

---

# A Stochastic Approximation Approach for Efficient Decentralized Optimization on Random Networks

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 A challenging problem in decentralized optimization is to develop algorithms with  
2 fast convergence on random and time varying topologies under unreliable and  
3 bandwidth-constrained communication network. This paper studies a stochastic  
4 approximation approach with a Fully Stochastic Primal Dual Algorithm (FSPDA)  
5 framework. Our framework relies on a novel observation that randomness in time  
6 varying topology can be incorporated in a stochastic augmented Lagrangian for-  
7 mulation, whose expected value admits saddle points that coincide with stationary  
8 solutions of the decentralized optimization problem. With the FSPDA framework,  
9 we develop two new algorithms supporting efficient sparsified communication on  
10 random time varying topologies — FSPDA-SA allows agents to execute multiple  
11 local gradient steps depending on the time varying topology to accelerate conver-  
12 gence, and FSPDA-STORM further incorporates a variance reduction step to improve  
13 sample complexity. For problems with smooth (possibly non-convex) objective  
14 function, within  $T$  iterations, we show that FSPDA-SA (resp. FSPDA-STORM) finds  
15 an  $\mathcal{O}(1/\sqrt{T})$ -stationary (resp.  $\mathcal{O}(1/T^{2/3})$ ) solution. Numerical experiments show  
16 the benefits of the FSPDA algorithms.

## 17 1 Introduction

18 Consider  $n$  agents that communicate on an undirected and connected graph/network  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  with  
19  $\mathcal{V} = [n] := \{1, \dots, n\}$ ,  $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ . Each agent  $i \in [n]$  has access to a continuously differentiable  
20 (possibly non-convex) local objective function  $f_i : \mathbb{R}^d \rightarrow \mathbb{R}$  and maintains a local decision variable  
21  $\mathbf{x}_i \in \mathbb{R}^d$ . Denote  $\mathbf{x} = [\mathbf{x}_1^\top, \dots, \mathbf{x}_n^\top]^\top \in \mathbb{R}^{nd}$ . Our aim is to tackle:

$$\min_{\mathbf{x} \in \mathbb{R}^{nd}} \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x}_i) \quad \text{s.t.} \quad \mathbf{x}_i = \mathbf{x}_j, \forall (i, j) \in \mathcal{E}. \quad (1)$$

22 In other words, (1) seeks a  $\mathbf{x}^* \in \mathbb{R}^d$  that minimizes  $F(\mathbf{x}) := (1/n) \sum_{i=1}^n f_i(\mathbf{x})$ . We are interested  
23 in the stochastic optimization setting where each  $f_i(\mathbf{x}_i)$  is given by (with slight abuse of notation)

$$f_i(\mathbf{x}_i) := \mathbb{E}_{\xi_i \sim \mathbb{P}_i} [f_i(\mathbf{x}_i; \xi_i)] \quad (2)$$

24 where  $\mathbb{P}_i$  represents the  $i$ -th data distribution. Problem (1) is relevant to the distributed learning  
25 problem especially in the decentralized case where a central server is absent. Prior works [Nedic and  
26 Ozdaglar, 2009, Lian et al., 2017, Nedic et al., 2017, Qu and Li, 2017] demonstrated that *decentralized*  
27 algorithms can tackle (1) efficiently through repeated message exchanges among the neighbors and  
28 local stochastic gradient updates.

29 Towards an efficient decentralized algorithm for (1), an important direction is to consider a *time*  
30 *varying graph topology* setting where the *active edge set* in  $\mathcal{G}$  changes over time. This is a generic  
31 setting covering cases when the communication links are unreliable, or the agents choose not to  
32 communicate in a certain round (a.k.a. local updates) [Koloskova et al., 2019a, Nadiradze et al., 2021].

Prior Works	SG	TV	w/o BH	Rate
Prox-GPDA [Hong et al., 2017]	✗	✗	✓	Asympt.
NEXT [Lorenzo and Scutari, 2016]	✗	✓	✓	Asympt.
DSGD [Koloskova et al., 2020]	✓	✓	✗	$\mathcal{O}(\sigma/\sqrt{nT})$
Swarm-SGD [Nadiradze et al., 2021]	✓	✓	✗	$\mathcal{O}(\sigma^2/\sqrt{T})$
CHOCO-SGD [Koloskova et al., 2019a]	✓	✗ <sup>‡</sup>	✗	$\mathcal{O}(\sigma/\sqrt{nT})$
Decen-Scaffnew [Mishchenko et al., 2022]	✓	✗ <sup>†</sup>	✓	$\mathcal{O}(\sigma/\sqrt{nT})$
Local-GT [Liu et al., 2024]	✓	✗ <sup>†</sup>	✓	$\mathcal{O}(\sigma/\sqrt{nT})$
LED [Alghunaim, 2024]	✓	✗ <sup>†</sup>	✓	$\mathcal{O}(\sigma/\sqrt{nT})$
FSPDA-SA (This Work)	✓	✓	✓	$\mathcal{O}(\sigma/\sqrt{nT})$
FSPDA-STORM (This Work)	✓	✓	✓	$\mathcal{O}(\sigma^{2/3}/T^{2/3})$

Table 1: Comparison of decentralized algorithms for **non-convex** optimization. In the table, ‘SG’ is ‘Stochastic Gradient’, ‘TV’ is ‘Time Varying Graph’, ‘w/o BH’ is ‘Without Bounded Heterogeneity’, and ‘Rate’ is the expected squared gradient norm  $\mathbb{E}[\|\nabla F(\bar{\mathbf{x}})\|^2]$  after  $T$  iterations. Note that  $\sigma^2$  is the variance of stochastic gradient. <sup>‡</sup>CHOCO-SGD incorporates broadcast gossip as a special case of compression. <sup>†</sup>ProxSkip, Local-GT, LED consider local updates with periodic communication.

By assuming that a random topology is drawn at each iteration, the convergence of decentralized stochastic gradient (DSGD) has been studied in [Lobel and Ozdaglar, 2010, Nadiradze et al., 2021] and is later on unified by [Koloskova et al., 2020] with tighter bounds for local updates, periodic sampling, etc. An alternative [Ram et al., 2010] is to analyze DSGD for the  $B$ -connectivity setting which requires the union of every  $B$  consecutive time varying topologies to yield a connected graph. Nevertheless, these works focused on vanilla DSGD that may have slow convergence (in transient stage) and is limited to bounded data heterogeneity. The prior restrictions can be relaxed using advanced algorithms such as gradient tracking [Qu and Li, 2017], EXTRA [Shi et al., 2015] and primal-dual framework [Hong et al., 2017, Hajinezhad and Hong, 2019, Yi et al., 2021].

As noted by [Koloskova et al., 2021], analyzing the convergence of sophisticated algorithms with time varying topology, such as gradient tracking [Qu and Li, 2017] is challenging due to the non-symmetric product of two (or more) mixing matrices. Existing works considered various restrictions on the time varying topology  $\mathcal{G}^{(t)} = (\mathcal{V}, \mathcal{E}^{(t)})$  and/or the problem (1): [Koloskova et al., 2021, Liu et al., 2024] studied gradient tracking with local updates that essentially takes  $\mathcal{E}^{(t)} = \mathcal{E}$  periodically and  $\mathcal{E}^{(t)} = \emptyset$  otherwise, also see [Mishchenko et al., 2022, Guo et al., 2023, Alghunaim, 2024] for a similar result and note that such algorithms require extra synchronization overhead; [Kovalev et al., 2021, 2024] considered a setting where  $\mathcal{G}^{(t)}$  is connected for any  $t$ ; [Nedic et al., 2017, Li and Lin, 2024] focused on (accelerated) gradient tracking with deterministic gradient when  $F(\mathbf{x})$  is (strongly) convex; [Lorenzo and Scutari, 2016] also considered deterministic gradient with possibly non-convex  $F(\mathbf{x})$  but only provides asymptotic convergence guarantees; [Lei et al., 2018, Yau and Wai, 2023] considered asymptotic convergence guarantees in the case of strictly (or strongly) convex  $F(\mathbf{x})$ . We provide a non-exhaustive list summarizing the convergence of existing works in Table 1.

