi_0\right) &= \xi_0^\top S_\alpha \xi_0 \\ &= \left\|x^{(0)} - x^\infty\right\|^2 \left(\frac{1}{2\alpha} + \left(\sqrt{\frac{\mu}{2}} - \sqrt{\frac{1}{2\alpha}}\right)^2 + \frac{\sqrt{2}}{\sqrt{\alpha}}\left(\sqrt{\frac{\mu}{2}} - \sqrt{\frac{1}{2\alpha}}\right)\right) \\ &= \frac{\mu}{2}\left\|x^{(0)} - x^\infty\right\|^2 \\ &\leq \mu\left\|x^{(0)} - x^*\right\|^2 + \mu\left\|x^\infty - x^*\right\|^2 \\ &\leq \mu\left(\left\|x^{(0)} - x^*\right\|^2 + \frac{C_1^2 N\alpha^2}{(1-\gamma)^2}\right),\end{aligned}$$

where the last inequality follows from (8). Plugging this bound in (103) along with these straightforward inequalities

$$1-\sqrt{\alpha\mu}\leq\exp(-\sqrt{\alpha\mu}),\quad 2-\lambda_N^W+\alpha L\leq 3,\quad 2+4\left(1-\sqrt{\alpha\mu}\right)^k\leq 6,$$

completes the proof. ■

D.0.2 Proof of Corollary 13

Proof First of all, notice that $x\mapsto\frac{\log x}{x}$ is decreasing for any $x\geq e$. To simplify the notation, let $\hat{k}=\max\left\{2\log(p\sqrt{\tilde{\kappa}}),e\right\}$. First note that, since $k\geq p\sqrt{\tilde{\kappa}}\hat{k}$, we have

$$\frac{p\sqrt{\tilde{\kappa}}\log k}{k}\leq\frac{p\sqrt{\tilde{\kappa}}\log(p\sqrt{\tilde{\kappa}}\hat{k})}{p\sqrt{\tilde{\kappa}}\hat{k}}=\frac{\log(p\sqrt{\tilde{\kappa}})+\log\hat{k}}{\hat{k}}\leq\frac{1}{2}+\frac{\log\hat{k}}{\hat{k}}\leq 1,\tag{104}$$

where the second inequality follows from $\hat{k}\geq 2\log(p\sqrt{\tilde{\kappa}})$ and the last inequality is obtained using $\hat{k}\geq e$. Hence, α_1 satisfies the condition $\alpha_1\leq\min\{\lambda_N^W/L,1/(L+\mu)\}$ in Proposition 12. In addition, note that α_1 can be written as

$$\alpha_1=\frac{\lambda_N^W}{L+\mu}\left(p\sqrt{\tilde{\kappa}}\log k/k\right)^2=\frac{1}{\mu}\left(p\log k/k\right)^2.$$

Plugging this into Proposition 12 completes the proof. ■

D.0.3 Proof of Proposition 14

Proof We show this result by using induction. First note that, for $t=0$, the argument holds using Proposition 12. Now, assume it holds for t and we show it for $t+1$. Using Proposition 12, and taking expectation from both sides, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|x^{(L_{t+2})}-x^*\right\|^2\right]\\
\leq&4\exp\left(-k_{t+1}\sqrt{\alpha_{t+1}\mu}\right)\mathbb{E}\left[\left\|x^{(L_{t+1})}-x^*\right\|^2\right]+6N\left(\frac{\sqrt{\alpha_{t+1}}}{\mu\sqrt{\mu}}\sigma^2+\frac{C_1^2\alpha_{t+1}^2}{(1-\gamma)^2}\right)\\
=&\frac{1}{2^{p-2}}\mathbb{E}\left[\left\|x^{(L_{t+1})}-x^*\right\|^2\right]+\frac{6N}{2^{t+1}}\sqrt{\frac{\lambda_N^W}{(L+\mu)\mu^3}}\sigma^2+\frac{6N}{2^{4(t+1)}}\left(\frac{C_1\lambda_N^W}{(L+\mu)(1-\gamma)}\right)^2 &(105)\\
\leq&\frac{4}{2^{(p-2)(t+1)}}\exp\left(-\frac{k_1}{\sqrt{\tilde{\kappa}}}\right)\left\|x^{(0)}-x^*\right\|^2+12N\left(\frac{1/2^{(p-2)}}{2^t}+\frac{1/2}{2^{t+1}}\right)\frac{\sigma^2}{\mu^2\sqrt{\tilde{\kappa}}}\\
&+12N\left(\frac{1/2^{(p-2)}}{2^{4t}}+\frac{1/2}{2^{4(t+1)}}\right)\left(\frac{C_1\lambda_N^W}{(L+\mu)(1-\gamma)}\right)^2 &(106)\\
\leq&\frac{4}{2^{(p-2)(t+1)}}\exp\left(-\frac{k_1}{\sqrt{\tilde{\kappa}}}\right)\left\|x^{(0)}-x^*\right\|^2+\frac{12N}{2^{t+1}}\frac{\sigma^2}{\mu^2\sqrt{\tilde{\kappa}}}+\frac{12N}{2^{4(t+1)}}\left(\frac{C_1\lambda_N^W}{L(1-\gamma)}\right)^2, &(107)
\end{aligned}$$

where (105) follows from substituting α_{t+1} and k_{t+1} and (106) is obtained using the induction hypothesis for t. Finally, (107) is obtained by replacing $L+\mu$ by L in (106) along with the assumption $p \geq 7$ so that the term $12N\left(\frac{1/2^{(p-2)}}{2^{4t}} + \frac{1/2}{2^{4(t+1)}}\right)$ in network effect in (106) is bounded by $\frac{12N}{2^{4(t+1)}}$ in (107), which completes the proof. ■

D.0.4 Proof of Proposition 15

Proof Let T denote the largest t such that $k \geq L_t$. In particular, we have

$$L_T \leq k < L_{T+1}.$$

Now, using Proposition 12, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|x^{(k)} - x^*\right\|^2\right] \leq 4\mathbb{E}\left[\left\|x^{(L_T)} - x^*\right\|^2\right] + 6N\left(\frac{\sqrt{\alpha_{T+1}}}{\mu\sqrt{\mu}}\sigma^2 + \frac{C_1^2\alpha_{T+1}^2}{(1-\gamma)^2}\right) \\
=&4\mathbb{E}\left[\left\|x^{(L_T)} - x^*\right\|^2\right] + \frac{6N}{2^{T+1}}\sqrt{\frac{\lambda_N^W}{(L+\mu)\mu^3}}\sigma^2 + \frac{6N}{2^{4(T+1)}}\left(\frac{C_1\lambda_N^W}{(L+\mu)(1-\gamma)}\right)^2 \\
\leq&\mathcal{O}(1)\left(\frac{1}{2^{(p-2)T}}\exp\left(-\frac{k_1}{\sqrt{\tilde{\kappa}}}\right)\left\|x^{(0)} - x^*\right\|^2 + \frac{N\sigma^2}{2^T\mu^2\sqrt{\tilde{\kappa}}} + \frac{N}{2^{4T}}\left(\frac{C_1\lambda_N^W}{L(1-\gamma)}\right)^2\right), \qquad (108)
\end{aligned}$$

where the last inequality follows from Proposition 14. Next, note that

$$k - k_1 \leq L_{T+1} - k_1 \leq 2(L_T - k_1) \leq p2^{T+3}\log(2)\sqrt{\tilde{\kappa}},$$

where the last two inequalities follows from the special pattern of the sequence $\{k_i\}_i$. Therefore, we have

$$\frac{1}{2^T} \leq \frac{8p\log(2)\sqrt{\tilde{\kappa}}}{k - k_1} \leq \frac{6p\sqrt{\tilde{\kappa}}}{k - k_1}.$$

Plugging this bound in (108) completes the proof. ■

D.0.5 Proof of Corollary 18

Proof By Proposition 10, for any k, we have

$$\begin{aligned}
&\mathbb{E}\left\|\bar{x}^{(k)} - x_*\right\|^2 \\
&\leq \left(1 - \frac{\sqrt{\alpha\mu}}{2}\right)^k \frac{2V_{\bar{S},\alpha}\left(\bar{\xi}_0\right)}{\mu} + \frac{4}{\mu\sqrt{\mu}}\left(\left(1 + \frac{\sqrt{\alpha\mu}}{2}\right)\frac{\sigma^2\sqrt{\alpha}}{2N}\left(1 + \alpha L\right) + \frac{1}{2\sqrt{\mu}}\alpha H_1 H_2\right) \\
&\quad + \frac{8}{\gamma^2\mu\sqrt{\mu}}\sqrt{\alpha}H_1H_3\frac{\gamma^{2k} - (1 - \sqrt{\alpha\mu}/2)^k}{\gamma^2 - (1 - \sqrt{\alpha\mu}/2)}.
\end{aligned}$$

where $V_{\bar{S},\alpha}$ is defined by (33). As $\alpha \to 0$, one can check that $H_1 = \mathcal{O}(1)$, $H_2 = \mathcal{O}(1)\frac{L^2}{N}\frac{C_0}{(1-\gamma)^2}$ and $H_3 = \mathcal{O}(1)\frac{L^2}{N}$ since it follows from the proof of Proposition 12 that $V_{S,\alpha}(\xi_0) \leq \mu\|x^{(0)} - x^*\|^2 + \mu\frac{C_1^2 N\alpha^2}{(1-\gamma)^2}$. When α is sufficiently small,

$$\begin{aligned}\frac{8}{\gamma^2\mu\sqrt{\mu}}\sqrt{\alpha}H_1H_3\frac{\gamma^{2k} - (1-\sqrt{\alpha\mu}/2)^k}{\gamma^2 - (1-\sqrt{\alpha\mu}/2)} &\leq \frac{8}{\gamma^2\mu\sqrt{\mu}}\sqrt{\alpha}H_1H_3\frac{(1-\sqrt{\alpha\mu}/2)^k}{(1-\sqrt{\alpha\mu}/2) - \gamma^2} \\ &\leq \left(1 - \frac{\sqrt{\alpha\mu}}{2}\right)^k \frac{2V_{\bar{S},\alpha}\left(\bar{\xi}_0\right)}{\mu}.\end{aligned}$$

Moreover, it follows similarly as in the proof of Proposition 12 that $V_{\bar{S},\alpha}(\bar{\xi}_0) \leq \frac{\mu}{N}\|x^{(0)} - x^*\|^2 + \mu\frac{C_1^2\alpha^2}{(1-\gamma)^2}$. Hence, as $\alpha \to 0$, we have

$$\begin{aligned}&\mathbb{E}\left\|\bar{x}^{(k)} - x_*\right\|^2 \\ &\leq \mathcal{O}(1)\left(\left(1 - \frac{\sqrt{\alpha\mu}}{2}\right)^k \frac{1}{N}\|x^{(0)} - x^*\|^2 + \frac{1}{\mu\sqrt{\mu}}\left(\frac{\sigma^2\sqrt{\alpha}}{N} + \frac{1}{\sqrt{\mu}}\alpha\frac{L^2}{N}\frac{C_0}{(1-\gamma)^2}\right)\right).\end{aligned}$$

Then, similar as in Corollary 16, we can show that by choosing $k_1 = \lceil (p-2)\log(6p\tilde{\kappa})\sqrt{\tilde{\kappa}}\rceil$, we have

$$\mathbb{E}\left[\left\|\bar{x}^{(k)} - x_*\right\|^2\right] \leq \mathcal{O}(1)\left(\frac{1}{k^{p-2}}\frac{\left\|x^{(0)} - x^*\right\|^2}{N} + \frac{p\sigma^2}{N\mu\sqrt{\mu}k} + \frac{p^4 C_0 L^2(1-\gamma)^{-2}}{N\mu^2 k^4}\right),$$

for any $k \geq 2k_1$. Also, for a given number of iterations, k, by choosing $p = 7$ and $k_1 = \lceil \frac{k}{C}\rceil$ for some constant $C \geq 2$, we have

$$\mathbb{E}\left[\left\|\bar{x}^{(k)} - x_*\right\|^2\right] \leq \mathcal{O}(1)\left(\exp\left(-\frac{k}{C\sqrt{\tilde{\kappa}}}\right)\frac{\left\|x^{(0)} - x^*\right\|^2}{N} + \frac{\sigma^2}{N\mu\sqrt{\mu}k} + \frac{C_0 L^2(1-\gamma)^{-2}}{N\mu^2 k^4}\right),$$

for any $k \geq 2\sqrt{\tilde{\kappa}}$, where C_1, γ are given in (8) and $\tilde{\kappa}$ is given in (38).

Moreover, we recall from Proposition 10 that for every $i = 1, 2, \ldots, N$ and any k,

$$\begin{aligned}&\mathbb{E}\left\|x_i^{(k)} - x_*\right\|^2 \\ &\leq \left(1 - \frac{\sqrt{\alpha\mu}}{2}\right)^k \frac{4V_{\bar{S},\alpha}\left(\bar{\xi}_0\right)}{\mu} + \frac{8}{\mu\sqrt{\mu}}\left(\left(1 + \frac{\sqrt{\alpha\mu}}{2}\right)\frac{\sigma^2\sqrt{\alpha}}{2N}(1+\alpha L) + \frac{1}{2\sqrt{\mu}}\alpha H_1 H_2\right) \\ &\quad + \frac{16}{\gamma^2\mu\sqrt{\mu}}\sqrt{\alpha}H_1H_3\frac{\gamma^{2k} - (1-\sqrt{\alpha\mu}/2)^k}{\gamma^2 - (1-\sqrt{\alpha\mu}/2)} \\ &\quad + 16\gamma^{2k}\left(4\frac{V_{S,\alpha}\left(\xi_0\right)}{\mu} + \frac{2\sigma^2 N\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(2 - \lambda_N^W + \alpha L\right) + \frac{2C_1^2 N\alpha^2}{(1-\gamma)^2} + \|x^*\|^2\right) \\ &\quad + \frac{8D_y^2\alpha^2}{(1-\gamma)^2} + \frac{8\sigma^2 N\alpha^2}{(1-\gamma)^2} + \frac{16C_0\alpha}{(1-\gamma)^2}.\end{aligned}$$

When α is sufficiently small,

$$16\gamma^{2k}\left(4\frac{V_{S,\alpha}\left(\xi_0\right)}{\mu}+\frac{2\sigma^2N\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(2-\lambda_N^W+\alpha L\right)+\frac{2C_1^2N\alpha^2}{(1-\gamma)^2}+\|x^*\|^2\right)$$
$$\leq\left(1-\frac{\sqrt{\alpha\mu}}{2}\right)^k\frac{4V_{\bar{S},\alpha}\left(\bar{\xi}_0\right)}{\mu}.$$

Similar as before, we can show that as $\alpha\to 0$, we have

$$\mathbb{E}\left\|x_i^{(k)}-x_*\right\|^2$$
$$\leq\mathcal{O}(1)\left(\left(1-\frac{\sqrt{\alpha\mu}}{2}\right)^k\frac{\|x^{(0)}-x^*\|^2}{N}+\frac{1}{\mu\sqrt{\mu}}\left(\frac{\sigma^2\sqrt{\alpha}}{N}+\frac{1}{\sqrt{\mu}}\alpha\frac{L^2}{N}\frac{C_0}{(1-\gamma)^2}\right)+\frac{C_0\alpha}{(1-\gamma)^2}\right),$$

and thus similar as in Corollary 16, we can show that by choosing $k_1=\lceil(p-2)\log(6p\tilde{\kappa})\sqrt{\tilde{\kappa}}\rceil$, we have

$$\mathbb{E}\left[\left\|x_i^{(k)}-x_*\right\|^2\right]$$
$$\leq\mathcal{O}(1)\left(\frac{1}{k^{p-2}}\frac{\left\|x^{(0)}-x^*\right\|^2}{N}+\frac{p\sigma^2}{N\mu\sqrt{\mu}k}+\left(\frac{L^2}{N\mu^2}+1\right)\frac{p^4C_0(1-\gamma)^{-2}}{k^4}\right),$$

for any $k\geq 2k_1$ and $i=1,2,\ldots,N$. Also, for a given number of iterations, k, by choosing $p=7$ and $k_1=\lceil\frac{k}{C}\rceil$ for some constant $C\geq 2$, we have

$$\mathbb{E}\left[\left\|x_i^{(k)}-x_*\right\|^2\right]$$
$$\leq\mathcal{O}(1)\left(\exp\left(-\frac{k}{C\sqrt{\tilde{\kappa}}}\right)\frac{\left\|x^{(0)}-x^*\right\|^2}{N}+\frac{\sigma^2}{N\mu\sqrt{\mu}k}+\left(\frac{L^2}{N\mu^2}+1\right)\frac{C_0(1-\gamma)^{-2}}{k^4}\right),$$

for any $k\geq 2\sqrt{\tilde{\kappa}}$ and $i=1,2,\ldots,N$, where C_1,γ are given in (8) and $\tilde{\kappa}$ is given in (38). The proof is complete. ■

Appendix E. Results for More General Noise Setting

Consider the following assumption on noise which is more general than Assumption 1, and we will show that the main results in this paper for D-SG and D-ASG still hold under this more general assumption on gradient noise.