The above discussion highlights a gap in the existing literature —

*Is there any algorithm that achieves fast convergence on time varying (random) topology?*

This paper gives an affirmative answer through developing the Fully Stochastic Primal Dual Algorithm (FSPDA) framework that leads to efficient decentralized algorithms tackling (1) in its general form. The framework features the design of a new stochastic augmented Lagrangian function.

As pointed out by [Chang et al., 2020], many decentralized algorithms (including gradient tracking) can be interpreted as primal-dual algorithms finding a saddle point of the augmented Lagrangian function. However, its extension to time varying topology is not straightforward due to the inconsistency in dual variables updates. To overcome this challenge, we propose a stochastic equality constrained reformulation of (1) to model randomness in topology. Then, the latter yields a stochastic augmented Lagrangian function. Applying stochastic approximation (SA) to solve the latter leads to the FSPDA framework. Our contributions are

- We propose two new algorithms: (i) FSPDA-SA is derived by vanilla SA that applies primal-dual stochastic gradient descent-ascent on the stochastic augmented Lagrangian, (ii) FSPDA-STORM uses an additional control variate / momentum term to reduce the drift term's variance in a recursive manner. Both algorithms are fully stochastic as the random time varying topology is treated as a part of randomness. Additionally, our framework supports sparsified communication, i.e., the agents can choose to communicate a subset of primal coordinates at each iteration.
- We show that after  $T$  iterations, FSPDA-SA (resp. FSPDA-STORM) finds in expectation a solution whose squared gradient norm is  $\mathcal{O}(1/\sqrt{T})$  (resp.  $\mathcal{O}(1/T^{2/3})$ ). The convergence analysis is derived from a new Lyapunov function design that involves an unsigned inner product term and incorporates a variance condition on the random time varying topologies. Interestingly, we show empirically that using momentum in dual updates benefits the consensus error convergence.
- We also demonstrate that both FSPDA-SA and FSPDA-STORM can be implemented in a fully asynchronous manner, i.e., the agents can communicate and compute at different time slots, and supports local update as the algorithms allow for arbitrary time varying topology. That said, we remark that the convergence rates with local updates of FSPDA-SA and FSPDA-STORM are only suboptimal.

We provide numerical experiments to show that FSPDA-SA and FSPDA-STORM outperform existing algorithms in terms of iteration and communication complexity.

**Notations.** Let  $\mathbf{W} \in \mathbb{R}^{d \times d}$  be a symmetric (not necessarily positive semidefinite) matrix, the  $\mathbf{W}$ -weighted (semi) inner product of vectors  $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$  is denoted as  $\langle \mathbf{a} | \mathbf{b} \rangle_{\mathbf{W}} := \mathbf{a}^\top \mathbf{W} \mathbf{b}$ . Similarly, the  $\mathbf{W}$ -weighted (semi) norm is denoted by  $\|\mathbf{a}\|_{\mathbf{W}}^2 := \langle \mathbf{a} | \mathbf{a} \rangle_{\mathbf{W}}$ . The subscript notation is omitted for  $\mathbf{I}$ -weighted inner products. For any square matrix  $\mathbf{X}$ ,  $(\mathbf{X})^\dagger$  denotes its pseudo inverse.

## 2 The Fully Stochastic Primal Dual Algorithm (FSPDA) Framework

This section develops the FSPDA framework for tackling (1) and describes two variants of the framework leading to decentralized stochastic optimization of (1). Let  $\tilde{\mathbf{A}} \in \{-1, 0, 1\}^{|\mathcal{E}| \times n}$  be an incidence matrix of  $\mathcal{G}$ . By defining  $\mathbf{A} = \tilde{\mathbf{A}} \otimes \mathbf{I}_d \in \{-1, 0, 1\}^{|\mathcal{E}|d \times nd}$ , we observe that the consensus constraint in (1) is equivalent to  $\mathbf{A}\mathbf{x} = \mathbf{0}$ .

Our first step is to model the randomness in the time varying topology using the random variable (r.v.)  $\xi_a \sim \mathbb{P}_a$ . For each realization  $\xi_a$ , we define the random incidence matrix  $\mathbf{A}(\xi_a) := \mathbf{I}(\xi_a)\mathbf{A} \in \{-1, 0, 1\}^{|\mathcal{E}|d \times nd}$  where  $\mathbf{I}(\xi_a) \in \{0, 1\}^{|\mathcal{E}|d \times |\mathcal{E}|d}$  is a binary diagonal matrix. In addition to selecting each edge of  $\mathcal{G}$  randomly,  $\mathbf{I}(\xi_a)$  selects a random subset of  $d$  coordinates. As we will see later, this allows our approach to simultaneously achieve random sparsification for communication compression.

Assume that  $\mathbb{E}_{\xi_a \sim \mathbb{P}_a}[\mathbf{I}(\xi_a)]$  is a positive diagonal matrix, (1) is equivalent to:

$$\min_{\mathbf{x} \in \mathbb{R}^{nd}} \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\xi_i \sim \mathbb{P}_i} [f_i(\mathbf{x}_i; \xi_i)] \quad \text{s.t.} \quad \mathbb{E}_{\xi_a \sim \mathbb{P}_a} [\mathbf{A}(\xi_a)] \mathbf{x} = \mathbf{0}. \quad (3)$$

Denote  $\xi = (\xi_1, \dots, \xi_n, \xi_a)$ , FSPDA hinges on the following *augmented Lagrangian* function of (3):

$$\begin{aligned} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) &:= \mathbb{E}_{\xi} [\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}; \xi)] \\ \text{with } \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}; \xi) &:= \sum_{i=1}^n f_i(\mathbf{x}_i; \xi_i) + \tilde{\eta} \langle \boldsymbol{\lambda} | \mathbf{A}(\xi_a) \mathbf{x} \rangle + \frac{\tilde{\gamma}}{2} \|\mathbf{A}(\xi_a) \mathbf{x}\|^2, \end{aligned} \quad (4)$$

where  $\tilde{\eta} > 0, \tilde{\gamma} > 0$  are penalty parameters. It can be verified that the saddle points of  $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$  correspond to the KKT points of (1) [Bertsekas, 2016]. For brevity, in the rest of this paper, we may drop the subscript in  $\xi$  whenever the notation is clear from the context.

FSPDA is developed from applying stochastic approximation (SA) to seek a saddle point of (4). By recognizing  $\mathbf{A}(\xi)^\top \mathbf{A}(\xi) = \mathbf{A}^\top \mathbf{A}(\xi)$ , we consider the stochastic gradients:

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}; \xi) := \nabla \mathbf{f}(\mathbf{x}; \xi) + \tilde{\eta} \mathbf{A}^\top \boldsymbol{\lambda} + \tilde{\gamma} \mathbf{A}^\top \mathbf{A}(\xi) \mathbf{x}, \quad \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}; \xi) := \tilde{\eta} \mathbf{A}(\xi) \mathbf{x}, \quad (5)$$

where  $\nabla \mathbf{f}(\mathbf{x}; \xi) = [\nabla f_1(\mathbf{x}_1; \xi_1); \dots; \nabla f_n(\mathbf{x}_n; \xi_n)] \in \mathbb{R}^{nd}$ . Notice that to facilitate algorithm development, we have taken a deterministic  $\mathbf{A}$  for the term in  $\nabla_{\mathbf{x}} \mathcal{L}$  related to  $\boldsymbol{\lambda}$ . Now observe the  $i$ th  $d$ -dimensional block of  $\mathbf{A}^\top \mathbf{A}(\xi) \mathbf{x}$  which can be aggregated within  $\mathcal{N}_i(\xi)$  the neighborhood of the  $i$ th agent as:

$$[\mathbf{A}^\top \mathbf{A}(\xi) \mathbf{x}]_i = \sum_{j \in \mathcal{N}_i(\xi)} \mathbf{C}_{ij}(\xi) (\mathbf{x}_j - \mathbf{x}_i), \quad (6)$$

where  $\mathbf{C}_{ij}(\xi) \in \{0, 1\}^{d \times d}$  is diagonal and depends on the selected coordinates for the edge  $(i, j)$  under randomness  $\xi$ . Eq. (6) only relies on  $\mathbf{x}_j$  from neighbor  $j$  that is connected on the time varying

topology  $\mathcal{G}(\xi)$ . For illustration, an example of the above random graph model is given by Figure 3 in Appendix A. Importantly, (5) shows that with the stochastic augmented Lagrangian function, the time varying topology can be treated implicitly as a part of the randomness in the stochastic primal-dual gradients. The framework is thus described as being *fully stochastic* as in [Bianchi et al., 2021], and departs from [Liu et al., 2024, Alghunaim, 2024] that treat the topology as fixed during the derivation of primal-dual algorithm(s). From (5), (6), we derive *two* variants of FSPDA.

**FSPDA-SA Algorithm.** The first variant of FSPDA is derived from a direct application of stochastic gradient descent-ascent (SGDA) updates. Take  $\alpha > 0, \beta > 0$  as the step sizes, we have

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t; \xi^t), \quad \boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \beta \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t; \xi^t). \quad (7)$$

Taking the variable substitution  $\hat{\boldsymbol{\lambda}} := \mathbf{A}^\top \boldsymbol{\lambda}$  yields the following recursion:

**FSPDA-SA:** for any  $t \geq 0$  and any  $i \in [n]$ ,

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t - \alpha \nabla f_i(\mathbf{x}_i^t; \xi_i^t) - \eta \hat{\boldsymbol{\lambda}}_i^t + \gamma \sum_{j \in \mathcal{N}_i(\xi_a^t)} \mathbf{C}_{ij}(\xi_a^t)(\mathbf{x}_j^t - \mathbf{x}_i^t), \quad (8a)$$

$$\hat{\boldsymbol{\lambda}}_i^{t+1} = \hat{\boldsymbol{\lambda}}_i^t + \beta \sum_{j \in \mathcal{N}_i(\xi_a^t)} \mathbf{C}_{ij}(\xi_a^t)(\mathbf{x}_j^t - \mathbf{x}_i^t). \quad (8b)$$

Note that  $\mathbf{x}^0, \hat{\boldsymbol{\lambda}}^0$  can be initialized arbitrarily.

**FSPDA-STORM Algorithm.** The second variant of FSPDA reduces the variance of the stochastic gradient term in (5) using the recursive momentum variance reduction technique [Cutkosky and Orabona, 2019]. Herein, the key idea is to utilize a control variate in estimating the (primal-dual) gradients of  $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ . Take  $\alpha, \beta > 0$  and  $a_x, a_\lambda \in [0, 1]$  as the momentum parameters, we have  $\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha \mathbf{m}_x^t, \boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \beta \mathbf{m}_\lambda^t$  as the primal-dual updates, and

$$\begin{aligned} \mathbf{m}_x^{t+1} &= \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}; \xi^{t+1}) + (1 - a_x)(\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t; \xi^{t+1})), \\ \mathbf{m}_\lambda^{t+1} &= \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}; \xi^{t+1}) + (1 - a_\lambda)(\mathbf{m}_\lambda^t - \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t; \xi^{t+1})). \end{aligned} \quad (9)$$

The aim of  $\mathbf{m}_x^{t+1}$  is to estimate  $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1})$ . Now, instead of the straightforward estimator  $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}; \xi^{t+1})$ , we include an extra zero-mean term  $\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t; \xi^{t+1})$  to reduce the variance of the stochastic gradient estimation. The latter is a control variate that is computed recursively. Particularly, it has been shown in [Cutkosky and Orabona, 2019] that it can effectively reduce variance with a carefully designed parameter  $a_x$ , provided that the stochastic gradient map satisfies a mean-square Lipschitz condition. We summarize the algorithm as follows.

**FSPDA-STORM:** for any  $t \geq 0$  and any  $i \in [n]$ ,

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t - \alpha \mathbf{m}_{x,i}^t, \quad (10a)$$

$$\hat{\boldsymbol{\lambda}}_i^{t+1} = \hat{\boldsymbol{\lambda}}_i^t + \beta \mathbf{m}_{\lambda,i}^t, \quad (10b)$$

$$\begin{aligned} \mathbf{m}_{x,i}^{t+1} &= (1 - a_x) [\mathbf{m}_{x,i}^t + \nabla f_i(\mathbf{x}_i^t; \xi_i^{t+1}) - \eta \hat{\boldsymbol{\lambda}}_i^t + \gamma \sum_{j \in \mathcal{N}_i(\xi_a^{t+1})} \mathbf{C}_{ij}(\xi_a^{t+1})(\mathbf{x}_j^t - \mathbf{x}_i^t)] \\ &\quad + \nabla f_i(\mathbf{x}_i^{t+1}; \xi_i^{t+1}) - \eta \hat{\boldsymbol{\lambda}}_i^{t+1} + \gamma \sum_{j \in \mathcal{N}_i(\xi_a^{t+1})} \mathbf{C}_{ij}(\xi_a^{t+1})(\mathbf{x}_j^{t+1} - \mathbf{x}_i^{t+1}) \end{aligned} \quad (10c)$$

$$\begin{aligned} \mathbf{m}_{\lambda,i}^{t+1} &= (1 - a_\lambda) [\mathbf{m}_{\lambda,i}^t + \sum_{j \in \mathcal{N}_i(\xi_a^{t+1})} \mathbf{C}_{ij}(\xi_a^{t+1})(\mathbf{x}_j^t - \mathbf{x}_i^t)] \\ &\quad + \sum_{j \in \mathcal{N}_i(\xi_a^{t+1})} \mathbf{C}_{ij}(\xi_a^{t+1})(\mathbf{x}_j^{t+1} - \mathbf{x}_i^{t+1}) \end{aligned} \quad (10d)$$

Note that to achieve the theoretical performance (see later in Sec. 3),  $\mathbf{x}^0, \hat{\boldsymbol{\lambda}}^0, \mathbf{m}_x^0, \mathbf{m}_\lambda^0$  shall be initialized as  $\mathbf{x}_i^0 = \bar{\mathbf{x}}^0, \hat{\boldsymbol{\lambda}}_i^0 = (\alpha/\eta)n^{-1}(\nabla F(\bar{\mathbf{x}}^0) - \nabla f_i(\bar{\mathbf{x}}^0)), \mathbf{m}_{x,i}^0 = \nabla F(\bar{\mathbf{x}}^0), \mathbf{m}_{\lambda,i}^0 = \mathbf{0}$  according to (23). We remark that a simple initialization choice  $\hat{\boldsymbol{\lambda}}^0 = \mathbf{m}_{x,i}^0 = \mathbf{m}_{\lambda,i}^0 = \mathbf{0}$  works well in practice.