Assumption 5 *Recall that $x_i^{(k)}$ denotes the decision variable of node i at iteration k. We assume at iteration k, node i has access to $\tilde{\nabla}f_i\left(x_i^{(k)},w_i^{(k)}\right)$ which is an estimate of $\nabla f_i\left(x_i^{(k)}\right)$ where $w_i^{(k)}$ is a random variable independent of $\left\{w_j^{(t)}\right\}_{j=1,\ldots,N,t=1,\ldots,k-1}$ and*

$\left\{w_j^{(k)}\right\}_{j\neq i}$. *Moreover, we assume* $\mathbb{E}\left[\tilde{\nabla} f_i\left(x_i^{(k)}, w_i^{(k)}\right)\middle| x_i^{(k)}\right] = \nabla f_i\left(x_i^{(k)}\right)$ *and*

$$\mathbb{E}\left[\left\|\tilde{\nabla} f_i\left(x_i^{(k)}, w_i^{(k)}\right) - \nabla f_i\left(x_i^{(k)}\right)\right\|^2 \middle| x_i^{(k)}\right] \leq \sigma^2 + \frac{\eta^2}{2}\left\|x_i^{(k)} - x_*\right\|^2.$$

for some constant $\eta > 0$. *To simplify the notation, we suppress the* $w_i^{(k)}$ *dependence, and denote* $\tilde{\nabla} f_i\left(x_i^{(k)}, w_i^{(k)}\right)$ *by* $\tilde{\nabla} f_i\left(x_i^{(k)}\right)$.

Such assumptions could hold if gradients are estimates from batches (randomly selected subset of data points) in the context of empirical risk minimization problems (Jain et al., 2018; Gürbüzbalaban et al., 2021). The constant η^2 is often inversely proportional to the batch size (see e.g. Raginsky et al. (2017)).

E.1 Distributed Stochastic Gradient (D-SG)

Let us recall the D-SG in (6), which takes the equivalent form (7). Then it follows from Assumption 5 that $\mathbb{E}\left[\tilde{\nabla} F\left(x^{(k)}\right)\middle| x^{(k)}\right] = \nabla F\left(x^{(k)}\right)$ and

$$\mathbb{E}\left[\left\|\tilde{\nabla} F\left(x^{(k)}\right) - \nabla F\left(x^{(k)}\right)\right\|^2 \middle| x^{(k)}\right] \leq \sigma^2 N + \frac{\eta^2}{2}\left\|x^{(k)} - x^*\right\|^2. \tag{109}$$

We recall that $\|x^\infty - x^*\| \leq \frac{\alpha C_1\sqrt{N}}{(1-\gamma)}$ from (8). Therefore, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\tilde{\nabla} F\left(x^{(k)}\right) - \nabla F\left(x^{(k)}\right)\right\|^2 \middle| x^{(k)}\right] &\leq \sigma^2 N + \eta^2\left\|x^{(k)} - x^\infty\right\|^2 + \eta^2\left\|x^\infty - x^*\right\|^2 \\ &\leq \left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N + \eta^2\left\|x^{(k)} - x^\infty\right\|^2. \end{aligned} \tag{110}$$

We have the following explicit performance bounds on the convergence and the robustness of D-SG iterates.

Theorem 42 *Assume that* $\alpha \leq \frac{\frac{1}{2}+\lambda_N^W}{L+\frac{\eta^2}{\mu}}$. *For any* $k \geq 0$,

$$\mathbb{E}\left\|x^{(k)} - x^*\right\|^2 \leq 2\left(1-\alpha\mu/2\right)^k \mathbb{E}\left\|x^{(0)} - x^\infty\right\|^2 + \frac{4\alpha}{\mu}\left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N + \frac{2\alpha^2(C_1)^2 N}{(1-\gamma)^2}.$$

Proof The D-SG iterates are given by

$$x^{(k+1)} = x^{(k)} - \alpha\nabla F_{\mathcal{W},\alpha}\left(x^{(k)}\right) - \alpha\xi^{(k+1)}, \tag{111}$$

where $\mathbb{E}\left[\xi^{(k+1)}|\mathcal{F}_k\right] = 0$ and

$$\mathbb{E}\left[\left\|\xi^{(k+1)}\right\|^2 \middle| \mathcal{F}_k\right] \leq \left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N + \eta^2\left\|x^{(k)} - x^\infty\right\|^2. \tag{112}$$

Therefore, we can compute that

$$\begin{aligned}
\mathbb{E}\left\|x^{(k+1)}-x^{\infty}\right\|^2 &= \mathbb{E}\left\|x^{(k)}-x^{\infty}-\alpha\nabla F_{\mathcal{W},\alpha}\left(x^{(k)}\right)-\alpha\xi^{(k+1)}\right\|^2 \\
&= \mathbb{E}\left\|x^{(k)}-x^{\infty}-\alpha\nabla F_{\mathcal{W},\alpha}\left(x^{(k)}\right)\right\|^2+\alpha^2\mathbb{E}\left\|\xi^{(k+1)}\right\|^2 \\
&= \mathbb{E}\left\|x^{(k)}-x^{\infty}\right\|^2+\alpha^2\mathbb{E}\left\|\nabla F_{\mathcal{W},\alpha}\left(x^{(k)}\right)\right\|^2 \\
&\qquad -2\alpha\mathbb{E}\left\langle x^{(k)}-x^{\infty},\nabla F_{\mathcal{W},\alpha}\left(x^{(k)}\right)\right\rangle+\alpha^2\mathbb{E}\left\|\xi^{(k+1)}\right\|^2 \\
&= \mathbb{E}\left\|x^{(k)}-x^{\infty}\right\|^2+\alpha^2L_{\alpha}\mathbb{E}\left\langle x^{(k)}-x^{\infty},\nabla F_{\mathcal{W},\alpha}\left(x^{(k)}\right)\right\rangle \\
&\qquad -2\alpha\mathbb{E}\left\langle x^{(k)}-x^{\infty},\nabla F_{\mathcal{W},\alpha}\left(x^{(k)}\right)\right\rangle+\alpha^2\mathbb{E}\left\|\xi^{(k+1)}\right\|^2 \\
&\leq \left(1-2\alpha\mu\left(1-\frac{\alpha L_{\alpha}}{2}\right)\right)\mathbb{E}\left\|x^{(k)}-x^{\infty}\right\|^2+\alpha^2\mathbb{E}\left\|\xi^{(k+1)}\right\|^2,
\end{aligned}$$

where we used the fact that $\mathcal{F}_{\mathcal{W},\alpha}$ is L_{α}-smooth and μ-strongly convex and the assumption $\alpha\leq\frac{\frac{1}{2}+\lambda_N^W}{L+\frac{\eta^2}{\mu}}<\frac{1+\lambda_N^W}{L}$ so that $\alpha L_{\alpha}=1-\lambda_N^W+\alpha L<2$. By applying (112), we get

$$\begin{aligned}
\mathbb{E}\left\|x^{(k+1)}-x^{\infty}\right\|^2 \leq \left(1-2\alpha\mu\left(1-\frac{\alpha L_{\alpha}}{2}-\frac{\alpha\eta^2}{2\mu}\right)\right)\mathbb{E}\left\|x^{(k)}-x^{\infty}\right\|^2 \\
+\alpha^2\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N.
\end{aligned}$$

We recall that assumption $\alpha\leq\frac{\frac{1}{2}+\lambda_N^W}{L+\frac{\eta^2}{\mu}}$ so that $1-\frac{\alpha L_{\alpha}}{2}-\frac{\alpha\eta^2}{2\mu}\geq\frac{1}{4}$. Therefore, we have

$$\mathbb{E}\left\|x^{(k+1)}-x^{\infty}\right\|^2\leq(1-\alpha\mu/2)\,\mathbb{E}\left\|x^{(k)}-x^{\infty}\right\|^2+\alpha^2\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N,$$

which implies that

$$\mathbb{E}\left\|x^{(k)}-x^{\infty}\right\|^2\leq(1-\alpha\mu/2)^k\,\mathbb{E}\left\|x^{(0)}-x^{\infty}\right\|^2+\frac{2\alpha}{\mu}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N.$$

Hence, we conclude that

$$\mathbb{E}\left\|x^{(k)}-x^{*}\right\|^2\leq 2\,(1-\alpha\mu/2)^k\,\mathbb{E}\left\|x^{(0)}-x^{\infty}\right\|^2+\frac{4\alpha}{\mu}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N+\frac{2\alpha^2(C_1)^2N}{(1-\gamma)^2}.$$

■

Next, we will provide the performance bounds for the average iterates and individual iterates. Before we proceed, let us first introduce and prove a few technical lemmas. Let us recall that

$$\mathcal{E}_{k+1}:=\nabla f\left(\bar{x}^{(k)}\right)-\frac{1}{N}\sum_{i=1}^{N}\nabla f_i\left(x_i^{(k)}\right).$$

Next, we will show that the error term $\mathcal{E}_{k+1}$ is small and for every $i = 1, 2, \ldots, N$, $x_i^{(k)}$ is close to the average $\bar{k}^{(k)}$.

Lemma 43 *Assume that* $\alpha \leq \frac{\frac{1}{2}+\lambda_N^W}{L+\frac{\eta^2}{\mu}}$ *and* $\alpha\mu(1+\lambda_N^W - \alpha L) < 1$. *For any* k *and* $i = 1, 2, \ldots, N$, *we have*

$$\mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 \leq \sum_{i=1}^N \mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 \leq 4\gamma^{2k}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{4D_1^2\alpha^2}{(1-\gamma)^2} + \frac{4N\alpha^2}{(1-\gamma^2)}D_2^2, \quad (113)$$

and for any k, *we have*

$$\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 \leq \frac{4L^2\gamma^{2k}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{4L^2D_1^2\alpha^2}{N(1-\gamma)^2} + \frac{4L^2\alpha^2}{(1-\gamma^2)}D_2^2,$$

where

$$D_1^2 := L^2\mathbb{E}\left\|x^{(0)} - x^\infty\right\|^2 + L^2\frac{2\alpha}{\mu}\left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N, \quad (114)$$

$$D_2^2 := \left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)\frac{\mu+2\alpha}{\mu} + \frac{\eta^2}{N}\mathbb{E}\left\|x^{(0)} - x^\infty\right\|^2. \quad (115)$$

Proof The proof of Lemma 43 will be provided in Appendix F. ■

Let us define x_k as the iterates of the decentralized algorithm:

$$x_{k+1} = x_k - \alpha\nabla f\left(x_k\right) - \alpha\bar{\xi}^{(k+1)},$$

with $x_0 = \bar{x}^{(0)}$, where we recall that

$$\bar{\xi}^{(k+1)} := \frac{1}{N}\sum_{i=1}^N\left(\tilde{\nabla}f_i\left(x_i^{(k)}\right) - \nabla f_i\left(x_i^{(k)}\right)\right),$$

so that we have

$$\mathbb{E}\left[\bar{\xi}^{(k+1)}\middle|\mathcal{F}_k\right] = 0, \qquad \mathbb{E}\left\|\bar{\xi}^{(k+1)}\middle|\mathcal{F}_k\right\|^2 \leq \frac{\sigma^2}{N} + \frac{\eta^2}{2N^2}\left\|x^{(k)} - x^*\right\|^2. \quad (116)$$

Next, we will show that x_k and the average iterates $\bar{x}^{(k)}$ are close to each other in the L^2 norm.

Lemma 44 *Assume that* $\alpha \leq \frac{\frac{1}{2}+\lambda_N^W}{L+\frac{\eta^2}{\mu}}$ *and* $\alpha\mu(1+\lambda_N^W - \alpha L) < 1$. *For any* k, *we have*

$$\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 \leq \alpha\left(\frac{\alpha}{\mu(1-\frac{\alpha L}{2})} + \frac{(1+\alpha L)^2}{\mu^2(1-\frac{\alpha L}{2})^2}\right)\left(\frac{4L^2D_1^2\alpha}{N(1-\gamma)^2} + \frac{4L^2\alpha}{(1-\gamma^2)}D_2^2\right)$$
$$+ \frac{\gamma^{2k} - \left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)^k}{\gamma^2 - 1 + \alpha\mu\left(1-\frac{\alpha L}{2}\right)}\frac{4L^2\gamma^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2,$$

where D_1, D_2 *are defined in Lemma 43.*

Proof The proof is similar to that of Lemma 24 and is hence omitted here. ■

Finally, we are ready to present the performance bounds for the average iterates and individual iterates.

Proposition 45 *Assume that* $\alpha \leq \frac{\frac{1}{2}+\lambda_N^W}{L+\frac{\eta^2}{\mu}}$ *and* $\alpha\mu(1+\lambda_N^W-\alpha L)<1$. *For any* $k \geq 0$,

$$\begin{aligned}\mathbb{E}\left\|\bar{x}^{(k)}-x_*\right\|^2 \leq\ & 2(1-\alpha\mu)^k\mathbb{E}\left\|\bar{x}^{(0)}-x_*\right\|^2+\frac{2\alpha}{\mu}\frac{\sigma^2}{N}\\ &+\frac{2\alpha\eta^2}{\mu N^2}\left(\mathbb{E}\left\|x^{(0)}-x^\infty\right\|^2+\frac{2\alpha}{\mu}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N+\frac{\alpha^2(C_1)^2N}{(1-\gamma)^2}\right)\\ &+\alpha\left(\frac{\alpha}{\mu(1-\frac{\alpha L}{2})}+\frac{(1+\alpha L)^2}{\mu^2(1-\frac{\alpha L}{2})^2}\right)\left(\frac{8L^2D_1^2\alpha}{N(1-\gamma)^2}+\frac{8L^2\alpha}{(1-\gamma^2)}D_2^2\right)\\ &+\frac{\gamma^{2k}-\left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)^k}{\gamma^2-1+\alpha\mu\left(1-\frac{\alpha L}{2}\right)}\frac{8L^2\gamma^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2,\end{aligned}$$

and for any $k \geq 0$, $i=1,2,\ldots,N$,

$$\begin{aligned}\mathbb{E}\left\|x_i^{(k)}-x_*\right\|^2 \leq\ & 4(1-\alpha\mu)^k\mathbb{E}\left\|\bar{x}^{(0)}-x_*\right\|^2+\frac{4\alpha}{\mu}\frac{\sigma^2}{N}\\ &+\frac{4\alpha\eta^2}{\mu N^2}\left(\mathbb{E}\left\|x^{(0)}-x^\infty\right\|^2+\frac{2\alpha}{\mu}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N+\frac{\alpha^2(C_1)^2N}{(1-\gamma)^2}\right)\\ &+\alpha\left(\frac{\alpha}{\mu(1-\frac{\alpha L}{2})}+\frac{(1+\alpha L)^2}{\mu^2(1-\frac{\alpha L}{2})^2}\right)\left(\frac{16L^2D_1^2\alpha}{N(1-\gamma)^2}+\frac{16L^2\alpha}{(1-\gamma^2)}D_2^2\right)\\ &+\frac{\gamma^{2k}-\left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)^k}{\gamma^2-1+\alpha\mu\left(1-\frac{\alpha L}{2}\right)}\frac{16L^2\gamma^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2\\ &+8\gamma^{2k}\mathbb{E}\left\|x^{(0)}\right\|^2+\frac{8D_1^2\alpha^2}{(1-\gamma)^2}+\frac{8N\alpha^2}{(1-\gamma^2)}D_2^2,\end{aligned}$$

where D_1, D_2 *are defined in Lemma 43.*

Proof By following the proof of Theorem 42, we get for any $\alpha \leq \frac{1}{L}$,

$$\begin{aligned}\mathbb{E}\left\|x_{k+1}-x_*\right\|^2 &\leq \left(1-2\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)\mathbb{E}\left\|x_k-x_*\right\|^2+\alpha^2\mathbb{E}\left\|\bar{\xi}^{(k+1)}\right\|^2\\ &\leq (1-\alpha\mu)\,\mathbb{E}\left\|x_k-x_*\right\|^2+\alpha^2\mathbb{E}\left\|\bar{\xi}^{(k+1)}\right\|^2.\end{aligned}$$

By (116) and Theorem 42, we get

$$\begin{aligned}\mathbb{E}\left\|\bar{\xi}^{(k+1)}\right\|^2 &\leq \frac{\sigma^2}{N}+\frac{\eta^2}{2N^2}\mathbb{E}\left\|x^{(k)}-x^*\right\|^2\\ &\leq \frac{\sigma^2}{N}+\frac{\eta^2}{2N^2}\left(2\mathbb{E}\left\|x^{(0)}-x^\infty\right\|^2+\frac{4\alpha}{\mu}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N+\frac{2\alpha^2(C_1)^2N}{(1-\gamma)^2}\right).\end{aligned}$$

Therefore, we obtain

$$\mathbb{E}\left\|x_k - x_*\right\|^2 \leq (1-\alpha\mu)^k \mathbb{E}\left\|\bar{x}^{(0)} - x_*\right\|^2 + \frac{\alpha}{\mu}\frac{\sigma^2}{N} + \frac{\alpha\eta^2}{\mu N^2}\left(\mathbb{E}\left\|x^{(0)} - x^\infty\right\|^2 + \frac{2\alpha}{\mu}\left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N + \frac{\alpha^2(C_1)^2 N}{(1-\gamma)^2}\right).$$

Finally, by applying

$$\mathbb{E}\left\|\bar{x}^{(k)} - x_*\right\|^2 \leq 2\mathbb{E}\left\|x_k - x_*\right\|^2 + 2\mathbb{E}\left\|x_k - \bar{x}^{(k)}\right\|^2, \tag{117}$$

and for every $i = 1, 2, \ldots, N$,

$$\begin{aligned}\mathbb{E}\left\|x_i^{(k)} - x_*\right\|^2 &\leq 2\mathbb{E}\left\|\bar{x}^{(k)} - x_*\right\|^2 + 2\mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 \\ &\leq 4\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 + 4\mathbb{E}\left\|x_k - x_*\right\|^2 + 2\mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2,\end{aligned}$$

and by applying Lemma 44, we complete the proof. ■

E.2 Distributed Accelerated Stochastic Gradient (D-ASG)

Let us recall the D-ASG (12). Define

$$\bar{\xi}^{(k+1)} := \frac{1}{N}\sum_{i=1}^N \left(\tilde{\nabla} f_i\left(y_i^{(k)}\right) - \nabla f_i\left(y_i^{(k)}\right)\right),$$

so that by Assumption 5, we have

$$\mathbb{E}\left[\bar{\xi}^{(k+1)}\middle|\mathcal{F}_k\right] = 0, \qquad \mathbb{E}\left\|\bar{\xi}^{(k+1)}\middle|\mathcal{F}_k\right\|^2 \leq \frac{\sigma^2}{N} + \frac{\eta^2}{2N^2}\left\|y^{(k)} - x^*\right\|^2. \tag{118}$$

Let us define

$$V_{Q,\alpha}(\xi) := \xi^T Q_\alpha \xi + F_{\mathcal{W},\alpha}(T\xi + x^\infty) - F_{\mathcal{W},\alpha}(x^\infty), \tag{119}$$

where $F_{\mathcal{W},\alpha}$ is defined in (15) and $Q_\alpha := \tilde{Q}_\alpha \otimes I_{Nd}$ with

$$\tilde{Q}_\alpha := \begin{bmatrix} \sqrt{\frac{1}{2\alpha}} \\ \sqrt{\frac{\mu}{2}} - \sqrt{\frac{1}{2\alpha}} \end{bmatrix}\begin{bmatrix} \sqrt{\frac{1}{2\alpha}} & \sqrt{\frac{\mu}{2}} - \sqrt{\frac{1}{2\alpha}} \end{bmatrix} + 2\alpha\eta^2 \begin{bmatrix} 1+\beta \\ -\beta \end{bmatrix}\begin{bmatrix} 1+\beta & -\beta \end{bmatrix}. \tag{120}$$

We have the following explicit performance bounds on the convergence and the robustness of D-ASG iterates.

Theorem 46 *Assume $\kappa \geq 4$. Consider running D-ASG method with $\alpha \in (0, \hat{\alpha}]$ and $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$ with*

$$\hat{\alpha} := \begin{cases} \min\left\{\frac{\lambda_N^W}{L}, \frac{\mu^3}{(60\eta^2)^2}\right\} & \text{if } \eta > 0, \\ \frac{\lambda_N^W}{L} & \text{if } \eta = 0. \end{cases} \tag{121}$$

Then, for any $k \geq 0$, we have

$$\mathbb{E}\left[\left\|x^{(k)} - x^{\infty}\right\|^2\right] \leq 2\left(1 - \sqrt{\alpha\mu}/3\right)^k \frac{V_{Q,\alpha}\left(\xi_0\right)}{\mu} + \frac{12\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N. \tag{122}$$

In addition, if $\alpha \leq \frac{1}{L+\mu}$, we have

$$\mathbb{E}\left[\left\|x^{(k)} - x^{*}\right\|^2\right] \leq 4\left(1 - \sqrt{\alpha\mu}/3\right)^k \frac{V_{Q,\alpha}\left(\xi_0\right)}{\mu} + \frac{24\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N + \frac{2C_1^2N\alpha^2}{(1-\gamma)^2}, \tag{123}$$

where C_1, γ are given in (8).

Proof D-ASG reduces to the iterations (18) which are equivalent to applying non-distributed ASG to minimize the function $F_{\mathcal{W},\alpha} \in S_{\mu,L_\alpha}(\mathbb{R}^{Nd})$, where $F_{\mathcal{W},\alpha}$ is defined in (15) and $L_\alpha = \frac{1-\lambda_N^W}{\alpha} + L$. Therefore, applying Theorem K.1 in Aybat et al. (2019) from the literature for non-distributed ASG, for any $\alpha \in (0, \bar{\alpha}]$ and $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$ with

$$\bar{\alpha} := \begin{cases} \min\left\{\frac{1}{L_\alpha}, \frac{\mu^3}{(60\eta^2)^2}\right\} & \text{if } \eta > 0, \\ \frac{1}{L_\alpha} & \text{if } \eta = 0, \end{cases} \tag{124}$$

where $L_\alpha = \frac{1-\lambda_N^W}{\alpha} + L$, we obtain

$$\mathbb{E}\left[V_{Q,\alpha}\left(\xi_{k+1}\right)\right] \leq \left(1 - \sqrt{\alpha\mu}/3\right)\mathbb{E}V_{Q,\alpha}\left(\xi_k\right) + 2\alpha\left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N, \tag{125}$$

which yields

$$\mathbb{E}\left[V_{Q,\alpha}\left(\xi_k\right)\right] \leq \left(1 - \sqrt{\alpha\mu}/3\right)^k V_{Q,\alpha}\left(\xi_0\right) + \frac{6\sqrt{\alpha}}{\sqrt{\mu}}\left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N.$$

Note that the condition $\alpha \in (0, \bar{\alpha}]$ is equivalent to $\alpha \in (0, \hat{\alpha}]$, where

$$\hat{\alpha} := \begin{cases} \min\left\{\frac{\lambda_N^W}{L}, \frac{\mu^3}{(60\eta^2)^2}\right\} & \text{if } \eta > 0, \\ \frac{\lambda_N^W}{L} & \text{if } \eta = 0. \end{cases} \tag{126}$$

The rest of the proof is similar to that of Theorem 7 and is omitted here. ■

Next, we show that the individual iterates and the average iterates are close.

Lemma 47 *Consider running D-ASG method under the assumptions in Theorem 46. For any k and $i = 1, \ldots, N$, we have*

$$\begin{aligned}\mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 &\leq \sum_{i=1}^N \mathbb{E}\left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 \\ &\leq 8\gamma^{2k}\left(4\frac{V_{Q,\alpha}\left(\xi_0\right)}{\mu} + \frac{24\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N + \frac{2C_1^2N\alpha^2}{(1-\gamma)^2} + \|x^*\|^2\right) \\ &\quad + \frac{4\tilde{D}_y^2\alpha^2}{(1-\gamma)^2} + \frac{4\tilde{E}_y^2\alpha^2}{(1-\gamma^2)} + \frac{8\tilde{C}_0\alpha}{(1-\gamma)^2},\end{aligned}$$

and for any k, we have

$$\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 \leq \frac{2}{N}L^2\left((1+\beta)^2+\beta^2\right)\Bigg[\frac{4\tilde{D}_y^2\alpha^2}{(1-\gamma)^2}+\frac{4\tilde{E}_y^2\alpha^2}{(1-\gamma^2)}+\frac{8\tilde{C}_0\alpha}{(1-\gamma)^2} + 8\gamma^{2(k-1)}\left(4\frac{V_{Q,\alpha}\left(\xi_0\right)}{\mu}+\frac{24\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N+\frac{2C_1^2N\alpha^2}{(1-\gamma)^2}+\|x^*\|^2\right)\Bigg],$$

where $\tilde{C}_0$ is defined in (132)*, $\tilde{D}_y$ is defined in* (130) *and $\tilde{E}_y$ is defined in* (131)*.*

Proof The proof of Lemma 47 will be provided in Appendix F. ■

Finally, we provide the performance bounds for the average iterates and individual iterates. We first define

$$V_{\bar{Q},\alpha}(\bar{\xi}) := \bar{\xi}^T\bar{Q}_\alpha\bar{\xi}+f\left(T\bar{\xi}+x_*\right)-f(x_*), \tag{127}$$

where $\bar{Q}_\alpha := \tilde{Q}_\alpha \otimes I_d$ and $\tilde{Q}_\alpha$ is defined in (120). Before we proceed, let us first prove a technical lemma.

Lemma 48 *Consider running D-ASG method under the assumptions in Theorem 46. For any $\epsilon > 0$, there exists $M_\epsilon > 0$ such that*

$$\mathbb{E}\left[V_{\bar{Q},\alpha}\left(\bar{\xi}_{k+1}\right)\right] \leq (1+\epsilon)\mathbb{E}\left[V_{\bar{Q},\alpha}\left(\bar{\xi}_{k+1}-D_{k+1}\right)\right]+M_\epsilon\mathbb{E}\|D_{k+1}\|^2, \tag{128}$$

where

$$M_\epsilon := \frac{\max(4/m_2, L^2/\mu)}{2\epsilon}+\frac{L}{2}+\frac{1}{\alpha}+\frac{\mu}{2}-\frac{\sqrt{\mu}}{\sqrt{\alpha}}+2\alpha\eta^2\left((1+\beta)^2+\beta^2\right), \tag{129}$$

where $m_2 > 0$ is the smallest eigenvalue of $\bar{Q}_\alpha$.

Proof The proof of Lemma 48 will be provided in Appendix F. ■

Finally, we are ready to present the performance bounds for the average iterates and individual iterates.

Proposition 49 *Consider running D-ASG method under the assumptions in Theorem 46. For any k, we have*

$$\begin{aligned}&\mathbb{E}\left\|\bar{x}^{(k)}-x_*\right\|^2\\ &\leq\left(1-\frac{\sqrt{\alpha\mu}}{6}\right)^k\frac{2V_{\bar{Q},\alpha}\left(\bar{\xi}_0\right)}{\mu}+\frac{12}{\mu\sqrt{\mu}}\left(\left(1+\frac{\sqrt{\alpha\mu}}{6}\right)\frac{2\sqrt{\alpha}}{N}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)+\frac{\alpha\tilde{H}_1\tilde{H}_2}{2\sqrt{\mu}}\right)\\ &\quad+\frac{8}{\gamma^2\mu\sqrt{\mu}}\sqrt{\alpha}\tilde{H}_1\tilde{H}_3\frac{\gamma^{2k}-(1-\sqrt{\alpha\mu}/6)^k}{\gamma^2-(1-\sqrt{\alpha\mu}/6)},\end{aligned}$$

and for every $i = 1, 2, \ldots, N$ *and any* k,

$$\begin{aligned}
&\mathbb{E}\left\|x_i^{(k)} - x_*\right\|^2 \\
&\leq \left(1 - \frac{\sqrt{\alpha\mu}}{6}\right)^k \frac{4V_{\bar{Q},\alpha}\left(\bar{\xi}_0\right)}{\mu} + \frac{24}{\mu\sqrt{\mu}}\left(\left(1 + \frac{\sqrt{\alpha\mu}}{6}\right)\frac{2\sqrt{\alpha}}{N}\left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right) + \frac{\alpha\tilde{H}_1\tilde{H}_2}{2\sqrt{\mu}}\right) \\
&\quad + \frac{16}{\gamma^2\mu\sqrt{\mu}}\sqrt{\alpha}\tilde{H}_1\tilde{H}_3\frac{\gamma^{2k} - (1 - \sqrt{\alpha\mu}/6)^k}{\gamma^2 - (1 - \sqrt{\alpha\mu}/6)} \\
&\quad + 16\gamma^{2k}\left(4\frac{V_{Q,\alpha}\left(\xi_0\right)}{\mu} + \frac{24\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N + \frac{2C_1^2N\alpha^2}{(1-\gamma)^2} + \|x^*\|^2\right) \\
&\quad + \frac{8\tilde{D}_y^2\alpha^2}{(1-\gamma)^2} + \frac{8\tilde{E}_y^2\alpha^2}{(1-\gamma^2)} + \frac{16\tilde{C}_0\alpha}{(1-\gamma)^2},
\end{aligned}$$

where $\tilde{C}_0$ *is some constant such that* $\tilde{C}_0 = \mathcal{O}(1)$ *as* $\alpha \to 0$, *and*

$$\tilde{D}_y^2 := 4L^2\left((1+\beta)^2 + \beta^2\right)\left(\frac{2V_{Q,\alpha}\left(\xi_0\right)}{\mu} + \frac{12\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N\right) + 2\|\nabla F(x^*)\|^2, \tag{130}$$

$$\begin{aligned}
\tilde{E}_y^2 := &\left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N \\
&+ 2\eta^2\left((1+\beta)^2 + \beta^2\right)\left(\frac{2V_{Q,\alpha}\left(\xi_0\right)}{\mu} + \frac{12\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N\right),
\end{aligned} \tag{131}$$

and

$$\begin{aligned}
\tilde{H}_1 &:= 6\alpha\max(4/m_2, L^2/\mu) + (L\alpha + 2 + \mu\alpha)\sqrt{\alpha\mu} - 2\mu\alpha + 4\alpha^2\sqrt{\alpha\mu}\eta^2\left((1+\beta)^2 + \beta^2\right), \\
\tilde{H}_2 &:= \frac{2}{N}L^2\left((1+\beta)^2 + \beta^2\right)\left(\frac{4\tilde{D}_y^2\alpha}{(1-\gamma)^2} + \frac{4\tilde{E}_y^2\alpha}{(1-\gamma)^2} + \frac{8\tilde{C}_0}{(1-\gamma)^2}\right), \\
\tilde{H}_3 &:= \frac{2}{N}L^2\left((1+\beta)^2 + \beta^2\right) \\
&\quad \cdot \left(4\frac{V_{Q,\alpha}\left(\xi_0\right)}{\mu} + \frac{24\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(\sigma^2 + \eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N + \frac{2C_1^2N\alpha^2}{(1-\gamma)^2} + \|x^*\|^2\right),
\end{aligned}$$

where $m_2 > 0$ *is the smallest eigenvalue of* $\bar{Q}_\alpha$.

Proof The proof is similar to the proof of Proposition 10. Similar to Lemma 29, we can show that

$$\sup_k \mathbb{E}\left\|x^{(k)} - x^{(k-1)}\right\|^2 \leq \frac{2\tilde{C}_0}{\beta^2}\alpha, \tag{132}$$

for some $\tilde{C}_0$ such that $\tilde{C}_0 = \mathcal{O}(1)$ as $\alpha \to 0$. By applying (132) and Theorem K.1 in Aybat et al. (2019) from the literature for non-distributed ASG, for any $\alpha \in (0, \bar{\alpha}]$ and $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$,

as well as Lemma 48, we have

$$\begin{aligned}
&\mathbb{E}\left[V_{\bar{Q},\alpha}\left(\bar{\xi}_{k+1}\right)\right] \\
&\leq (1+\epsilon)\mathbb{E}\left[V_{\bar{Q},\alpha}\left(\bar{\xi}_{k+1}-D_{k+1}\right)\right]+M_\epsilon\mathbb{E}\|D_{k+1}\|^2 \\
&\leq (1+\epsilon)\left((1-\sqrt{\alpha\mu}/3)\,\mathbb{E}V_{\bar{Q},\alpha}\left(\bar{\xi}_k\right)+\frac{2\alpha}{N}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)\right)+M_\epsilon\mathbb{E}\|D_{k+1}\|^2.
\end{aligned}$$

Let us take $\epsilon=\frac{1}{6}\sqrt{\alpha\mu}$, then we get

$$\begin{aligned}
&\mathbb{E}\left[V_{\bar{Q},\alpha}\left(\bar{\xi}_{k+1}\right)\right] \\
&\leq \left(1-\frac{\sqrt{\alpha\mu}}{6}\right)\mathbb{E}V_{\bar{Q},\alpha}\left(\bar{\xi}_k\right)+\left(1+\frac{\sqrt{\alpha\mu}}{6}\right)\frac{2\alpha}{N}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)+M_{\frac{\sqrt{\alpha\mu}}{6}}\mathbb{E}\|D_{k+1}\|^2,
\end{aligned}$$

and by Lemma 48, we have

$$\begin{aligned}
M_{\frac{\sqrt{\alpha\mu}}{6}} &= \frac{3\max(4/m_2,L^2/\mu)}{\sqrt{\alpha\mu}}+\frac{L}{2}+\frac{1}{\alpha}+\frac{\mu}{2}-\frac{\sqrt{\mu}}{\sqrt{\alpha}}+2\alpha\eta^2\left((1+\beta)^2+\beta^2\right) \\
&\leq \frac{1}{2\alpha\sqrt{\alpha\mu}}\tilde{H}_1,
\end{aligned}$$

where

$$\tilde{H}_1=6\alpha\max(4/m_2,L^2/\mu)+(L\alpha+2+\mu\alpha)\sqrt{\alpha\mu}-2\mu\alpha+4\alpha^2\sqrt{\alpha\mu}\eta^2\left((1+\beta)^2+\beta^2\right). \quad (133)$$