Both FSPDA-SA and FSPDA-STORM are decentralized algorithms that can be implemented on random time varying topology, and support randomized sparsification for further communication compression. The key is to observe that in (8), (10), the only information required for agent  $i$  is to obtain  $\sum_{j \in \mathcal{N}_i(\xi_a^t)} \mathbf{C}_{ij}(\xi_a^t)(\mathbf{x}_j^t - \mathbf{x}_i^t)$ , and in addition  $\sum_{j \in \mathcal{N}_i(\xi_a^t)} \mathbf{C}_{ij}(\xi_a^t)(\mathbf{x}_j^{t-1} - \mathbf{x}_i^{t-1})$  for FSPDA-STORM, at iteration  $t$ .

## 2.1 Implementation Details and Connection to Existing Works

We discuss several features of the FSPDA algorithms and their connections to existing works.

**Local & Asynchronous Updates.** The *local update* scheme where each agent  $i$  is allowed to update its own local variables  $\mathbf{x}_i, \lambda_i$  for multiple iterations without a communication step is a common practice in decentralized optimization [Liu et al., 2024, Li and Lin, 2024, Alghunaim, 2024, Mishchenko et al., 2022]. As discussed before, such scheme can be seen as a special case of the FSPDA framework where the time varying topology  $\mathcal{E}^{(t)}$  is chosen such that the latter alternates between  $\mathcal{E}^{(t)} = \mathcal{E}$  and  $\mathcal{E}^{(t)} = \emptyset$ .

Furthermore, FSPDA-SA allows for the general case of *asynchronous* updates. This is done so by taking the stochastic gradient as  $\nabla f_i(\mathbf{x}_i^t; \xi^t) = b_i(\xi^t) \bar{b}_i \nabla f_i(\mathbf{x}_i^t; \xi^t)$  such that  $b_i(\xi^t) \in \{0, 1\}$  with  $\mathbb{E}[b_i(\xi^t)] = 1/\bar{b}_i$  for some constant  $\bar{b}_i > 0$ . Detailed discussions for a fully asynchronous implementation of FSPDA-SA can be found in Appendix A.

**Connection to Existing Works.** Evaluating  $\mathbf{x}^{t+2} - \mathbf{x}^{t+1}$  from the FSPDA-SA sequence and observe that the combination of (8a) and (8b) is equivalent to the second order recursion:

$$\begin{aligned} \mathbf{x}^{t+2} = & 2 \left( \mathbf{I} - \frac{\gamma}{2} \mathbf{A}^\top \mathbf{A}(\xi^{t+1}) \right) \mathbf{x}^{t+1} - (\mathbf{I} - (\gamma - \eta\beta) \mathbf{A}^\top \mathbf{A}(\xi^t)) \mathbf{x}^t \\ & - \alpha (\nabla \mathbf{f}(\mathbf{x}^{t+1}; \xi^{t+1}) - \nabla \mathbf{f}(\mathbf{x}^t; \xi^t)). \end{aligned} \quad (11)$$

This reduces the FSPDA-SA recursion into a primal-only sequence by eliminating the dual sequence  $\lambda^t$ . In the deterministic optimization setting when  $\mathbf{A}(\xi) \equiv \mathbf{A}$  and  $\nabla \mathbf{f}(\mathbf{x}; \xi) \equiv \nabla \mathbf{f}(\mathbf{x})$ , (11) is equivalent to the EXTRA algorithm [Shi et al., 2015] using the mixing matrix  $\tilde{\mathbf{W}} = \mathbf{I} - \gamma \text{Diag}(\tilde{\mathbf{W}}\mathbf{1}) + \gamma \tilde{\mathbf{W}}$  where  $\tilde{\mathbf{W}}$  is the 0-1 adjacency matrix of  $\mathcal{G}$ . Here, with an appropriate choice of  $\gamma$ ,  $\tilde{\mathbf{W}}$  will be doubly stochastic and satisfies the convergence requirement in [Shi et al., 2015]. Similar observations have been made in [Nedic et al., 2017] for the gradient tracking and DIGing algorithms.

On the other hand, for stochastic optimization on random networks, (11) suggests each agent to keep the current and previous iterates received from neighbors in the corresponding time varying topology. In this case, (11) yields an extension of the EXTRA/GT algorithms to time varying topology.

## 3 Convergence Analysis of FSPDA

This section presents the convergence rate analysis of FSPDA for (1). Unless otherwise specified, we focus on the case with smooth but possibly non-convex objective function. Specifically, we consider:

**Assumption 3.1.** Each  $f_i$  is  $L$ -smooth, i.e., for  $i = 1, \dots, n$ ,

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \quad (12)$$

There exists  $f_\star > -\infty$  such that  $f_i(\mathbf{x}) \geq f_\star$  for any  $\mathbf{x} \in \mathbb{R}^d$ .

Note this implies that the global objective function  $F(\cdot)$  is  $L$ -smooth but possibly non-convex.

We further assume that the random network  $\mathcal{G}(\xi_a)$  is connected in expectation, yet each realization  $\mathcal{G}(\xi_a)$  may not be connected. Let  $\mathbf{R} = \mathbb{E}[\mathbf{I}(\xi_a)]$ , this leads to the following property concerning the expected graph Laplacian matrix  $\mathbf{A}^\top \mathbf{R} \mathbf{A} = \mathbb{E}[\mathbf{A}(\xi_a)^\top \mathbf{A}]$ . Defining the matrix  $\mathbf{K} := (\mathbf{I}_n - \mathbf{1}\mathbf{1}^\top/n) \otimes \mathbf{I}_d$ , we have

**Assumption 3.2.** There exists  $\rho_{\max} \geq \rho_{\min} > 0$  and  $\bar{\rho}_{\max} \geq \bar{\rho}_{\min} > 0$  such that

$$\rho_{\min} \mathbf{K} \preceq \mathbf{A}^\top \mathbf{R} \mathbf{A} \preceq \rho_{\max} \mathbf{K} \quad \text{and} \quad \bar{\rho}_{\min} \mathbf{K} \preceq \mathbf{A}^\top \mathbf{A} \preceq \bar{\rho}_{\max} \mathbf{K}. \quad (13)$$

It holds that  $\mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{K} = \mathbf{A}^\top \mathbf{R} \mathbf{A} = \mathbf{K} \mathbf{A}^\top \mathbf{R} \mathbf{A}$ . The above assumption can be satisfied if  $\mathcal{G}$  is connected [Yi et al., 2021], [Yi et al., 2018, Lemma 2] and  $\text{diag}(\mathbf{R}) > \mathbf{0}$  such that each edge is selected with a positive probability. As an important consequence, if  $\gamma \leq \rho_{\min}/\rho_{\max}^2$ , we have

$$\|(\mathbf{I} - \gamma \mathbf{A}^\top \mathbf{R} \mathbf{A}) \mathbf{x}\|_{\mathbf{K}}^2 \leq (1 - \gamma \rho_{\min}) \|\mathbf{x}\|_{\mathbf{K}}^2, \quad \forall \mathbf{x} \in \mathbb{R}^{nd}.$$

We thus observe that the operator  $(\mathbf{I} - \gamma \mathbf{A}^\top \mathbf{R} \mathbf{A})$  serves a similar purpose as the mixing matrix in a average consensus algorithms and  $\rho_{\min}$  can be interpreted as the spectral radius of  $\mathcal{G}$  similar

182 to [Koloskova et al., 2020, Eq. (12)]. Moreover, if we define  $\mathbf{Q} := (\mathbf{A}^\top \mathbf{R} \mathbf{A})^\dagger$  such that it holds  
 183  $\mathbf{Q} \mathbf{A}^\top \mathbf{R} \mathbf{A} = \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{Q} = \mathbf{K}$ , Assumption 3.2 implies that  $\rho_{\max}^{-1} \mathbf{K} \preceq \mathbf{Q} \preceq \rho_{\min}^{-1} \mathbf{K}$ .

184 Next we consider several assumptions on the noise variance of the random quantities in FSPDA:

185 **Assumption 3.3.** For any fixed  $\mathbf{x}_i \in \mathbb{R}^d$ ,  $i \in [n]$ , there exists  $\sigma_i \geq 0$  such that

$$\mathbb{E}_{\xi_i \sim \mathbb{P}_i} [\|\nabla f_i(\mathbf{x}_i; \xi_i) - \nabla f_i(\mathbf{x}_i)\|^2] \leq \sigma_i^2. \quad (14)$$

186 To simplify notations, we define  $\bar{\sigma}^2 := (1/n) \sum_{i=1}^n \sigma_i^2$ .

187 **Assumption 3.4.** For any fixed  $\mathbf{x} \in \mathbb{R}^{nd}$ , there exists  $\sigma_A \geq 0$  such that

$$\mathbb{E}_{\xi_a \sim \mathbb{P}_a} [\|\mathbf{A}(\xi_a)^\top \mathbf{A} \mathbf{x} - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}\|^2] \leq \sigma_A^2 \|\mathbf{x}\|_{\mathbf{K}}^2. \quad (15)$$

188 Assumption 3.3 is standard. Meanwhile for Assumption 3.4, the variance term  $\sigma_A^2$  measures the  
 189 quality of the random topology  $\mathcal{G}(\xi_a)$  in approximating the expected graph Laplacian  $\mathbf{A}^\top \mathbf{R} \mathbf{A}$ . The  
 190 latter is important as it contributes to the variance in the drift term of FSPDA. Observe that  $\sigma_A^2$   
 191 decreases with the proportion of edges selected in each random subgraph  $\mathcal{G}(\xi_a)$ .

192 To facilitate our discussions, we define the following quantities:

$$\bar{\mathbf{x}}^t := \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^t, \quad \sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2 = \|\mathbf{x}^t\|_{\mathbf{K}}^2. \quad (16)$$

193 **Convergence of FSPDA-SA.** We summarize the convergence rate for FSPDA-SA as follows. The proof  
 194 can be found in Appendix C:

**Theorem 3.5.** Under Assumptions 3.1, 3.2, 3.3, 3.4. Suppose that the step sizes satisfy the conditions defined in (46). Then, for any  $T \geq 1$  with the random stopping iteration  $\mathsf{T} \sim \text{Unif}\{0, \dots, T-1\}$ , the iterates generated by FSPDA-SA satisfy

$$\mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^{\mathsf{T}})\|^2] \leq \frac{F_0 - f_\star}{\alpha T/8} + 8\alpha \mathbb{C}_\sigma \frac{\bar{\sigma}^2}{n}, \quad (17)$$

$$\mathbb{E} [\sum_{i=1}^n \|\mathbf{x}_i^{\mathsf{T}} - \bar{\mathbf{x}}^{\mathsf{T}}\|^2] \leq \frac{F_0 - f_\star}{\alpha \gamma \rho_{\min} T/8} + \frac{8\alpha^2 \mathbb{C}_\sigma \bar{\sigma}^2}{\alpha \gamma \rho_{\min} n}, \quad (18)$$

for any  $\alpha > 0$ , where  $F_0, \mathbb{C}_\sigma$  are defined in (44), (50).

195  
 196 Setting  $\alpha = \mathcal{O}(n/\sqrt{T\bar{\sigma}^2})$ ,  $\alpha = \sqrt{n/(T\bar{\sigma}^2)}$  (and assuming  $\bar{\sigma} > 0$ ), we have

$$\mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^{\mathsf{T}})\|^2] = \mathcal{O}(\bar{\sigma}/\sqrt{nT}), \quad (19)$$

197 which is the same asymptotic convergence rate as a centralized SGD algorithm that takes  $n$  stochastic  
 198 gradient samples uniformly from each agent, i.e., linear speedup [Lian et al., 2017]. Also, using  
 199  $\alpha = 1$ , the consensus error converges as a rate of  $\mathbb{E} [\sum_{i=1}^n \|\mathbf{x}_i^{\mathsf{T}} - \bar{\mathbf{x}}^{\mathsf{T}}\|^2] = \mathcal{O}(n^2 \sigma_A^2 \rho_{\max}/(T \rho_{\min}^2))$   
 200 under the same step size choice used in (19). Notice that for  $T \gg 1$ , the effect of random topology  
 201 only degrades the convergence of consensus error, keeping the transient rate in (19) unaffected. If  
 202 the gradients are deterministic ( $\bar{\sigma} = 0$ ), setting  $\alpha = (L^2 \eta_\infty \rho_{\min})^{1/3}$ ,  $\alpha = \alpha_\infty$  will yield a better  
 203 convergence rate as  $\mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^{\mathsf{T}})\|^2] = \mathcal{O}(\sigma_A^4 \sqrt{n}/T)$ . Without a transient phase, the error due to  
 204 random graph and coordinate sparsification is persistent through  $\sigma_A^4$  in the above convergence rate.

205 We further show that the convergence of FSPDA-SA can be accelerated if the objective function of (1)  
 206 satisfies the Polyak-Lojasiewicz (PL) condition:

207 **Assumption 3.6.** There exists a constant  $\mu > 0$  such that  $2\mu(F(\mathbf{x}) - f_\star) \leq \|\nabla F(\mathbf{x})\|^2$ ,  $\forall \mathbf{x} \in \mathbb{R}^d$ .

208 Assumption 3.6 includes strongly convex functions as a special case, but also includes other non-  
 209 convex functions; see [Karimi et al., 2016]. We observe:

**Corollary 3.7.** Suppose the assumptions and step size conditions in Theorem 3.5 hold. Furthermore, with Assumption 3.6, there exists  $\delta \in (0, 1)$  such that for any  $t \geq 0$ ,

$$\mathbb{E}_t [F_{t+1} - f_\star] \leq (1 - \delta)(F_t - f_\star) + \mathbb{C}_\sigma \alpha^2 \bar{\sigma}^2 / n \quad (20)$$

for  $F_t, \mathbb{C}_\sigma$  defined in (44), (70), and  $\delta = \min\{\alpha\mu/4, \gamma\rho_{\min}/16, \eta\beta/(3\rho_{\min}), \eta/12\}$ .



The proof can be found in Appendix C.6. By setting  $\alpha = c \ln(T)/(n^2 T)$  in (20), with a carefully chosen  $c$  and a sufficiently large  $T$  such that  $\alpha \leq \alpha_\infty$ , we can ensure that

$$\mathbb{E} [F(\bar{\mathbf{x}}^T) - f_\star + \|\mathbf{x}^T\|_{\mathbf{K}}^2] = \mathcal{O}(\bar{\sigma}^2 \ln(T)/(\mu n T)) \quad (21)$$

In the case of deterministic gradient, i.e.,  $\bar{\sigma}^2 = 0$ , by setting  $\alpha = \alpha_\infty$ , (20) ensures a linear convergence rate of  $\mathbb{E} [F(\bar{\mathbf{x}}^T) - f_\star + \|\mathbf{x}^T\|_{\mathbf{K}}^2] = \mathcal{O}((1 - \delta)^T)$ , which shows that the performance of FSPDA-SA is on par with [Nedic et al., 2017, Xu et al., 2017], despite it only requires one round of (sparsified) transmission per iteration.

**Convergence of FSPDA-STORM.** To exploit the benefits of control variates, we need an additional assumption on the stochastic gradient map:

**Assumption 3.8.** Each stochastic function  $f_i(\cdot; \xi)$  is  $L_\xi$ -smooth in expectation, i.e., for  $i = 1, \dots, n$ ,

$$\mathbb{E}_\xi [\|\nabla f_i(\mathbf{x}; \xi) - \nabla f_i(\mathbf{y}; \xi)\|^2] \leq L_\xi^2 \|\mathbf{x} - \mathbf{y}\|^2 \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d. \quad (22)$$

The above assumption is also known as the mean-square smoothness condition, see [Cutkosky and Orabona, 2019], which is strictly stronger than Assumption 3.1. We observe the following convergence guarantee for FSPDA-STORM, whose proof can be found in Appendix D.