By Lemma 47, we have

$$\mathbb{E}\|D_{k+1}\|^2\leq\alpha^2\left[\alpha\tilde{H}_2+8\gamma^{2(k-1)}\tilde{H}_3\right],$$

where

$$\begin{aligned}
\tilde{H}_2 &= \frac{2}{N}L^2\left((1+\beta)^2+\beta^2\right)\left(\frac{4\tilde{D}_y^2\alpha}{(1-\gamma)^2}+\frac{4\tilde{E}_y^2\alpha}{(1-\gamma)^2}+\frac{8\tilde{C}_0}{(1-\gamma)^2}\right), \\
\tilde{H}_3 &= \frac{2}{N}L^2\left((1+\beta)^2+\beta^2\right) \\
&\quad\cdot\left(4\frac{V_{Q,\alpha}\left(\xi_0\right)}{\mu}+\frac{24\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N+\frac{2C_1^2N\alpha^2}{(1-\gamma)^2}+\|x^*\|^2\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[V_{\bar{Q},\alpha}\left(\bar{\xi}_{k+1}\right)\right] &\leq \left(1-\frac{\sqrt{\alpha\mu}}{6}\right)\mathbb{E}V_{\bar{Q},\alpha}\left(\bar{\xi}_k\right)+\left(1+\frac{\sqrt{\alpha\mu}}{6}\right)\frac{2\alpha}{N}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right) \\
&\quad+\frac{1}{2\sqrt{\mu}}\alpha^2\tilde{H}_1\tilde{H}_2+\frac{4}{\gamma^2\sqrt{\mu}}\sqrt{\alpha}\tilde{H}_1\tilde{H}_3\gamma^{2k},
\end{aligned}$$

which by following the similar argument in the proof of Proposition 10 implies that

$$\begin{aligned}
\mathbb{E}\left[V_{\bar{Q},\alpha}\left(\bar{\xi}_k\right)\right] &\leq \left(1-\frac{\sqrt{\alpha\mu}}{6}\right)^k V_{\bar{Q},\alpha}\left(\bar{\xi}_0\right)+\frac{4}{\gamma^2\sqrt{\mu}}\sqrt{\alpha}\tilde{H}_1\tilde{H}_3\frac{\gamma^{2k}-(1-\sqrt{\alpha\mu}/6)^k}{\gamma^2-(1-\sqrt{\alpha\mu}/6)} \\
&\quad+\frac{6}{\sqrt{\alpha\mu}}\left(\left(1+\frac{\sqrt{\alpha\mu}}{6}\right)\frac{2\alpha}{N}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)+\frac{1}{2\sqrt{\mu}}\alpha\sqrt{\alpha}\tilde{H}_1\tilde{H}_2\right).
\end{aligned}$$

The rest of the proof is similar to that of Proposition 10. ■

Appendix F. Proofs of Technical Lemmas

F.1 Proofs of Technical Results in Appendix B

F.1.1 Proof of Lemma 24

Proof We can compute that

$$\bar{x}^{(k+1)} - x_{k+1} = \bar{x}^{(k)} - x_k - \alpha \left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right] + \alpha \mathcal{E}_{k+1}, \tag{134}$$

where $\mathcal{E}_{k+1} := \nabla f\left(\bar{x}^{(k)}\right) - \frac{1}{N}\sum_{i=1}^{N} \nabla f_i\left(x_i^{(k)}\right)$.

If the term $\mathcal{E}_{k+1}$ were not present in the recursion (134), we could rely on standard analysis techniques for analyzing a gradient step in order to bound $\left\|\bar{x}^{(k+1)} - x_{k+1}\right\|^2$ with $\left\|\bar{x}^{(k)} - x_k\right\|^2$. However, in the presence of $\mathcal{E}_{k+1}$, we need to control this error term based on Lemma 23. To be more precise, we have

$$\begin{aligned}
&\left\|\bar{x}^{(k+1)} - x_{k+1}\right\|^2 \\
&= \left\|\bar{x}^{(k)} - x_k - \alpha \left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right]\right\|^2 + \alpha^2 \left\|\mathcal{E}_{k+1}\right\|^2 \\
&\qquad + 2\left\langle \bar{x}^{(k)} - x_k - \alpha \left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right], \alpha \mathcal{E}_{k+1}\right\rangle \\
&= \left\|\bar{x}^{(k)} - x_k\right\|^2 + \alpha^2 \left\|\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right]\right\|^2 \\
&\quad - 2\left\langle \bar{x}^{(k)} - x_k, \alpha \left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right]\right\rangle + \alpha^2 \left\|\mathcal{E}_{k+1}\right\|^2 \\
&\qquad + 2\left\langle \bar{x}^{(k)} - x_k - \alpha \left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right], \alpha \mathcal{E}_{k+1}\right\rangle \\
&\le \left\|\bar{x}^{(k)} - x_k\right\|^2 + \alpha^2 L \left\langle \bar{x}^{(k)} - x_k, \left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right]\right\rangle \\
&\quad - 2\left\langle \bar{x}^{(k)} - x_k, \alpha \left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right]\right\rangle + \alpha^2 \left\|\mathcal{E}_{k+1}\right\|^2 \\
&\qquad + 2\left\langle \bar{x}^{(k)} - x_k - \alpha \left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right], \alpha \mathcal{E}_{k+1}\right\rangle \\
&= \left\|\bar{x}^{(k)} - x_k\right\|^2 - 2\alpha\left(1 - \frac{\alpha L}{2}\right)\left\langle \bar{x}^{(k)} - x_k, \left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right]\right\rangle \\
&\qquad + \alpha^2 \left\|\mathcal{E}_{k+1}\right\|^2 + 2\left\langle \bar{x}^{(k)} - x_k - \alpha \left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right], \alpha \mathcal{E}_{k+1}\right\rangle \\
&\le \left(1 - 2\alpha\mu\left(1 - \frac{\alpha L}{2}\right)\right)\left\|\bar{x}^{(k)} - x_k\right\|^2 + \alpha^2 \left\|\mathcal{E}_{k+1}\right\|^2 \\
&\qquad + 2\left\langle \bar{x}^{(k)} - x_k - \alpha \left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right], \alpha \mathcal{E}_{k+1}\right\rangle,
\end{aligned} \tag{135}$$

where we used Nesterov (2003, Theorem 2.1.5) on L-smooth and convex functions to obtain the second term after the first inequality above and μ-strong convexity of f and the

assumption that $\alpha < 2/L$ to obtain the first term after the second inequality above. By taking expectations in (135), we get

$$\begin{aligned}
&\mathbb{E}\left\|\bar{x}^{(k+1)} - x_{k+1}\right\|^2 \\
&\leq \left(1 - 2\alpha\mu\left(1 - \frac{\alpha L}{2}\right)\right)\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 + \alpha^2\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 \\
&\qquad + \mathbb{E}\left[2\left\langle \bar{x}^{(k)} - x_k - \alpha\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right], \alpha\mathcal{E}_{k+1}\right\rangle\right] \\
&= \left(1 - 2\alpha\mu\left(1 - \frac{\alpha L}{2}\right)\right)\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 + \alpha^2\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 \\
&\qquad + \mathbb{E}\left[2\left\langle \bar{x}^{(k)} - x_k - \alpha\left[\nabla f\left(\bar{x}^{(k)}\right) - \nabla f(x_k)\right], \alpha\mathcal{E}_{k+1}\right\rangle\right] \\
&\leq \left(1 - 2\alpha\mu\left(1 - \frac{\alpha L}{2}\right)\right)\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 + \alpha^2\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 \\
&\qquad + 2(1+\alpha L)\alpha\mathbb{E}\left[\left\|\bar{x}^{(k)} - x_k\right\| \cdot \left\|\mathcal{E}_{k+1}\right\|\right],
\end{aligned}$$

where we used L-smoothness of f.

For any $x, y \geq 0$ and $c > 0$, we have the inequality $2xy \leq cx^2 + \frac{y^2}{c}$, which implies that

$$\begin{aligned}
&\mathbb{E}\left\|\bar{x}^{(k+1)} - x_{k+1}\right\|^2 \\
&\leq \left(1 - 2\alpha\mu\left(1 - \frac{\alpha L}{2}\right)\right)\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 + \alpha^2\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2 \\
&\qquad + (1+\alpha L)\alpha\left(\frac{\mu(1-\frac{\alpha L}{2})}{1+\alpha L}\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 + \frac{1+\alpha L}{\mu(1-\frac{\alpha L}{2})}\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2\right) \\
&= \left(1 - \alpha\mu\left(1 - \frac{\alpha L}{2}\right)\right)\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 + \alpha\left(\alpha + \frac{(1+\alpha L)^2}{\mu(1-\frac{\alpha L}{2})}\right)\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2.
\end{aligned}$$

By applying Lemma 23, we get

$$\begin{aligned}
&\mathbb{E}\left\|\bar{x}^{(k+1)} - x_{k+1}\right\|^2 \\
&\leq \left(1 - \alpha\mu\left(1 - \frac{\alpha L}{2}\right)\right)\mathbb{E}\left\|\bar{x}^{(k)} - x_k\right\|^2 \\
&\quad + \alpha\left(\alpha + \frac{(1+\alpha L)^2}{\mu(1-\frac{\alpha L}{2})}\right)\left(\frac{4L^2\gamma^{2k}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2 + \frac{4L^2D^2\alpha^2}{N(1-\gamma)^2} + \frac{4L^2\sigma^2\alpha^2}{(1-\gamma^2)}\right),
\end{aligned}$$

for every k. Note that $\mathbb{E}\left\|\bar{x}^{(0)}-x_0\right\|^2=0$. By iterating the above equation, we get

$$\begin{aligned}
&\mathbb{E}\left\|\bar{x}^{(k)}-x_k\right\|^2\\
&\leq\sum_{i=0}^{k-1}\left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)^i\cdot\alpha\left(\alpha+\frac{(1+\alpha L)^2}{\mu(1-\frac{\alpha L}{2})}\right)\left(\frac{4L^2D^2\alpha^2}{N(1-\gamma)^2}+\frac{4L^2\sigma^2\alpha^2}{(1-\gamma^2)}\right)\\
&\quad+\sum_{i=0}^{k-1}\left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)^i\alpha\left(\alpha+\frac{(1+\alpha L)^2}{\mu(1-\frac{\alpha L}{2})}\right)\frac{4L^2\gamma^{2(k-i)}}{N}\mathbb{E}\left\|x^{(0)}\right\|^2\\
&=\frac{1-\left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)^k}{1-\left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)}\cdot\alpha\left(\alpha+\frac{(1+\alpha L)^2}{\mu(1-\frac{\alpha L}{2})}\right)\left(\frac{4L^2D^2\alpha^2}{N(1-\gamma)^2}+\frac{4L^2\sigma^2\alpha^2}{(1-\gamma^2)}\right)\\
&\quad+\frac{\gamma^{2k}-\left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)^k}{1-\left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)(\gamma)^{-2}}\frac{4L^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2.
\end{aligned}$$

By our assumption on stepsize α, we have $1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\in[0,1)$. Hence, we conclude that for every k,

$$\begin{aligned}
\mathbb{E}\left\|\bar{x}^{(k)}-x_k\right\|^2&\leq\frac{\alpha\left(\alpha+\frac{(1+\alpha L)^2}{\mu(1-\frac{\alpha L}{2})}\right)\left(\frac{4L^2D^2\alpha^2}{N(1-\gamma)^2}+\frac{4L^2\sigma^2\alpha^2}{(1-\gamma^2)}\right)}{1-\left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)}\\
&\quad+\frac{\gamma^{2k}-\left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)^k}{1-\left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)(\gamma)^{-2}}\frac{4L^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2\\
&=\frac{\alpha\left(\alpha+\frac{(1+\alpha L)^2}{\mu(1-\frac{\alpha L}{2})}\right)\left(\frac{4L^2D^2\alpha}{N(1-\gamma)^2}+\frac{4L^2\sigma^2\alpha}{(1-\gamma^2)}\right)}{\mu\left(1-\frac{\alpha L}{2}\right)}\\
&\quad+\frac{\gamma^{2k}-\left(1-\alpha\mu\left(1-\frac{\alpha L}{2}\right)\right)^k}{\gamma^2-1+\alpha\mu\left(1-\frac{\alpha L}{2}\right)}\frac{4L^2\gamma^2}{N}\mathbb{E}\left\|x^{(0)}\right\|^2.
\end{aligned}$$

The proof is complete. ■

F.1.2 Proof of Lemma 27

Proof By the definition of $x^{(k)}$ and $y^{(k)}$, we get

$$\begin{aligned}
x^{(k+1)}&=(W\otimes I_d)y^{(k)}-\alpha\nabla F\left(y^{(k)}\right)-\alpha\xi^{(k+1)},\\
y^{(k)}&=(1+\beta)x^{(k)}-\beta x^{(k-1)},
\end{aligned}$$

which implies that

$$x^{(k+1)}=(W\otimes I_d)x^{(k)}+(W\otimes I_d)\beta\left(x^{(k)}-x^{(k-1)}\right)-\alpha\nabla F\left(y^{(k)}\right)-\alpha\xi^{(k+1)}.$$

It follows that

$$x^{(k)} = \left(W^k \otimes I_d\right) x^{(0)} - \alpha \sum_{s=0}^{k-1} \left(W^{k-1-s} \otimes I_d\right) \nabla F\left(y^{(s)}\right) - \alpha \sum_{s=0}^{k-1} \left(W^{k-1-s} \otimes I_d\right) \xi^{(s+1)} + \sum_{s=0}^{k-1} \left(W^{k-s} \otimes I_d\right) \beta \left(x^{(s)} - x^{(s-1)}\right). \quad (136)$$

Let us define

$$\bar{\mathbf{x}}^{(k)} := \left[\left(\bar{x}^{(k)}\right)^T, \cdots, \left(\bar{x}^{(k)}\right)^T\right]^T \in \mathbb{R}^{Nd}, \quad (137)$$

and equivalently

$$\bar{\mathbf{x}}^{(k)} = \frac{1}{N}\left(\left(1_N 1_N^T\right) \otimes I_d\right) x^{(k)}, \quad (138)$$

where $1_N \in \mathbb{R}^N$ is a vector of ones; i.e. it is a column vector with all entries equal to one and the superscript ^{T} denotes the vector transpose.