**Theorem 3.9.** Under Assumptions 3.1, 3.2, 3.3, 3.4, 3.8. Suppose that the step sizes satisfy the conditions in (184) - (214). Then, for any  $T \geq 1$  with the random stopping iteration  $T \sim \text{Unif}\{0, \dots, T-1\}$ , the iterates generated by FSPDA-STORM satisfy

$$\mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^T)\|^2] \leq \frac{F_0 - f_\star}{T\alpha/4} + \frac{(\mathbf{e} \cdot 2a_x^2 + \mathbf{f} \cdot 4a_x^2 n)\bar{\sigma}^2}{\alpha/4}, \quad (23)$$

$$\mathbb{E} [\sum_{i=1}^n \|\mathbf{x}_i^T - \bar{\mathbf{x}}^T\|^2] \leq \frac{F_0 - f_\star}{T\mathbf{a}\gamma\rho_{\min}/8} + \frac{(\mathbf{e} \cdot 2a_x^2 + \mathbf{f} \cdot 4a_x^2 n)\bar{\sigma}^2}{\mathbf{a}\gamma\rho_{\min}/8}, \quad (24)$$

where the constants  $F_0, \mathbf{a}, \mathbf{e}, \mathbf{f}$  are defined in (110).

Setting  $\alpha = \mathcal{O}(\bar{\sigma}^{-2/3} T^{-1/3})$ ,  $\eta = \mathcal{O}(n)$ ,  $\gamma = \mathcal{O}(T^{-1/3})$ ,  $\beta = \mathcal{O}(n^{-1} T^{-2/3})$ ,  $a_x = \mathcal{O}(\bar{\sigma}^{-4/3} T^{-2/3})$ ,  $a_\lambda = \mathcal{O}(T^{-1/3})$ ,  $\mathbf{f} = \mathcal{O}(n^{-1} T^{1/3})$  (see (111) - (117)), and initializing the algorithm such that  $\|\mathbf{v}^0\|_{\mathbf{K}}^2 = \mathcal{O}(T^{-2/3})$ ,  $\|\bar{\mathbf{m}}_x^0 - (1/n)\mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^0)\|^2 = \mathcal{O}(T^{-1/3})$  and  $\|\bar{\mathbf{m}}_x^0 - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^0, \lambda^0)\|^2 = \mathcal{O}(T^{-1/3})$ , we have

$$\mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^T)\|^2] = \mathcal{O}(\bar{\sigma}^{2/3}/T^{2/3}). \quad (25)$$

In regard to the order of  $\bar{\sigma}$  and  $T$ , provided that  $n$  is small, the convergence rate of FSPDA-STORM matches the lower bound [Arjevani et al., 2023] for non-convex functions under the same smoothness assumption. Moreover, by the same choice of step sizes, the consensus error converges at the rate of  $\mathbb{E} [\sum_{i=1}^n \|\mathbf{x}_i^T - \bar{\mathbf{x}}^T\|^2] = \mathcal{O}(\bar{\sigma}^{2/3} n \rho_{\min}^{-1} T^{-2/3})$ . We remark that in (25), the rate remains constant as  $n$  increases such that FSPDA-STORM does not offer the same *linear speedup* observed in Theorem 3.5 for FSPDA-SA. Nevertheless, as  $T \gg 1$ , the rate of FSPDA-STORM will surpass that of FSPDA-SA and other decentralized algorithms on time varying topologies.

Lastly, we provide detailed discussions on the convergence rates above, e.g., transient time, effects of random topology, etc., in Appendix B.

### 3.1 Insight from Analysis: Fixed Point Iteration of FSPDA-SA

From (8a), the following recursive relationship holds for  $\bar{\mathbf{x}}^t$ : using the relation  $\mathbf{1}^\top \mathbf{A}^\top = \mathbf{0}$ , we have

$$\bar{\mathbf{x}}^{t+1} = \bar{\mathbf{x}}^t - \frac{\alpha}{n} \sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t; \xi_i^t). \quad (26)$$

This shows that the evolution of  $\{\bar{\mathbf{x}}^t\}_{t \geq 0}$  is similar to that of ‘centralized’ SGD applied on (1) except that the local gradients are evaluated on the local iterates. However, it is still not straightforward to analyze the convergence of FSPDA-SA as the update of  $\mathbf{x}^t$  involves the dual variable  $\lambda^t$  which lacks an intuitive interpretation for constructing the right Lyapunov function.

To this end, we study the fixed point(s) of (8) to gain insights. Suppose that for some  $t_\star$ , the fixed point conditions  $\mathbb{E}[\lambda^{t_\star+1} | \xi^{t_\star}] = \lambda^{t_\star}$ ,  $\mathbb{E}[\mathbf{x}^{t_\star+1} | \xi^{t_\star}] = \mathbf{x}^{t_\star}$  hold. Since  $\mathbf{R}$  is a diagonal matrix with positive diagonal elements, we observe

$$\mathbb{E}[\lambda^{t_\star+1} | \xi^{t_\star}] = \lambda^{t_\star} \iff \mathbf{R} \mathbf{A} \mathbf{x}^{t_\star} = \mathbf{0} \iff \mathbf{A} \mathbf{x}^{t_\star} = \mathbf{0}, \quad (27)$$

246 On the other hand, the primal update yields

$$\mathbb{E}[\mathbf{x}^{t_*+1} \mid \xi^{t_*}] = \mathbf{x}^{t_*} - \alpha \nabla \mathbf{f}(\mathbf{x}^{t_*}) - \eta \mathbf{A}^\top \boldsymbol{\lambda}^{t_*}. \quad (28)$$

247 Since  $\mathbf{x}_1^{t_*} = \mathbf{x}_2^{t_*} = \dots = \mathbf{x}_n^{t_*}$  at the fixed point (due to (27)), by the consensus condition across two  
248 time steps, it implies

$$\begin{aligned} \mathbb{E}[\mathbf{x}^{t_*+1} \mid \xi^{t_*}] - \mathbf{x}^{t_*} &= (\mathbf{1} \otimes \mathbf{I}_d)(\bar{\mathbf{x}}^{t_*+1} - \bar{\mathbf{x}}^{t_*}) \\ &\iff \alpha \nabla \mathbf{f}(\mathbf{x}^{t_*}) + \eta \mathbf{A}^\top \boldsymbol{\lambda}^{t_*} = \frac{\alpha}{n} (\mathbf{1}\mathbf{1}^\top \otimes \mathbf{I}_d) \nabla \mathbf{f}(\mathbf{x}^{t_*}) \\ &\iff \eta \mathbf{A}^\top \boldsymbol{\lambda}^{t_*} = \alpha \left( \frac{1}{n} \mathbf{1}\mathbf{1}^\top - \mathbf{I}_n \right) \otimes \mathbf{I}_d \nabla \mathbf{f}((\mathbf{1} \otimes \mathbf{I}) \bar{\mathbf{x}}^{t_*}). \end{aligned} \quad (29)$$

249 From (29), we see that  $\hat{\boldsymbol{\lambda}}^t$  shall converge to the difference between global and local gradient. Inspired  
250 by the above, to facilitate the analysis later, we define

$$\mathbf{v}^t := \mathbf{A}^\top \boldsymbol{\lambda}^t + \frac{\alpha}{\eta} \nabla \mathbf{f}((\mathbf{1} \otimes \mathbf{I}) \bar{\mathbf{x}}^t), \quad (30)$$

251 for any  $t \geq 0$ . In particular, we see that  $\|\mathbf{v}^t\|_{\mathbf{K}}^2$  measures the violation of (29) in tracking the average  
252 deterministic gradient using the dual variables. The latter will be instrumental in analyzing the  
253 consensus error bound, as revealed in Lemma C.2.