Therefore, we get from (137) and (138) that

$$\sum_{i=1}^{N} \left\|x_i^{(k)} - \bar{x}^{(k)}\right\|^2 = \left\|x^{(k)} - \bar{\mathbf{x}}^{(k)}\right\|^2 = \left\|x^{(k)} - \frac{1}{N}\left(\left(1_N 1_N^T\right) \otimes I_d\right) x^{(k)}\right\|^2.$$

Next, we notice that it follows from (136) that

$$\begin{aligned}
&x^{(k)} - \frac{1}{N}\left(\left(1_N 1_N^T\right) \otimes I_d\right) x^{(k)} \\
&= \left(W^k \otimes I_d\right) x^{(0)} - \frac{1}{N}\left(\left(1_N 1_N^T W^k\right) \otimes I_d\right) x^{(0)} \\
&- \alpha \sum_{s=0}^{k-1} \left(W^{k-1-s} \otimes I_d\right) \nabla F\left(y^{(s)}\right) + \alpha \sum_{s=0}^{k-1} \frac{1}{N}\left(\left(1_N 1_N^T W^{k-1-s}\right) \otimes I_d\right) \nabla F\left(y^{(s)}\right) \\
&\quad - \alpha \sum_{s=0}^{k-1} \left(W^{k-1-s} \otimes I_d\right) \xi^{(s+1)} + \alpha \sum_{s=0}^{k-1} \frac{1}{N}\left(\left(1_N 1_N^T W^{k-1-s}\right) \otimes I_d\right) \xi^{(s+1)} \\
&\quad + \sum_{s=0}^{k-1} \left(W^{k-s} \otimes I_d\right) \beta \left(x^{(s)} - x^{(s-1)}\right) - \sum_{s=0}^{k-1} \frac{1}{N}\left(\left(1_N 1_N^T W^{k-s}\right) \otimes I_d\right) \beta \left(x^{(s)} - x^{(s-1)}\right).
\end{aligned}$$

By the Cauchy-Schwarz inequality, we have

$$
\begin{aligned}
&\left\|x^{(k)}-\frac{1}{N}\left(\left(1_N1_N^T\right)\otimes I_d\right)x^{(k)}\right\|^2\\
&\leq 4\left\|(W^k\otimes I_d)x^{(0)}-\frac{1}{N}\left(\left(1_N1_N^TW^k\right)\otimes I_d\right)x^{(0)}\right\|^2\\
&\quad+4\left\|-\alpha\sum_{s=0}^{k-1}\left(W^{k-1-s}\otimes I_d\right)\nabla F\left(y^{(s)}\right)+\alpha\sum_{s=0}^{k-1}\frac{1}{N}\left(\left(1_N1_N^TW^{k-1-s}\right)\otimes I_d\right)\nabla F\left(y^{(s)}\right)\right\|^2\\
&\quad+4\left\|\alpha\sum_{s=0}^{k-1}\left(W^{k-1-s}\otimes I_d\right)\xi^{(s+1)}-\alpha\sum_{s=0}^{k-1}\frac{1}{N}\left(\left(1_N1_N^TW^{k-1-s}\right)\otimes I_d\right)\xi^{(s+1)}\right\|^2\\
&\quad+4\left\|\sum_{s=0}^{k-1}\left(W^{k-s}\otimes I_d\right)\beta\left(x^{(s)}-x^{(s-1)}\right)-\sum_{s=0}^{k-1}\frac{1}{N}\left(\left(1_N1_N^TW^{k-s}\right)\otimes I_d\right)\beta\left(x^{(s)}-x^{(s-1)}\right)\right\|^2\\
&=4\left\|(W^k\otimes I_d)x^{(0)}-\frac{1}{N}\left(\left(1_N1_N^T\right)\otimes I_d\right)x^{(0)}\right\|^2\\
&\quad+4\left\|-\alpha\sum_{s=0}^{k-1}\left(W^{k-1-s}\otimes I_d\right)\nabla F\left(y^{(s)}\right)+\alpha\sum_{s=0}^{k-1}\frac{1}{N}\left(\left(1_N1_N^T\right)\otimes I_d\right)\nabla F\left(y^{(s)}\right)\right\|^2\\
&\quad+4\left\|\alpha\sum_{s=0}^{k-1}\left(W^{k-1-s}\otimes I_d\right)\xi^{(s+1)}-\alpha\sum_{s=0}^{k-1}\frac{1}{N}\left(\left(1_N1_N^T\right)\otimes I_d\right)\xi^{(s+1)}\right\|^2\\
&\quad+4\left\|\sum_{s=0}^{k-1}\left(W^{k-s}\otimes I_d\right)\beta\left(x^{(s)}-x^{(s-1)}\right)-\sum_{s=0}^{k-1}\frac{1}{N}\left(\left(1_N1_N^T\right)\otimes I_d\right)\beta\left(x^{(s)}-x^{(s-1)}\right)\right\|^2,
\end{aligned}
$$

where we used the property that W is doubly stochastic. Therefore, we get

$$
\begin{aligned}
&\left\|x^{(k)}-\frac{1}{N}\left(\left(1_N1_N^T\right)\otimes I_d\right)x^{(k)}\right\|^2\\
&\leq 4\left\|\left(\left(W^k-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)x^{(0)}\right\|^2\\
&\qquad+4\alpha^2\left\|\sum_{s=0}^{k-1}\left(\left(W^{k-1-s}-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)\nabla F\left(y^{(s)}\right)\right\|^2\\
&\qquad\quad+4\alpha^2\left\|\sum_{s=0}^{k-1}\left(\left(W^{k-1-s}-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)\xi^{(s+1)}\right\|^2\\
&\qquad\qquad+4\left\|\sum_{s=0}^{k-1}\left(\left(W^{k-s}-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)\beta\left(x^{(s)}-x^{(s-1)}\right)\right\|^2. \qquad (139)
\end{aligned}
$$

Next, we notice that

$$
\begin{aligned}
&4\alpha^2 \left\| \sum_{s=0}^{k-1} \left(\left(W^{k-1-s} - \frac{1}{N} 1_N 1_N^T \right) \otimes I_d \right) \nabla F\left(y^{(s)}\right) \right\|^2 \\
&\leq 4\alpha^2 \left(\sum_{s=0}^{k-1} \left\| \left(W^{k-1-s} - \frac{1}{N} 1_N 1_N^T \right) \otimes I_d \right\| \cdot \left\| \nabla F\left(y^{(s)}\right) \right\| \right)^2 \\
&\leq 4\alpha^2 \left(\sum_{s=0}^{k-1} \left\| W^{k-1-s} - \frac{1}{N} 1_N 1_N^T \right\| \cdot \left\| \nabla F\left(y^{(s)}\right) \right\| \right)^2 \\
&= 4\alpha^2 \left(\sum_{s=0}^{k-1} \gamma^{k-1-s} \cdot \left\| \nabla F\left(y^{(s)}\right) \right\| \right)^2 \\
&= 4\alpha^2 \left(\sum_{s=0}^{k-1} \gamma^{k-1-s} \right)^2 \left(\frac{\sum_{s=0}^{k-1} \gamma^{k-1-s} \cdot \left\| \nabla F\left(y^{(s)}\right) \right\|}{\sum_{s=0}^{k-1} \gamma^{k-1-s}} \right)^2 \\
&\leq 4\alpha^2 \left(\sum_{s=0}^{k-1} \gamma^{k-1-s} \right)^2 \sum_{s=0}^{k-1} \frac{\gamma^{k-1-s}}{\sum_{s=0}^{k-1} \gamma^{k-1-s}} \left\| \nabla F\left(y^{(s)}\right) \right\|^2, \qquad (140)
\end{aligned}
$$

where we used Jensen's inequality in the last step above, and the fact that W^{k-1-s} has eigenvalues $(\lambda_i^W)^{k-1-s}$ with $1 = \lambda_1^W > \lambda_2^W \geq \cdots \geq \lambda_N^W > -1$, and hence

$$
\left\| W^{k-1-s} - \frac{1}{N} 1_N 1_N^T \right\| = \max\left\{ \left|\lambda_2^W\right|^{k-1-s}, \left|\lambda_N^W\right|^{k-1-s} \right\} = \gamma^{k-1-s}.
$$

Moreover, we can compute that for any k

$$
\begin{aligned}
\mathbb{E}\left\| \nabla F\left(y^{(k)}\right) \right\|^2 &\leq 2\mathbb{E}\left\| \nabla F\left(y^{(k)}\right) - \nabla F\left(x^*\right) \right\|^2 + 2\left\| \nabla F\left(x^*\right) \right\|^2 \\
&\leq 2L^2 \mathbb{E}\left\| y^{(k)} - x^* \right\|^2 + 2\left\| \nabla F\left(x^*\right) \right\|^2 \\
&= 2L^2 \mathbb{E}\left\| (1+\beta)\left(x^{(k)} - x^*\right) - \beta\left(x^{(k-1)} - x^*\right) \right\|^2 + 2\left\| \nabla F\left(x^*\right) \right\|^2 \\
&\leq 4L^2(1+\beta)^2 \mathbb{E}\left\| x^{(k)} - x^* \right\|^2 + 4L^2\beta^2 \mathbb{E}\left\| x^{(k-1)} - x^* \right\|^2 + 2\left\| \nabla F\left(x^*\right) \right\|^2 \\
&\leq D_y^2, \qquad (141)
\end{aligned}
$$

where D_y^2 is defined in (34) and we used Corollary 8 to obtain the last line above. Therefore, by (140), we have

$$
\begin{aligned}
&4\alpha^2 \mathbb{E}\left[\left\| \sum_{s=0}^{k-1} \left(\left(W^{k-1-s} - \frac{1}{N} 1_N 1_N^T \right) \otimes I_d \right) \nabla F\left(y^{(s)}\right) \right\|^2 \right] \\
&\leq 4D_y^2 \alpha^2 \left(\sum_{s=0}^{k-1} \gamma^{k-1-s} \right)^2 \sum_{s=0}^{k-1} \frac{\gamma^{k-1-s}}{\sum_{s=0}^{k-1} \gamma^{k-1-s}} \leq 4D_y^2 \alpha^2 \frac{1}{(1-\gamma)^2}.
\end{aligned}
$$

Similarly, we can show that

$$\begin{aligned}
&4\left\|\left(\left(W^k-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)x^{(0)}\right\|^2\\
&\leq 4\left\|\left(W^k-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right\|^2\left\|x^{(0)}\right\|^2\\
&\leq 4\gamma^{2k}\left\|x^{(0)}\right\|^2.
\end{aligned}$$

This implies that

$$\begin{aligned}
&4\mathbb{E}\left\|\left(\left(W^k-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)x^{(0)}\right\|^2\\
&\leq 8\gamma^{2k}\mathbb{E}\left\|x^{(0)}-x^*\right\|^2+8\gamma^{2k}\|x^*\|^2\\
&\leq 8\gamma^{2k}\left(4\frac{V_{S,\alpha}\left(\xi_0\right)}{\mu}+\frac{2\sigma^2N\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(2-\lambda_N^W+\alpha L\right)+\frac{2C_1^2N\alpha^2}{(1-\gamma)^2}+\|x^*\|^2\right),
\end{aligned}$$

where we used Corollary 8.

In addition, by applying Lemma 29, we can show that

$$\begin{aligned}
&4\sum_{s=0}^{k-1}\mathbb{E}\left\|\left(\left(W^{k-s}-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)\beta\left(x^{(s)}-x^{(s-1)}\right)\right\|^2\\
&\leq 4\frac{\beta^2}{(1-\gamma)^2}\sup_{s\geq 0}\mathbb{E}\left\|\left(x^{(s)}-x^{(s-1)}\right)\right\|^2\\
&\leq 4\frac{\beta^2}{(1-\gamma)^2}\frac{2C_0}{\beta^2}\alpha\\
&=\frac{8C_0}{(1-\gamma)^2}\alpha.
\end{aligned}$$

Finally, we can show that

$$\begin{aligned}
&4\alpha^2\sum_{s=0}^{k-1}\mathbb{E}\left\|\left(\left(W^{k-1-s}-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)\xi^{(s+1)}\right\|^2\\
&\leq 4\alpha^2\sum_{s=0}^{k-1}\left\|W^{k-1-s}-\frac{1}{N}1_N1_N^T\right\|^2\mathbb{E}\left\|\xi^{(s+1)}\right\|^2\\
&\leq\frac{4\sigma^2N\alpha^2}{(1-\gamma)^2}.
\end{aligned}$$

Hence, it follows from (139) that

$$\begin{aligned}
&\sum_{i=1}^{N}\mathbb{E}\left\|x_i^{(k)}-\bar{x}^{(k)}\right\|^2\\
&=\left\|x^{(k)}-\frac{1}{N}\left(\left(1_N1_N^T\right)\otimes I_d\right)x^{(k)}\right\|^2\\
&\leq 4\gamma^{2k}\mathbb{E}\left\|x^{(0)}\right\|^2+4D_y^2\alpha^2\frac{1}{(1-\gamma)^2}+4\alpha^2\sum_{s=0}^{k-1}\mathbb{E}\left\|\left(\left(W^{k-1-s}-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)\xi^{(s+1)}\right\|^2\\
&\quad+4\sum_{s=0}^{k-1}\mathbb{E}\left\|\left(\left(W^{k-s}-\frac{1}{N}1_N1_N^T\right)\otimes I_d\right)\beta\left(x^{(s)}-x^{(s-1)}\right)\right\|^2\\
&\leq 8\gamma^{2k}\left(4\frac{V_{S,\alpha}\left(\xi_0\right)}{\mu}+\frac{2\sigma^2N\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(2-\lambda_N^W+\alpha L\right)+\frac{2C_1^2N\alpha^2}{(1-\gamma)^2}+\|x^*\|^2\right)\\
&\quad+4D_y^2\alpha^2\frac{1}{(1-\gamma)^2}+\frac{4\sigma^2N\alpha^2}{(1-\gamma)^2}+\frac{8C_0}{(1-\gamma)^2}\alpha.
\end{aligned}$$

The proof is complete. ■

F.1.3 Proof of Lemma 28

Proof We notice that

$$\begin{aligned}
&\mathbb{E}\left\|\mathcal{E}_{k+1}\right\|^2\\
&=\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\left(\nabla f_i\left(y_i^{(k)}\right)-\nabla f_i\left(\bar{y}^{(k)}\right)\right)\right\|^2\\
&\leq\frac{1}{N^2}\sum_{i=1}^{N}N\mathbb{E}\left\|\nabla f_i\left(y_i^{(k)}\right)-\nabla f_i\left(\bar{y}^{(k)}\right)\right\|^2\\
&\leq\frac{1}{N}L^2\sum_{i=1}^{N}\mathbb{E}\left\|y_i^{(k)}-\bar{y}^{(k)}\right\|^2\\
&\leq\frac{2}{N}L^2\sum_{i=1}^{N}\left((1+\beta)^2\mathbb{E}\left\|x_i^{(k)}-\bar{x}^{(k)}\right\|^2+\beta^2\mathbb{E}\left\|x_i^{(k-1)}-\bar{x}^{(k-1)}\right\|^2\right)\\
&\leq\frac{2}{N}L^2\left((1+\beta)^2+\beta^2\right)\left[4D_y^2\alpha^2\frac{1}{(1-\gamma)^2}+\frac{4\sigma^2N\alpha^2}{(1-\gamma)^2}+\frac{8C_0}{(1-\gamma)^2}\alpha\right.\\
&\qquad\left.+8\gamma^{2(k-1)}\left(4\frac{V_{S,\alpha}\left(\xi_0\right)}{\mu}+\frac{2\sigma^2N\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(2-\lambda_N^W+\alpha L\right)+\frac{2C_1^2N\alpha^2}{(1-\gamma)^2}+\|x^*\|^2\right)\right],
\end{aligned}$$

where we used Lemma 27. The proof is complete. ■

F.1.4 Proof of Lemma 29

Proof We first rewrite the D-ASG iterations (18) as

$$z^{(k+1)} = \left(\tilde{M} \otimes I_d\right) z^{(k)} - \alpha \begin{bmatrix} \tilde{\nabla} F\left(C z^{(k)}\right) \\ 0 \end{bmatrix}, \quad z^{(k)} = \begin{bmatrix} x^{(k)} \\ y^{(k)} \end{bmatrix}, \tag{142}$$

where

$$\tilde{M} = \begin{bmatrix} (1+\beta)W & -\beta W \\ I_N & 0_N \end{bmatrix}. \tag{143}$$

By a reasoning similar to the proof of Proposition 36, we observe that $\tilde{M}$ is block diagonalizable with 2×2 blocks satisfying

$$\tilde{M} = \tilde{O} \mathbf{diag}\left(\{Z_i\}_{i=1}^N\right) \tilde{O}^T, \tag{144}$$

where

$$\tilde{Z}_i = \begin{bmatrix} (1+\beta)\lambda_i^W & -\beta\lambda_i^W \\ 1 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}, \qquad 1 \le i \le N,$$

λ_i^W are the eigenvalues of W in decreasing order, $\tilde{O} = \tilde{U}\tilde{P}_{\tilde{\pi}}$ is orthogonal with $\tilde{U}$ and $\tilde{P}_{\tilde{\pi}}$ are defined as $\tilde{U} = \mathbf{diag}(V, V)$ where $W = VDV^T$ is the eigenvalue decomposition of W and $\tilde{P}_{\tilde{\pi}}$ is the permutation matrix associated with the permutation $\tilde{\pi}$ over $\{1, 2, \ldots, 2N\}$ that satisfies

$$\tilde{\pi}(i) = \begin{cases} 2i-1 & \text{if } 1 \le i \le N, \\ 2(i-N) & \text{if } N+1 \le i \le 2N. \end{cases} \tag{145}$$

We also observe that $\tilde{Z}_i$ has eigenvalues

$$\mu_{i,\pm} := \frac{(1+\beta)\lambda_i^W \pm \sqrt{(1+\beta)^2(\lambda_i^W)^2 - 4\beta\lambda_i^W}}{2}.$$

In particular, in the special case when $i = 1$, we have $\lambda_1^W = 1$ and $\tilde{Z}_1$ has two eigenvalues $\mu_{1,+} = 1$ and $\mu_{1,-} = \beta < 1$ for $i = 1, 2, \ldots, d$, admitting the Jordan decomposition

$$\tilde{Z}_i = S_1 \begin{bmatrix} 1 & 0 \\ 0 & \beta \end{bmatrix} S_1^{-1}, \quad \text{for} \quad i = 1, 2, \ldots, d,$$

where

$$S_1 = \begin{bmatrix} 1 & \beta \\ 1 & 1 \end{bmatrix}, \quad S_1^{-1} = \frac{1}{1-\beta} \begin{bmatrix} 1 & -\beta \\ -1 & 1 \end{bmatrix}.$$

Similarly, for $i > 1$, we can also write the Jordan decomposition of $\tilde{Z}_i$ as

$$\tilde{Z}_i = S_i J_i S_i^{-1} \quad \text{for} \quad 1 < i \le N, \tag{146}$$

where

$$S_i = \begin{cases} \begin{bmatrix} \mu_{i,+} & \mu_{i,-} \\ 1 & 1 \end{bmatrix} & \text{if} \quad \mu_{i,+} \neq \mu_{i,-}, \\ \begin{bmatrix} \mu_{i,+} & 1 \\ 1 & 1 \end{bmatrix} & \text{if} \quad \mu_{i,+} = \mu_{i,-}, \end{cases}, \tag{147}$$

$$S_i^{-1} = \begin{cases} \frac{1}{\mu_{i,+}-\mu_{i,-}} \begin{bmatrix} 1 & -\mu_{i,-} \\ -1 & \mu_{i,+} \end{bmatrix} & \text{if} \quad \mu_{i,+} \neq \mu_{i,-}, \\ \frac{1}{\mu_{i,+}-1} \begin{bmatrix} 1 & -1 \\ -1 & \mu_{i,+} \end{bmatrix} & \text{if} \quad \mu_{i,+} = \mu_{i,-}, \end{cases} \tag{148}$$

and

$$J_i = \begin{cases} \begin{bmatrix} \mu_{i,+} & 0 \\ 0 & \mu_{i,-} \end{bmatrix} & \text{if} \quad \mu_{i,+} \neq \mu_{i,-}, \\ \begin{bmatrix} \mu_{i,+} & 1 \\ 0 & \mu_{i,+} \end{bmatrix} & \text{if} \quad \mu_{i,+} = \mu_{i,-}. \end{cases}$$

Basically, the structure of the Jordan blocks J_i will depend on the multiplicity of the eigenvalues $s\mu_{i,+}, \mu_{i,-}$ which itself depends on the stepsize chosen and the eigenvalues of the matrix W. Next, we introduce

$$\bar{Z}_i := \begin{cases} S_1 \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} S_1^{-1} & \text{if} \quad i = 1, \\ 0 & \text{if} \quad 1 < i \leq N, \end{cases}$$

as well as

$$Z := \mathbf{diag}\left(\{Z_i\}_{i=1}^N\right), \quad \bar{Z} := \mathbf{diag}\left(\{\bar{Z}_i\}_{i=1}^N\right).$$

Note that the matrices Z and $\bar{Z}$ are block diagonal with 2×2 blocks and they have both 1 as a simple eigenvalue with the same eigenvectors; however other eigenvalues of $\bar{Z}$ is set to zero. We have also

$$Z_i - \bar{Z}_i = \begin{cases} S_1 \begin{bmatrix} 0 & 0 \\ 0 & \beta \end{bmatrix} S_1^{-1} & \text{if} \quad i = 1, \\ \tilde{Z}_i & \text{if} \quad 1 < i \leq N. \end{cases} \tag{149}$$

It can also be computed that

$$\bar{M} := \tilde{O}\bar{Z}\tilde{O}^T = \frac{1}{1-\beta} \begin{bmatrix} v_1 v_1^T & -\beta v_1 v_1^T \\ v_1 v_1^T & -\beta v_1 v_1^T \end{bmatrix},$$

where $v_1 = \frac{1}{\sqrt{N}}\mathbf{1} \in \mathbb{R}^N$ is the eigenvector of W corresponding to the eigenvalue 1. Consider

$$\left(\bar{M} \otimes I_d\right) z^{(k)} = \frac{1}{1-\beta} \begin{bmatrix} \bar{x}^{(k)} - \beta \bar{x}^{(k-1)} \\ \vdots \\ \bar{x}^{(k)} - \beta \bar{x}^{(k-1)} \end{bmatrix} = \frac{1}{1-\beta} \begin{bmatrix} \bar{\mathbf{x}}^{(k)} - \beta \bar{\mathbf{x}}^{(k-1)} \\ \bar{\mathbf{x}}^{(k)} - \beta \bar{\mathbf{x}}^{(k-1)} \end{bmatrix} \in \mathbb{R}^{2Nd}, \tag{150}$$

which can be viewed as a weighted average of $\bar{x}^{(k)}$ and $\bar{x}^{(k-1)}$ with $\bar{\mathbf{x}}^{(k)}$ defined as in (138). From the definition of $\bar{M}$ and the decomposition (144), we have

$$\tilde{M} - \bar{M} = \tilde{O}\left(Z - \bar{Z}\right)\tilde{O}^T, \tag{151}$$

and this matrix has the spectral radius

$$r := \rho\left(\tilde{M} - \bar{M}\right) = \max\left(\beta, \max_{j\geq 2}\rho(J_j)\right) \tag{152}$$

$$= \max\left(\beta, \left|\frac{(1+\beta)\lambda_2^W + \sqrt{(1+\beta)^2(\lambda_2^W)^2 - 4\beta\lambda_2^W}}{2}\right|\right) \tag{153}$$

$$< r_3 := \max\left(\beta, \sqrt{\lambda_2^W}\right) < 1, \tag{154}$$

where in the last inequality we used the facts that the function

$$g(\lambda,\beta) := \left|\frac{(1+\beta)\lambda + \sqrt{(1+\beta)^2(\lambda)^2 - 4\beta\lambda}}{2}\right| \tag{155}$$

defined for $\lambda, \beta \in [0,1]$ is increasing in λ on the interval $[0,1]$ for fixed $\beta \in [0,1]$ and is increasing in β on the interval $[0,1]$ for fixed $\alpha \in [0,1]$ which results in the inequalities $g(\lambda_i^W,\beta) \leq g(\lambda_2^W,\beta) \leq g(\lambda_2^W,1) = \sqrt{\lambda_2^W}$ for $i \geq 2$. From (154), it follows that

$$\left\|(M-\bar{M})^k\right\| \leq C_3 r_3^k \quad \text{for all} \quad k \geq 0, \tag{156}$$

for some positive constant C_3. The constant C_3 will depend on the parameter α in general (as β and r_3 are functions of α) but it will not depend on k. A natural question would be how the constants C_3 and r_3 change as a function of α as $\alpha \to 0$. It follows from Lemma 50 that one can choose c_3 and r_3 such that

$$C_3 = \Theta\left(\frac{1}{\sqrt{\alpha}}\right), \quad r_3 = 1 - \Theta(\sqrt{\alpha}) \quad \text{as} \quad \alpha \to 0. \tag{157}$$

Furthermore,

$$\begin{aligned}
&\left(\left(I_{2N} - \bar{M}\right) \otimes I_d\right) z^{(k+1)} \\
&= \left(\left(I_{2N} - \bar{M}\right)\tilde{M} \otimes I_d\right) z^{(k)} - \left(\left(I_{2N} - \bar{M}\right) \otimes I_d\right) \alpha \begin{bmatrix} \tilde{\nabla} F\left(Cz^{(k)}\right) \\ 0 \end{bmatrix} \\
&= \left(\left(\tilde{M} - \bar{M}\right)\left(I_{2N} - \bar{M}\right) \otimes I_d\right) z^{(k)} - \left(\left(I_{2N} - \bar{M}\right) \otimes I_d\right) \alpha \begin{bmatrix} \tilde{\nabla} F\left(Cz^{(k)}\right) \\ 0 \end{bmatrix},
\end{aligned} \tag{158}$$

where we used the facts that $\tilde{M}\bar{M} = \bar{M}\tilde{M} = \bar{M}$ and $\bar{M}^2 = \bar{M}$. On the other hand,

$$\begin{aligned}
\left(\left(I_{2N} - \bar{M}\right) \otimes I_d\right) z^{(k+1)} = &\left((\tilde{M} - \bar{M})^{k+1}(I_{2N} - \bar{M}) \otimes I_d\right) z^{(0)} \\
&- \alpha \sum_{j=0}^{k} \left((\tilde{M} - \bar{M})^j (I_{2N} - \bar{M}) \otimes I_d\right) \begin{bmatrix} \tilde{\nabla} F\left(Cz^{(k-j)}\right) \\ 0 \end{bmatrix}.
\end{aligned} \tag{159}$$

Therefore, for $z^{(0)} = 0$,

$$
\begin{aligned}
&\mathbb{E}\left\|\left((I_{2N} - \bar{M}) \otimes I_d\right) z^{(k+1)}\right\|^2 \\
&= \alpha^2 \mathbb{E}\left\|\sum_{j=0}^{k} \left(\left(\tilde{M} - \bar{M}\right)^j \left(I_{2N} - \bar{M}\right) \otimes I_d \right) \begin{bmatrix} \tilde{\nabla} F\left(C z^{(k-j)}\right) \\ 0 \end{bmatrix} \right\|^2 \\
&\leq \alpha^2 \mathbb{E}\left\|\sum_{j=0}^{k} \left(\left(\tilde{M} - \bar{M}\right)^j \left(I_{2N} - \bar{M}\right) \otimes I_d \right) \begin{bmatrix} \tilde{\nabla} F\left(C z^{(k-j)}\right) \\ 0 \end{bmatrix} \right\|^2 \\
&= \alpha^2 \mathbb{E}\left\|\sum_{j=0}^{k} \left(\left(\left(\tilde{M} - \bar{M}\right)^{j+1} + \left(\tilde{M} - \bar{M}\right)^j \left(I_{2N} - \tilde{M}\right) \right) \otimes I_d \right) \begin{bmatrix} \tilde{\nabla} F\left(C z^{(k-j)}\right) \\ 0 \end{bmatrix} \right\|^2 \\
&\leq \alpha^2 \mathbb{E}\left| \sum_{j=0}^{k} c_3 r_3^j \left(1 + \left\|I_{2N} - \tilde{M}\right\|\right) \left\|\tilde{\nabla} F\left(C z^{(k-j)}\right)\right\| \right|^2 \\
&\leq \alpha^2 \left(\sum_{j=0}^{k} a_j\right)^2 \mathbb{E}\left| \sum_{j=0}^{k} \frac{a_j}{\left(\sum_{j=0}^{k} a_j\right)} \left\|\tilde{\nabla} F\left(C z^{(k-j)}\right)\right\| \right|^2 \\
&\leq \alpha^2 \left(\sum_{j=0}^{k} a_j\right)^2 \sum_{j=0}^{k} \frac{a_j}{\left(\sum_{j=0}^{k} a_j\right)} \mathbb{E}\left\|\tilde{\nabla} F\left(C z^{(k-j)}\right)\right\|^2 \\
&\leq \alpha^2 \left(\sum_{j=0}^{k} a_j\right)^2 D_y^2, \qquad (160)
\end{aligned}
$$

where we used (141), D_y^2 is defined in (34) and

$$a_j := c_3 r_3^j \left(1 + \left\|I_{2N} - \tilde{M}\right\|\right).$$

Note that

$$
\begin{aligned}
\left\|I_{2N} - \tilde{M}\right\| &\leq \|I_{2N}\| + \left\|\tilde{M}\right\| \\
&\leq 1 + \left\|\tilde{M}\right\|_F \\
&= 1 + \sqrt{\left((1+\beta)^2 + \beta^2\right) \|W\|_F^2 + N}, \qquad (161)
\end{aligned}
$$

where we used the definition of $\tilde{M}$. Consequently,

$$\sum_{j=0}^{\infty} a_j \leq \sum_{j \geq 0} c_3 r_3^j \left(1 + \left\|I_{2N} - \tilde{M}\right\|\right) \leq c_3 \frac{1}{(1 - r_3)} \left(1 + \sqrt{\left((1+\beta)^2 + \beta^2\right) \|W\|_F^2 + N}\right).$$

Therefore, we conclude from (150) and (160) that

$$\begin{aligned}\sup_k \mathbb{E}\left\|\left((I_{2N}-\bar{M})\otimes I_d\right)z^{(k)}\right\|^2 &= \sup_k \mathbb{E}\left\|\begin{bmatrix} x^{(k)}-\frac{1}{1-\beta}\left(\bar{\mathbf{x}}^{(k)}-\beta\bar{\mathbf{x}}^{(k-1)}\right)\\ y^{(k)}-\frac{1}{1-\beta}\left(\bar{\mathbf{x}}^{(k)}-\beta\bar{\mathbf{x}}^{(k-1)}\right)\end{bmatrix}\right\|^2\\ &\le D_y^2\alpha^2\frac{c_3}{(1-r_3)}\left(1+\sqrt{((1+\beta)^2+\beta^2)\left\|W\right\|_F^2+N}\right)\\ &\le D_y^2\alpha^2\frac{c_3}{(1-r_3)}\left(1+\sqrt{5\|W\|_F^2+N}\right), \end{aligned} \tag{162}$$

where $\bar{\mathbf{x}}^{(k)}$ is defined by (138) and we used the fact that $\beta \le 1$. From (157), we observe that the term $\frac{c_3}{1-r_3} = \mathcal{O}(\frac{1}{\alpha})$. We conclude that the right hand-side of (162) is $O(\alpha)$. Hence, we conclude that

$$\begin{aligned}&\sup_k \mathbb{E}\left\|x^{(k)}-y^{(k)}\right\|^2\\ &\le 2\sup_k\left(\mathbb{E}\left\|x^{(k)}-\frac{1}{1-\beta}\left(\bar{\mathbf{x}}^{(k)}-\beta\bar{\mathbf{x}}^{(k-1)}\right)\right\|^2+\mathbb{E}\left\|y^{(k)}-\frac{1}{1-\beta}\left(\bar{\mathbf{x}}^{(k)}-\beta\bar{\mathbf{x}}^{(k-1)}\right)\right\|^2\right)\\ &\le 2C_0\alpha,\end{aligned}$$

where C_0 is some constant such that $C_0 = \mathcal{O}(1)$ as $\alpha \to 0$. Finally, we notice that $y^{(k)} = (1+\beta)x^{(k)} - \beta x^{(k-1)}$, and this implies that,

$$\sup_k \mathbb{E}\left\|x^{(k)}-x^{(k-1)}\right\|^2 = \sup_k \mathbb{E}\left\|\frac{1}{\beta}\left(x^{(k)}-y^{(k)}\right)\right\|^2 \le \frac{2C_0}{\beta^2}\alpha, \tag{163}$$

which completes the proof. ■

Lemma 50 *In the setting of the proof of Lemma 29,*

$$c_3 = \Theta\left(\frac{1}{\sqrt{\alpha}}\right), \quad r_3 = 1-\Theta(\sqrt{\alpha}) \quad \textit{as} \quad \alpha \to 0.$$

Proof Note that we have from (151), (146), (149),

$$\left\|(M-\bar{M})^k\right\| = \left\|\tilde{O}\left(Z-\bar{Z}\right)^k\tilde{O}^T\right\| = \max_i\left\|(Z_i-\bar{Z}_i)^k\right\| \tag{164}$$

$$\le \max\left(\|S_1\|\left\|S_1^{-1}\right\|\beta^k, \max_{2\le i\le N}\|S_i\|\left\|J_i^k\right\|\left\|S_i^{-1}\right\|\right), \tag{165}$$

where we used the fact that $\tilde{O}$ is orthogonal. Also,

$$\|S_i\|_2 \le \|S_i\|_F \le 2 \quad \text{for} \quad 1 \le i \le N, \tag{166}$$

where we used the definition of S_i and the inequalities $|\mu_{i,+}| = g(\lambda_i^W, \beta) \leq g(1,1) = 1$ and $|\mu_{i,-}| \leq 1$ which follow from the definition (155) of the function $g(\lambda, \beta)$ and its monotonicity property with respect to λ and β. Similarly,

$$\left\|S_1^{-1}\right\|_2 \leq \left\|S_1^{-1}\right\|_F \leq \frac{2}{1-\beta}, \tag{167}$$

and for $1 < i \leq N$,

$$\left\|S_i^{-1}\right\|_2 \leq \left\|S_i^{-1}\right\|_F \leq c_i := \begin{cases} \frac{2}{|\mu_{i,+}-\mu_{i,-}|} = \frac{2}{\sqrt{|(1+\beta)^2(\lambda_i^W)^2 - 4\beta\lambda_i^W|}} & \text{if} \quad \mu_{i,+} \neq \mu_{i,-}, \\ \frac{2}{|\mu_{i,+}-1|} & \text{if} \quad \mu_{i,+} = \mu_{i,-}. \end{cases} \tag{168}$$

In addition,

$$J_i^k = \begin{cases} \begin{bmatrix} \mu_{i,+}^k & 0 \\ 0 & \mu_{i,-}^k \end{bmatrix} & \text{if} \quad \mu_{i,+} \neq \mu_{i,-}, \\ \begin{bmatrix} \mu_{i,+}^k & k\mu_{i,+}^{k-1} \\ 0 & \mu_{i,+}^k \end{bmatrix} & \text{if} \quad \mu_{i,+} = \mu_{i,-}, \end{cases}$$

used again the fact that $|\mu_{i,+}| \leq 1$, we have then

$$\begin{aligned} \left\|J_i^k\right\|_2 \leq \left\|J_i^k\right\|_F &= \begin{cases} |\mu_{i,+}|^k & \text{if} \quad \mu_{i,+} \neq \mu_{i,-} \\ \sqrt{2|\mu_{i,+}|^{2k} + k^2|\mu_{i,+}|^{2k-2}} & \text{if} \quad \mu_{i,+} = \mu_{i,-} \end{cases} \\ &\leq d_{i,k}|\mu_{i,+}|^{k-1}, \end{aligned} \tag{169}$$

with

$$d_{i,k} := \begin{cases} 1 & \text{if} \quad \mu_{i,+} \neq \mu_{i,-}, \\ \sqrt{2}+k & \text{if} \quad \mu_{i,+} = \mu_{i,-}. \end{cases}$$