## 254 4 Numerical Experiments

255 This section reports the numerical experiments on practical performance of FSPDA. For the time  
256 varying topology, we take an extreme setting where for each realization  $\mathcal{G}(\xi_a)$ , only one edge will  
257 be selected uniformly at random from  $\mathcal{G}$ . We evaluate the performance with the worst-agent metric,  
258 i.e., we present the training loss as  $\max_{i \in [n]} F(\mathbf{x}_i^t)$ , and the stationarity/gradient-norm measure as  
259  $\max_{i \in [n]} \|\nabla F(\mathbf{x}_i^t)\|^2$ . This captures the worst-case of the solutions produced by the algorithms.  
260 Unless otherwise specified, all algorithms are initialized with  $\mathbf{x}_i^0 = \bar{\mathbf{x}}^0$ , and for FSPDA we initialize  
261  $\hat{\boldsymbol{\lambda}}^0 = \mathbf{m}_{x,i}^0 = \mathbf{m}_{\lambda,i}^0 = \mathbf{0}$ , and the stochastic gradients are estimated with a batch size of 256. In the  
262 interest of space, omitted details and hyperparameters of the experiments can be found in Appendix F.

263 **MNIST Experiments.** The first set of experiments considers a moderate-scale setting of training a  
264 one hidden layer feed-forward neural network with 100 hidden neurons (total number of parameters  
265  $d = 79,510$ ) on the MNIST dataset with  $m = 60,000$  samples of 784-dimensional features.

266 In the first experiment, we consider the static topology  $\mathcal{G}$  as an Erdos-Renyi graph with connectivity of  
267  $p = 0.5$  and  $n = 10$  agents. We compare the proposed FSPDA-SA, FSPDA-STORM with six benchmark  
268 algorithms utilizing different types of time-varying topology. Among them, DSGD [Koloskova et al.,  
269 2020] and Swarm-SGD [Nadiradze et al., 2021] use the general time varying topology setting as FSPDA  
270 where each edge of  $\mathcal{G}(\xi_a)$  is active uniformly at random, in addition to random sparsification used  
271 FSPDA-SA and adaptive quantized used in Swarm-SGD; CHOCO-SGD [Koloskova et al., 2019b] takes  
272  $\mathcal{G}(\xi_a)$  as an broadcasting subgraph where one agent selects all his/her neighbors; Decen-Scaffnew  
273 [Mishchenko et al., 2022], LED [Alghunaim, 2024], and K-GT [Liu et al., 2024] utilize local updates  
274 where  $\mathcal{G}(\xi_a)$  is either taken as an empty topology, or as the static topology  $\mathcal{G}$ . We configure these  
275 algorithms such that they have the same communication cost (in terms of bits transmitted over  
276 network) *on average*. For instance, the local update algorithms (Decen-Scaffnew, LED, K-GT)  
277 only communicate once using  $\mathcal{G}$  every  $\mathcal{O}\left(\frac{|\mathcal{E}|d}{k}\right)$  iterations to match the communication cost of  
278  $k$ -coordinate sparse one-edge random graph used in FSPDA.

279 The local objective function held by each agent is the cross-entropy classification loss on a local  
280 dataset with  $m_i = 6000$  samples, plus a regularization loss  $\frac{\lambda}{2} \|\mathbf{x}_i\|^2$  with  $\lambda = 10^{-4}$ , where  $\mathbf{x}_i$  are the  
281 weight parameters of the feed-forward neural network classifier. We split the training set into  $n = 10$   
282 disjoint sets such that each set contains only one class label and assign each set to one agent as its  
283 local dataset. Note that as we do not shuffle the data samples across local datasets, the local objective  
284 function held by different agents will become highly heterogeneous.

285 Fig. 1 compares the squared gradient norm, training loss, consensus error of the benchmarked algo-  
286 rithms. We first note that both FSPDA algorithms have significantly outperformed DSGD, Swarm-SGD  
287 on the general time varying topology as well as CHOCO-SGD. Meanwhile, the performance of FSPDA  
288 is comparable to the local update algorithms Decen-Scaffnew, LED, K-GT. Notice that the latter



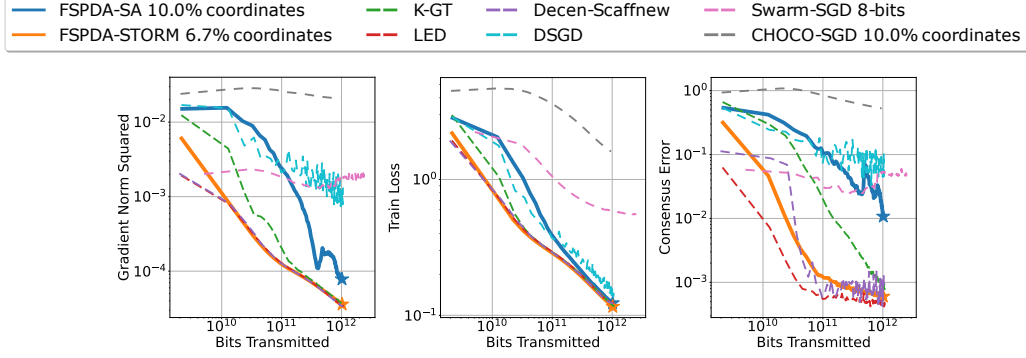


Figure 1: Feed-forward neural network classification training on MNIST using  $10^6$  iterations.

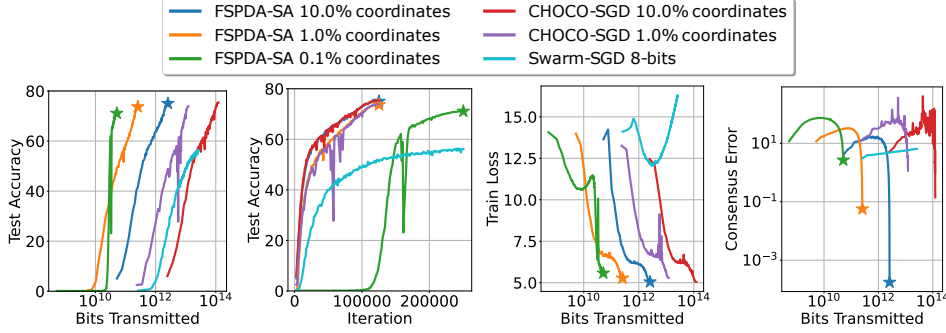


Figure 2: Resnet-50 classification training on Imagenet.

require additional synchronization steps which may not be suitable for random networks. Lastly, we notice that as  $T \gg 1$ , FSPDA-STORM can slightly outperform FSPDA-SA due to its  $\mathcal{O}(1/T^{2/3})$  rate as shown in our analysis. We further expand the experiments by a series of ablation studies over data heterogeneity, sparsity levels, graph topologies, gradient noise and dual momentum in Appendix E.

**Imagenet Experiments.** The second set of experiments consider a large-scale setting for training a Resnet-50 network (total number of parameters  $d = 25,557,032$ ) on the Imagenet dataset (training dataset of 1,281,168 images from 100 classes, re-scaled and cropped to  $256 \times 256$  image dimensions). We consider cross-entropy classification loss plus the same L2 norm regularization loss as in the previous setup. We split the dataset across a network of  $n = 8$  nodes where the static graph  $\mathcal{G}$  is taken as the fully connected topology. The performance metrics are measured at the network average iterate  $\bar{x}^t$ . Inspired by [Loshchilov and Hutter, 2016, Eq. (5)] we adopt a cosine learning rate scheduling with 5 epochs of linear warm up for every algorithm. In particular, the step sizes  $\alpha, \eta$  of FSPDA-SA are scheduled simultaneously such that  $\alpha_t/\eta_t$  remains constant, as illustrated in Appendix F. We draw a batch of 128 samples to estimate the stochastic gradient.