Combining all the estimates (166), (167), (168), (169) together, we obtain

$$\begin{aligned} \max_{2\leq i\leq N} \|S_i\| \left\|J_i^k\right\| \left\|S_i^{-1}\right\| &\leq 2\left(\max_{2\leq i\leq N} c_i d_{i,k}\right)\left(\max_{2\leq i\leq N} |\mu_{i,+}|^{k-1}\right) \\ &= 2\left(\max_{2\leq i\leq N} c_i d_{i,k}\right)|\mu_{2,+}|^{k-1}. \end{aligned} \tag{170}$$

Note that we have $|\mu_{i,+}| = g(\lambda_i^W, \beta) \leq g(\lambda_2^W, \beta) = \mu_{2,+} < 1$. Therefore,

$$|\mu_{2,+}| < r_2 := \frac{1+|\mu_{2,+}|}{2}.$$

Furthermore, $\max_{2\leq i\leq N} c_i(\alpha) d_{i,k} = \mathcal{O}(k)$. Consequently, we can bound (170) as

$$\max_{2\leq i\leq N} \|S_i\| \left\|J_i^k\right\| \left\|S_i^{-1}\right\| \leq C_2 r_2^{k-1} \tag{171}$$

for some constant $C_2 = \mathcal{O}(1)$ that does not depend on k. Then, from (165), it follows that

$$\begin{aligned} \left\|\left(M-\bar{M}\right)^k\right\| &\leq \max\left(\|S_1\|\left\|S_1^{-1}\right\|\beta^k, \max_{2\leq i\leq N}\|S_i\|\left\|J_i^k\right\|\left\|S_i^{-1}\right\|\right) \\ &\leq \max\left(\frac{4}{1-\beta}\beta^k, C_2 r_2^{k-1}\right) \\ &\leq C_3 r_3^k, \end{aligned} \tag{172}$$

where

$$C_3 := \max\left(\frac{4}{1-\beta}, C_2\right), \quad r_3 = \max(\beta, r_2).$$

Also $|\mu_{2,+}| = g(\lambda_2^W, \beta) < g(\lambda_2^W, 1) = \sqrt{\lambda_2^W}$. Therefore, $r_2 < \frac{1+\sqrt{\lambda_2^W}}{2}$. Since $\beta = 1-\Theta(\sqrt{\alpha})$, we observe that

$$C_3 = \Theta\left(\frac{1}{\alpha}\right), \quad r_3 = 1 - \Theta(\sqrt{\alpha})$$

as $\alpha \to 0$. The proof is complete. ■

F.1.5 Proof of Lemma 30

Proof The function $V_{\bar{S},\alpha}$ is L_α smooth where $\bar{L}_\alpha = L + 2\|\bar{S}_\alpha\|$. Note that

$$\bar{S}_\alpha = vv^T, \qquad v = \begin{bmatrix} \sqrt{\frac{1}{2\alpha}} \\ \sqrt{\frac{\mu}{2}} - \sqrt{\frac{1}{2\alpha}} \end{bmatrix}.$$

Therefore,

$$\left\|\bar{S}_\alpha\right\| = \|v\|^2 = \frac{1}{\alpha} + \frac{\mu}{2} - \frac{\sqrt{\mu}}{\sqrt{\alpha}}. \tag{173}$$

For any $\xi, \Delta \in \mathbb{R}^{2d}$, by the $\bar{L}_\alpha$-smoothness of $V_{\bar{S},\alpha}$ we have also

$$\begin{aligned} V_{\bar{S},\alpha}(\xi+\Delta) &\le V_{\bar{S},\alpha}(\xi) + \left\langle \nabla V_{\bar{S},\alpha}(\xi), \Delta \right\rangle + \frac{\bar{L}_\alpha}{2}\|\Delta\|^2 \\ &\le V_{\bar{S},\alpha}(\xi) + c_1 \left\|\nabla V_{\bar{S},\alpha}(\xi)\right\|^2 + \|\Delta\|^2/(4c_1) + \frac{\bar{L}_\alpha}{2}\|\Delta\|^2, \end{aligned} \tag{174}$$

for any $c_1 > 0$ where we used Cauchy-Schwarz in the last inequality. We have also

$$\begin{aligned} \left\|\nabla V_{\bar{S},\alpha}(\xi)\right\|^2 &\le 2\|2\bar{S}_\alpha\xi\|^2 + 2\left\|\nabla f\left(\bar{T}\xi + x_*\right)\right\|^2 \\ &\le 8\xi^T\bar{S}_\alpha^2\xi + 2L^2\left\|\bar{T}\xi\right\|^2 \\ &\le 8\left(\frac{1}{\alpha} + \frac{\mu}{2} - \frac{\sqrt{\mu}}{\sqrt{\alpha}}\right)\xi^T\bar{S}_\alpha\xi + 2L^2\left\|\bar{T}\xi\right\|^2 \\ &\le 8\left(\frac{1}{\alpha} + \frac{\mu}{2} - \frac{\sqrt{\mu}}{\sqrt{\alpha}}\right)\xi^T\bar{S}_\alpha\xi + 2L^2\frac{f(\bar{T}\xi + x_*) - f(x_*)}{\mu} \\ &\le c_2 V_{\bar{S},\alpha}(\xi), \end{aligned} \tag{175}$$

where

$$c_2 := 2\max\left(4\left(\frac{1}{\alpha} + \frac{\mu}{2} - \frac{\sqrt{\mu}}{\sqrt{\alpha}}\right), \frac{L^2}{\mu}\right),$$

and we used the facts that

$$\bar{S}_\alpha^2 = \|v\|^2\bar{S}_\alpha = \left(\frac{1}{\alpha} + \frac{\mu}{2} - \frac{\sqrt{\mu}}{\sqrt{\alpha}}\right)\bar{S}_\alpha.$$

Therefore, if choose $c_1 = \epsilon/c_2$; then we get from (174)

$$V_{\bar{S},\alpha}(\xi + \Delta) \leq (1+\epsilon)V_{\bar{S},\alpha}(\xi) + C_\epsilon \|\Delta\|^2, \tag{176}$$

with $C_\epsilon = \frac{c_2}{4\epsilon} + \bar{L}_\alpha/2$. Finally, if we choose $\xi = \bar{\xi}_{k+1} - D_{k+1}$ and $\Delta = D_{k+1}$; we obtain (63). This completes the proof. ■

F.2 Proofs of Technical Results in Appendix C

F.2.1 Proof of Lemma 41

Proof We recall from the proof of Proposition 36 that

$$\mathcal{W} - \alpha Q = R\mathbf{diag}\left([\mu_i]_{i=1}^{Nd}\right) R^T,$$

where R is real orthogonal and the eigenvalues μ_i are listed in non-increasing order. Then we can write

$$P_\pi U A_{\mathrm{dasg},Q} U^T P_\pi^T = \mathbf{diag}\left(\tilde{T}_i\right) \quad \text{where} \quad U = \mathbf{diag}(R, R)$$

is orthogonal and

$$\tilde{T}_i = \begin{bmatrix} (1+\beta)\mu_i & -\beta\mu_i \\ 1 & 0 \end{bmatrix} \in \mathbb{R}^{2\times 2}, \qquad 1 \leq i \leq Nd.$$

Therefore, we have

$$\left\| A_{\mathrm{dasg},Q}^k \right\| \leq \|V\|^2 \max_{1\leq i\leq Nd} \left\| \left(\tilde{T}_i\right)^k \right\| = \max_{1\leq i\leq Nd} \left\| \left(\tilde{T}_i\right)^k \right\|,$$

where we used the fact that $\|V\| = 1$ since V is orthogonal. The remainder of the proof is devoted to provide an upper bound on $\max_{1\leq i\leq Nd} \|(\tilde{T}_i)^k\|$.

Let $\gamma_{i,\pm} := \frac{(1+\beta)\mu_i \pm \sqrt{(1+\beta)^2\mu_i^2 - 4\beta\mu_i}}{2}$ be the eigenvalues of $\tilde{T}_i$.

(i) If $\gamma_{i,+} \neq \gamma_{i,-}$, then by the formula of k-th power of 2×2 matrix with distinct eigenvalues (see e.g. Williams (1992)), we get

$$\left(\tilde{T}_i\right)^k = \frac{\gamma_{i,+}^k}{\gamma_{i,+} - \gamma_{i,-}} \left(\tilde{T}_i - \gamma_{i,-} I\right) + \frac{\gamma_{i,-}^k}{\gamma_{i,-} - \gamma_{i,+}} \left(\tilde{T}_i - \gamma_{i,+} I\right).$$

This implies that

$$\left\| \left(\tilde{T}_i\right)^k \right\| \leq \frac{\max\left\{ \left\|\tilde{T}_i - \gamma_{i,-} I\right\|, \left\|\tilde{T}_i - \gamma_{i,+} I\right\| \right\}}{|\gamma_{i,+} - \gamma_{i,-}|} \max\left\{|\gamma_{i,+}|, |\gamma_{i,-}|\right\}^k.$$

We can compute that

$$\tilde{T}_i - \gamma_{i,-} I = \begin{bmatrix} \gamma_{i,+} & -\gamma_{i,+}\gamma_{i,-} \\ 1 & -\gamma_{i,-} \end{bmatrix},$$

which implies that

$$\left\|\tilde{T}_i - \gamma_{i,-}I\right\| \leq \left\|\begin{pmatrix} \gamma_{i,+} \\ 1 \end{pmatrix}\right\| \left\|\begin{pmatrix} 1 & -\gamma_{i,-} \end{pmatrix}\right\| \leq 1 + \max\{|\gamma_{i,+}|, |\gamma_{i,-}|\}^2.$$

Similarly, we have

$$\left\|\tilde{T}_i - \gamma_{i,+}I\right\| \leq \left\|\begin{pmatrix} \gamma_{i,-} \\ 1 \end{pmatrix}\right\| \left\|\begin{pmatrix} 1 & -\gamma_{i,+} \end{pmatrix}\right\| \leq 1 + \max\{|\gamma_{i,+}|, |\gamma_{i,-}|\}^2.$$

(ii) If $\gamma_{i,+} = \gamma_{i,-} = \frac{(1+\beta)\mu_i}{2}$, then by the formula for k-th power of 2×2 matrix with two identical eigenvalues (see e.g. Williams (1992)), we get

$$\left(\tilde{T}_i\right)^k = \left(\frac{(1+\beta)\mu_i}{2}\right)^{k-1} \left(k\tilde{T}_i - (k-1)\frac{(1+\beta)\mu_i}{2}I\right),$$

so that

$$\left\|\left(\tilde{T}_i\right)^k\right\| \leq \left(\frac{(1+\beta)|\mu_i|}{2}\right)^{k-1} \left(k\|\tilde{T}_i\| + (k-1)\frac{(1+\beta)|\mu_i|}{2}\right).$$

Also notice that $\mu_i = 0$, $\gamma_{i,+} = \gamma_{i,-} = 0$ and $\tilde{T}_i^k = 0$ for every $k \geq 2$.

Hence, we get

$$\max_{1\leq i\leq Nd} \left\|\left(\tilde{T}_i\right)^k\right\| \leq C_k \cdot \rho_{\text{dasg}}^k,$$

where

$$C_k := \max\left\{k \max_{i:\gamma_{i,+}=\gamma_{i,-},\mu_i\neq 0} \frac{2\|\tilde{T}_i\|}{(1+\beta)|\mu_i|} + k - 1, \max_{i:\gamma_{i,+}\neq\gamma_{i,-}} \frac{1 + \max\{|\gamma_{i,+}|, |\gamma_{i,-}|\}^2}{|\gamma_{i,+} - \gamma_{i,-}|}\right\},$$

and

$$\rho_{\text{dasg}} = \max_{1\leq i\leq Nd} \max\{|\gamma_{i,+}|, |\gamma_{i,-}|\}.$$

Moreover, when $\mu_i \neq 0$,

$$\left\|\tilde{T}_i\right\| = \max\left\{|\gamma_{i,-}|, |\gamma_{i,+}|\right\} = \frac{1+\beta}{2}\mu_i.$$

Hence

$$C_k := \max\left\{2k - 1, \max_{i:\gamma_{i,+}\neq\gamma_{i,-}} \frac{1 + \max\{|\gamma_{i,+}|, |\gamma_{i,-}|\}^2}{|\gamma_{i,+} - \gamma_{i,-}|}\right\}.$$

Next, let us assume that $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$, $\lambda_N^W > 0$ and $\alpha \in (0, \frac{\lambda_N^W}{L}]$. Therefore, we get

$$0 \leq \lambda_N^W - \alpha L \leq \mu_i \leq 1 - \alpha\mu.$$

When $0 < \mu_i < 1 - \alpha\mu$, we claim that

$$\Delta_i := (1+\beta)^2\mu_i^2 - 4\beta\mu_i < 0.$$

To see this, note that since $\mu_i > 0$ it is equivalent to

$$\mu_i < \frac{4\beta}{(1+\beta)^2} = \frac{4\frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}}{(\frac{2}{1+\sqrt{\alpha\mu}})^2} = 1-\alpha\mu.$$

Therefore, when $0 < \mu_i < 1-\alpha\mu$, we have $\Delta_i < 0$ and both $\gamma_{i,+}$ and $\gamma_{i,-}$ are complex numbers. In this case,

$$|\gamma_{i,-}| = |\gamma_{i,+}| = \sqrt{\beta\mu_i},$$

and

$$\begin{aligned}\max_{i:\gamma_{i,+}\neq\gamma_{i,-}} \frac{1+\max\{|\gamma_{i,+}|,|\gamma_{i,-}|\}^2}{|\gamma_{i,+}-\gamma_{i,-}|} &= \max_{i:0<\mu_i<1-\alpha\mu} \frac{1+\beta\mu_i}{\sqrt{-(1+\beta)^2\mu_i^2+4\beta\mu_i}} \\ &= \max_{i:0<\mu_i<1-\alpha\mu} \frac{1+\sqrt{\alpha\mu}+(1-\sqrt{\alpha\mu})\mu_i}{2\sqrt{\mu_i(1-\alpha\mu-\mu_i)}}.\end{aligned}$$

Moreover, when $0 < \mu_i < 1-\alpha\mu$

$$|\gamma_{i,+}| = |\gamma_{i,-}| = \sqrt{\beta\mu_i} \leq \sqrt{\frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}(1-\alpha\mu)} = 1-\sqrt{\alpha\mu}.$$

Next, $\gamma_{i,-} = \gamma_{i,+}$ if and only if $\mu_i = 0$ or $\mu_i = 1-\alpha\mu$. When $\mu_i = 0$, $|\gamma_{i,-}| = |\gamma_{i,+}| = 0$, and when $\mu_i = 1-\alpha\mu$,

$$|\gamma_{i,-}| = |\gamma_{i,+}| = \frac{(1+\beta)\mu_i}{2} \leq \frac{1}{1+\sqrt{\alpha\mu}}(1-\alpha\mu) = 1-\sqrt{\alpha\mu}.$$

Hence, we conclude that when $\beta = \frac{1-\sqrt{\alpha\mu}}{1+\sqrt{\alpha\mu}}$, $\lambda_N^W > 0$ and $\alpha \in (0, \frac{\lambda_N^W}{L}]$, we have $\rho_{\text{dasg}} = 1-\sqrt{\alpha\mu}$, and

$$C_k = \max\left\{2k-1, \max_{i:0<\mu_i<1-\alpha\mu} \frac{1+\sqrt{\alpha\mu}+(1-\sqrt{\alpha\mu})\mu_i}{2\sqrt{\mu_i(1-\alpha\mu-\mu_i)}}\right\}.$$

The proof is complete. ■

F.3 Proofs of Technical Results in Appendix E

F.3.1 Proof of Lemma 43

Proof The proof follows from Lemma 6 and Lemma 7 in Gürbüzbalaban et al. (2021). The difference is that in our case, by applying the proof of Theorem 42

$$\mathbb{E}\left\|\nabla F\left(x^{(k)}\right)\right\|^2 \leq L^2\mathbb{E}\left\|x^{(k)}-x^\infty\right\|^2 \leq D_1^2, \tag{177}$$

where

$$D_1^2 := L^2\mathbb{E}\left\|x^{(0)}-x^\infty\right\|^2 + L^2\frac{2\alpha}{\mu}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N, \tag{178}$$

and moreover, by (112) and the proof of Theorem 42, we get

$$\begin{aligned}\mathbb{E}\left[\left\|\xi^{(k+1)}\right\|^2\right] &\leq \left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N+\eta^2\mathbb{E}\left\|x^{(k)}-x^\infty\right\|^2\\ &\leq \left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N+\eta^2\left(\mathbb{E}\left\|x^{(0)}-x^\infty\right\|^2+\frac{2\alpha}{\mu}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N\right)\\ &= \left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)\frac{\mu+2\alpha}{\mu}N+\eta^2\mathbb{E}\left\|x^{(0)}-x^\infty\right\|^2.\end{aligned}$$

The rest of the proof is similar to Lemma 6 and Lemma 7 in Gürbüzbalaban et al. (2021) and is omitted here. ■

F.3.2 Proof of Lemma 47

Proof The proof follows from Lemma 27 and Lemma 28. The difference is that in our case, by applying Theorem 46

$$\begin{aligned}\mathbb{E}\left\|\nabla F\left(y^{(k)}\right)\right\|^2 &\leq 4L^2(1+\beta)^2\mathbb{E}\left\|x^{(k)}-x^\infty\right\|^2+4L^2\beta^2\mathbb{E}\left\|x^{(k-1)}-x^\infty\right\|^2+2\|\nabla F(x^*)\|^2\\ &\leq \tilde{D}_y^2,\end{aligned}\tag{179}$$

where we recall that $\tilde{D}_y^2$ is defined in (130) as follows:

$$\tilde{D}_y^2=4L^2\left((1+\beta)^2+\beta^2\right)\left(\frac{2V_{Q,\alpha}\left(\xi_0\right)}{\mu}+\frac{12\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N\right)+2\|\nabla F(x^*)\|^2,\tag{180}$$

and moreover, by applying (112) to the context of D-ASG and Theorem 46, we get

$$\begin{aligned}&\mathbb{E}\left[\left\|\xi^{(k+1)}\right\|^2\right]\\ &\leq \left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N+\eta^2\mathbb{E}\left\|y^{(k)}-x^\infty\right\|^2\\ &\leq \left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N+2\eta^2(1+\beta)^2\mathbb{E}\left\|x^{(k)}-x^\infty\right\|^2+2\eta^2\beta^2\mathbb{E}\left\|x^{(k-1)}-x^\infty\right\|^2\\ &\leq \left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N\\ &\qquad+2\eta^2\left((1+\beta)^2+\beta^2\right)\left(\frac{2V_{Q,\alpha}\left(\xi_0\right)}{\mu}+\frac{12\sqrt{\alpha}}{\mu\sqrt{\mu}}\left(\sigma^2+\eta^2\frac{\alpha^2(C_1)^2}{(1-\gamma)^2}\right)N\right).\end{aligned}$$

The rest of the proof is similar to Lemma 27 and Lemma 28 and is omitted here. ■

F.3.3 Proof of Lemma 48

Proof The functions $V_{\bar{Q},\alpha}$ and $V_{\bar{S},\alpha}$ have a similar structure and can be written as the sum of a quadratic term and a term involving the objective. Therefore, the proof of Lemma 30

applies with minor modifications. We next provide the details for the sake of completeness. We start with observing that the function $V_{\bar{Q},\alpha}$ is $\bar{M}_\alpha$ smooth where $\bar{M}_\alpha = L + 2\|\bar{Q}_\alpha\|$ and

$$\|\bar{Q}_\alpha\| = \left\|\tilde{Q}_\alpha\right\| \leq \|\bar{S}_\alpha\| + 2\alpha\eta^2\left((1+\beta)^2+\beta^2\right) = \frac{1}{\alpha} + \frac{\mu}{2} - \frac{\sqrt{\mu}}{\sqrt{\alpha}} + 2\alpha\eta^2\left((1+\beta)^2+\beta^2\right),$$

where we used (173). For any $\xi, \Delta \in \mathbb{R}^{2d}$, by the $\bar{M}_\alpha$-smoothness of $V_{\bar{Q},\alpha}$ we have also

$$\begin{aligned} V_{\bar{Q},\alpha}(\xi+\Delta) &\leq V_{\bar{Q},\alpha}(\xi) + \left\langle \nabla V_{\bar{Q},\alpha}(\xi), \Delta \right\rangle + \frac{\bar{M}_\alpha}{2}\|\Delta\|^2 \\ &\leq V_{\bar{Q},\alpha}(\xi) + m_1\|\nabla V_{\bar{S},\alpha}(\xi)\|^2 + \|\Delta\|^2/(4m_1) + \frac{\bar{M}_\alpha}{2}\|\Delta\|^2, \end{aligned} \tag{181}$$

for any $m_1 > 0$ where we used Cauchy-Schwarz inequality in (181). Note that the matrix $\tilde{Q}_\alpha$ has some special structure as a sum of two rank-one matrices, i.e. we can write

$$\tilde{Q}_\alpha = \sigma_1 a a^T + \sigma_2 b b^T,$$

with

$$\sigma_1 := \frac{1}{\alpha} + \frac{\mu}{2} - \frac{\sqrt{\mu}}{\sqrt{\alpha}}, \qquad a := \frac{1}{\sqrt{\sigma_1}} \begin{bmatrix} \sqrt{\frac{1}{2\alpha}} \\ \sqrt{\frac{\mu}{2}} - \sqrt{\frac{1}{2\alpha}} \end{bmatrix},$$

$$\sigma_2 := 2\alpha\eta^2\left((1+\beta)^2+\beta^2\right), \qquad b := \frac{1}{\sqrt{(1+\beta)^2+\beta^2}} \begin{bmatrix} 1+\beta \\ -\beta \end{bmatrix},$$

where a and b are both unit vectors with norm $\|a\| = \|b\| = 1$. Using the definition of β, one can see that a cannot be equal to b up to a multiplicative constant. Therefore, $\bar{Q}_\alpha$ cannot be of rank one; and has to be of rank two and hence is positive definite. In other words, the smallest eigenvalue m_2 of $\tilde{Q}_\alpha$ is positive. In this case, we have

$$\frac{1}{m_2}\tilde{Q}_\alpha \succeq I.$$

We have also

$$\begin{aligned} \left\|\nabla V_{\bar{Q},\alpha}(\xi)\right\|^2 &\leq 2\left\|2\bar{Q}_\alpha\xi\right\|^2 + 2\left\|\nabla f\left(\bar{T}\xi + x_*\right)\right\|^2 \\ &\leq 8\xi^T\bar{Q}_\alpha^2\xi + 2L^2\left\|\bar{T}\xi\right\|^2 \\ &\leq 8\xi^T\bar{Q}_\alpha^2\xi + 2L^2\frac{f(\bar{T}\xi + x_*) - f(x_*)}{\mu} \qquad (182) \\ &\leq m_3 V_{\bar{Q},\alpha}(\xi), \qquad (183) \end{aligned}$$

where

$$m_3 := 2\max\left(\frac{4}{m_2}, \frac{L^2}{\mu}\right).$$

Therefore, if choose $m_1 = \epsilon/m_3$; then we get from (181)

$$V_{\bar{Q},\alpha}(\xi+\Delta) \leq (1+\epsilon)V_{\bar{S},\alpha}(\xi) + M_\epsilon\|\Delta\|^2, \tag{184}$$

with $M_\epsilon = \frac{m_3}{4\epsilon} + \bar{M}_\alpha/2$. Finally, if we choose $\xi = \bar{\xi}_{k+1} - D_{k+1}$ and $\Delta = D_{k+1}$; we obtain (128). This completes the proof. ■

References

A. Agarwal and J. C. Duchi. Distributed delayed stochastic optimization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 873–881. Curran Associates, Inc., 2011.

S. A. Alghunaim and A. H. Sayed. Distributed coupled learning over adaptive networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6353–6357. IEEE, 2018.

S. A. Alghunaim and A. H. Sayed. Distributed coupled multi-agent stochastic optimization. *IEEE Transactions on Automatic Control*, 65(1):175–190, 2020.

Y. Arjevani and O. Shamir. On the iteration complexity of oblivious first-order optimization algorithms. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 908–916, New York, New York, USA, 20–22 Jun 2016. PMLR.

Y. Arjevani, J. Bruna, B. Can, M. Gurbuzbalaban, S. Jegelka, and H. Lin. IDEAL: Inexact DEcentralized accelerated augmented Lagrangian method. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

N. S. Aybat, A. Fallah, M. Gurbuzbalaban, and A. Ozdaglar. A universally optimal multistage accelerated stochastic gradient method. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019.

N. S. Aybat, A. Fallah, M. Gürbüzbalaban, and A. Ozdaglar. Robust accelerated gradient methods for smooth strongly convex functions. *SIAM Journal on Optimization*, 30(1): 717–751, 2020.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

D. Blatt, A. O. Hero, and H. Gauchman. A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51, 2007.

S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE/ACM Transactions on Networking (TON)*, 14(SI):2508–2530, 2006.

B. Can, M. Gürbüzbalaban, and L. Zhu. Accelerated linear convergence of stochastic momentum methods in Wasserstein distances. In *Proceedings of the 34th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 891–901. PMLR, 2019.

A. Chapman. *Semi-Autonomous Networks: Effective Control of Networked Systems Through Protocols, Design, and Modeling*. Springer, 2015.

F. R. Chung. *Spectral Graph Theory*. American Mathematical Society, Providence, Rhode Island, 1997.

A. d'Aspremont. Smooth optimization with approximate gradient. *SIAM Journal on Optimization*, 19(3):1171–1183, 2008.

P. Di Lorenzo and G. Scutari. Distributed nonconvex optimization over networks. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 229–232. IEEE, 2015.

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

A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *Annals of Statistics*, 48(3):1348–1382, 2020.

J. C. Duchi, A. Agarwal, and M. J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57 (3):592–606, 2012a.

J. C. Duchi, P. L. Bartlett, and M. J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012b.

D. Dvinskikh and A. Gasnikov. Decentralized and parallelized primal and dual accelerated methods for stochastic convex programming problems. *arXiv preprint arXiv:1904.09015*, 2019.

M. Fazlyab, A. Ribeiro, M. Morari, and V. Preciado. Analysis of optimization algorithms via integral quadratic constraints: Nonstrongly convex problems. *SIAM Journal on Optimization*, 28(3):2654–2689, 2018.

N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In *Conference on Learning Theory*, pages 658–695. PMLR, 2015.

M. Grant, S. Boyd, and Y. Ye. CVX: Matlab software for disciplined convex programming, 2008.

M. Gürbüzbalaban, X. Gao, Y. Hu, and L. Zhu. Decentralized stochastic gradient Langevin dynamics and Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 22:1–69, 2021.

I. Hakimi, S. Barkai, M. Gabel, and A. Schuster. DANA: Scalable out-of-the-box distributed ASGD without retuning, 2019.

M. Hardt. Robustness versus acceleration, Aug. 2014. URL `http://blog.mrtz.org/2014/08/18/robustness-versus-acceleration`.

B. Hu and L. Lessard. Dissipativity theory for Nesterov's accelerated method. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1549–1557, International Convention Centre, Sydney, Australia, 2017. PMLR.

M. Jaggi, V. Smith, M. Takác, J. Terhorst, S. Krishnan, T. Hofmann, and M. I. Jordan. Communication-efficient distributed dual coordinate ascent. In *Advances in Neural Information Processing Systems*, pages 3068–3076, 2014.

P. Jain, S. M. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating stochastic gradient descent for least squares regression. In *Conference On Learning Theory*, pages 545–604. PMLR, 2018.

D. Jakovetić. A unification and generalization of exact distributed first-order methods. *IEEE Transactions on Signal and Information Processing over Networks*, 5(1):31–46, 2019.

D. Jakovetić, J. Xavier, and J. M. Moura. Fast distributed gradient methods. *IEEE Transactions on Automatic Control*, 59(5):1131–1146, 2014.

K. Kaya, F. Öztoprak, Ş. İ. Birbil, A. T. Cemgil, U. Şimşekli, N. Kuru, H. Koptagel, and M. K. Öztürk. A framework for parallel second order incremental optimization algorithms for solving partially separable problems. *Computational Optimization and Applications*, 72(3):675–705, 2019.

A. Koloskova, S. U. Stich, and M. Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3478–3487. PMLR, 2019.

J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

G. Lan, S. Lee, and Y. Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *Mathematical Programming*, 180:237–284, 2020.

C.-p. Lee, C. H. Lim, and S. J. Wright. A distributed quasi-Newton algorithm for empirical risk minimization with nonsmooth regularization. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1646–1655. ACM, 2018.

L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

D. A. Levin, Y. Peres, and E. L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, Rhode Island, 2009.

H. Li, C. Fang, W. Yin, and Z. Lin. Decentralized accelerated gradient methods with increasing penalty parameters. *IEEE Transactions on Signal Processing*, 68:4855–4870, 2020.

F. Mansoori and E. Wei. Superlinearly convergent asynchronous distributed network Newton method. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 2874–2879. IEEE, 2017.

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

Q. Meng, W. Chen, J. Yu, T. Wang, Z. Ma, and T.-Y. Liu. Asynchronous accelerated stochastic gradient descent. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1853–1859, 2016.

K. Mishchenko, F. Iutzeler, J. Malick, and M.-R. Amini. A delay-tolerant proximal-gradient algorithm for distributed learning. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3587–3595, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

A. Mokhtari and A. Ribeiro. DSA: Decentralized double stochastic averaging gradient algorithm. *Journal of Machine Learning Research*, 17(1):2165–2199, 2016.

A. Nedic and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

A. Nedić, A. Olshevsky, and M. G. Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5): 953–976, 2018.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer, 2003.

M. Pirani, E. M. Shahrivar, B. Fidan, and S. Sundaram. Robustness of leader-follower networked dynamical systems. *IEEE Transactions on Control of Network Systems*, 5(4): 1752–1763, 2018.

S. Pu and A. Nedić. A distributed stochastic gradient tracking method. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 963–968. IEEE, 2018.

S. Pu and A. Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187:409–457, 2021.

S. Pu, A. Olshevsky, and I. C. Paschalidis. A sharp estimate on the transient time of distributed stochastic gradient descent. *IEEE Transactions on Automatic Control*, 2021. Forthcoming.

G. Qu and N. Li. Accelerated distributed nesterov gradient descent for smooth and strongly convex functions. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 209–216. IEEE, 2016.

G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2018.

G. Qu and N. Li. Accelerated distributed Nesterov gradient descent. *IEEE Transactions on Automatic Control*, 65(6):2566–2581, 2020.

M. Rabbat. Multi-agent mirror descent for decentralized stochastic optimization. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 517–520. IEEE, 2015.

M. Raginsky, A. Rakhlin, and M. Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703, Amsterdam, Netherlands, 07–10 Jul 2017. PMLR.

T. Sarkar, M. Roozbehani, and M. A. Dahleh. Asymptotic network robustness. *IEEE Transactions on Control of Network Systems*, 6(2):812–821, 2018.

K. Scaman, F. Bach, S. Bubeck, L. Massoulié, and Y. T. Lee. Optimal algorithms for non-smooth distributed optimization in networks. In *Advances in Neural Information Processing Systems*, pages 2740–2749, 2018.

K. Scaman, F. Bach, S. Bubeck, Y. T. Lee, and L. Massoulié. Optimal convergence rates for convex distributed optimization in networks. *Journal of Machine Learning Research*, 20(159):1–31, 2019.

M. Schmidt, R. Babanezhad, M. Ahmed, A. Defazio, A. Clifton, and A. Sarkar. Non-uniform stochastic average gradient method for training conditional random fields. In *Artificial Intelligence and Statistics*, pages 819–828, 2015.

O. Shamir and N. Srebro. Distributed stochastic optimization and learning. In *2014 52nd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 850–857. IEEE, 2014.

W. Shi, Q. Ling, G. Wu, and W. Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

U. Şimşekli, Ç. Yıldız, T. H. Nguyen, G. Richard, and A. T. Cemgil. Asynchronous stochastic quasi-Newton MCMC for non-convex optimization. In *International Conference on Machine Learning*, pages 4674–4683. PMLR, 2018.

A. Sundararajan, B. Hu, and L. Lessard. Robust convergence analysis of distributed optimization algorithms. In *2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1206–1212. IEEE, 2017.

A. Sundararajan, B. Van Scoy, and L. Lessard. A canonical form for first-order distributed optimization algorithms. In *2019 American Control Conference (ACC)*, pages 4075–4080. IEEE, 2019.

A. Sundararajan, B. Van Scoy, and L. Lessard. Analysis and design of first-order distributed optimization algorithms over time-varying graphs. *IEEE Transactions on Control of Network Systems*, 7(4):1597–1608, 2020.

T. M. D. Tran and A. Y. Kibangou. Distributed estimation of graph Laplacian eigenvalues by the alternating direction of multipliers method. *IFAC Proceedings Volumes*, 47(3): 5526–5531, 2014.

K. I. Tsianos and M. G. Rabbat. Distributed strongly convex optimization. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 593–600. IEEE, 2012.

C. A. Uribe, S. Lee, A. Gasnikov, and A. Nedić. A dual approach for optimal algorithms in distributed optimization over networks. *Optimization Methods and Software*, 36(1): 171–210, 2021.

K. S. Williams. The nth power of a 2×2 matrix. *Mathematics Magazine*, 65(5):336–336, 1992.

C. Xi, R. Xin, and U. A. Khan. Add-opt: Accelerated distributed directed optimization. *IEEE Transactions on Automatic Control*, 63(5):1329–1339, 2017.

R. Xin and U. A. Khan. Distributed heavy-ball: A generalization and acceleration of first-order methods with gradient tracking. *IEEE Transactions on Automatic Control*, 65(6): 2627–2633, 2020.

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

K. Zhou, J. C. Doyle, and K. Glover. *Robust and Optimal Control*, volume 40. Prentice Hall New Jersey, 1996.