We focus on the communication efficiency and only compare FSPDA-SA, CHOCO-SGD, Swarm-SGD in this experiment due to limited resources. The results are reported in Figure 2 that compare the test accuracy and training loss against iteration number and bits transmitted. When compared with CHOCO-SGD, FSPDA-SA achieves almost the same accuracy using one-edge random graphs with at least 100x reduction in communication cost on 100 epoch training. Also notice that further compressing the communication to 0.1% sparse coordinates in FSPDA-SA requires more training epochs to recover the same level of accuracy.

**Conclusions.** This paper proposed a fully stochastic primal dual gradient algorithm (FSPDA) framework for decentralized optimization over arbitrarily time varying random networks. We utilize a new stochastic augmented Lagrangian function and apply SA to search for its saddle point. We develop two algorithms, one is by plain SA (FSPDA-SA), and one uses control variates for variance reduction (FSPDA-STORM). We prove that both algorithms achieve state-of-the-art convergence rates, while relaxing assumptions on both bounded heterogeneity and the type of time varying topologies.

## References

- Sulaiman A Alghunaim. Local exact-diffusion for decentralized optimization and learning. *IEEE Transactions on Automatic Control*, 2024.
- Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1): 165–214, 2023.
- Dimitri Bertsekas. *Nonlinear Programming*, volume 4. Athena Scientific, 2016.
- Pascal Bianchi, Walid Hachem, and Adil Salim. A fully stochastic primal-dual algorithm. *Optimization Letters*, 15(2):701–710, 2021.
- Tsung-Hui Chang, Mingyi Hong, Hoi-To Wai, Xinwei Zhang, and Songtao Lu. Distributed learning in the nonconvex world: From batch data to streaming and beyond. *IEEE Signal Processing Magazine*, 37(3):26–38, 2020.
- Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.
- Luyao Guo, Sulaiman A Alghunaim, Kun Yuan, Laurent Condat, and Jinde Cao. Revisiting decentralized proxskip: Achieving linear speedup. *arXiv preprint arXiv:2310.07983*, 2023.
- Davood Hajinezhad and Mingyi Hong. Perturbed proximal primal–dual algorithm for nonconvex nonsmooth optimization. *Mathematical Programming*, 176(1):207–245, 2019.
- Mingyi Hong, Davood Hajinezhad, and Ming-Min Zhao. Prox-pda: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *International Conference on Machine Learning*, pages 1529–1538. PMLR, 2017.
- Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.
- Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer, 2016.
- Anastasia Koloskova, Tao Lin, Sebastian U Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. In *International Conference on Learning Representations*, 2019a.
- Anastasia Koloskova, Sebastian Stich, and Martin Jaggi. Decentralized stochastic optimization and gossip algorithms with compressed communication. In *International Conference on Machine Learning*, pages 3478–3487. PMLR, 2019b.
- Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pages 5381–5393. PMLR, 2020.
- Anastasiia Koloskova, Tao Lin, and Sebastian U Stich. An improved analysis of gradient tracking for decentralized machine learning. *Advances in Neural Information Processing Systems*, 34: 11422–11435, 2021.
- Dmitry Kovalev, Elnur Gasanov, Alexander Gasnikov, and Peter Richtarik. Lower bounds and optimal algorithms for smooth and strongly convex decentralized optimization over time-varying networks. *Advances in Neural Information Processing Systems*, 34:22325–22335, 2021.
- Dmitry Kovalev, Ekaterina Borodich, Alexander Gasnikov, and Dmitrii Feoktistov. Lower bounds and optimal algorithms for non-smooth convex decentralized optimization over time-varying networks. *arXiv preprint arXiv:2405.18031*, 2024.

363 Jinlong Lei, Han-Fu Chen, and Hai-Tao Fang. Asymptotic properties of primal-dual algorithm for  
364 distributed stochastic optimization over random networks with imperfect communications. *SIAM*  
365 *Journal on Control and Optimization*, 56(3):2159–2188, 2018.

366 Huan Li and Zhouchen Lin. Accelerated gradient tracking over time-varying graphs for decentralized  
367 optimization. *Journal of Machine Learning Research*, 25(274):1–52, 2024.

368 Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized  
369 algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic  
370 gradient descent. *Advances in neural information processing systems*, 30, 2017.

371 Yue Liu, Tao Lin, Anastasia Koloskova, and Sebastian U Stich. Decentralized gradient tracking with  
372 local steps. *Optimization Methods and Software*, pages 1–28, 2024.

373 Ilan Lobel and Asuman Ozdaglar. Distributed subgradient methods for convex optimization over  
374 random networks. *IEEE Transactions on Automatic Control*, 56(6):1291–1306, 2010.

375 Paolo Di Lorenzo and Gesualdo Scutari. Next: In-network nonconvex optimization. *IEEE Transac-*  
376 *tions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

377 Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv*  
378 *preprint arXiv:1608.03983*, 2016.

379 Songtao Lu, Xinwei Zhang, Haoran Sun, and Mingyi Hong. Gnsd: A gradient-tracking based  
380 nonconvex stochastic algorithm for decentralized optimization. In *2019 IEEE Data Science*  
381 *Workshop (DSW)*, pages 315–321. IEEE, 2019.

382 Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes!  
383 local gradient steps provably lead to communication acceleration! finally! In *International*  
384 *Conference on Machine Learning*, pages 15750–15769. PMLR, 2022.

385 Giorgi Nadiradze, Amirmojtaba Sabour, Peter Davies, Shigang Li, and Dan Alistarh. Asynchronous  
386 decentralized sgd with quantized and local updates. *Advances in Neural Information Processing*  
387 *Systems*, 34:6829–6842, 2021.

388 Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization.  
389 *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

390 Angelia Nedic, Alex Olshevsky, and Wei Shi. Achieving geometric convergence for distributed  
391 optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

392 Shi Pu, Alex Olshevsky, and Ioannis Ch Paschalidis. A sharp estimate on the transient time of  
393 distributed stochastic gradient descent. *IEEE Transactions on Automatic Control*, 67(11):5900–  
394 5915, 2021.

395 Tiancheng Qin, S Rasoul Etesami, and César A Uribe. Communication-efficient decentralized local  
396 sgd over undirected networks. In *2021 60th IEEE Conference on Decision and Control (CDC)*,  
397 pages 3361–3366. IEEE, 2021.

398 Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE*  
399 *Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

400 S Sundhar Ram, Angelia Nedić, and Venugopal V Veeravalli. Distributed stochastic subgradient  
401 projection algorithms for convex optimization. *Journal of optimization theory and applications*,  
402 147:516–545, 2010.

403 Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized  
404 consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

405 Jinming Xu, Shanying Zhu, Yeng Chai Soh, and Lihua Xie. Convergence of asynchronous distributed  
406 gradient methods over stochastic networks. *IEEE Transactions on Automatic Control*, 63(2):  
407 434–448, 2017.

- 408 Chung-Yiu Yau and Hoi-To Wai. Fully stochastic distributed convex optimization on time-varying  
409 graph with compression. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages  
410 145–150. IEEE, 2023.
- 411 Xinlei Yi, Lisha Yao, Tao Yang, Jemin George, and Karl H Johansson. Distributed optimization for  
412 second-order multi-agent systems with dynamic event-triggered communication. In *2018 IEEE*  
413 *Conference on Decision and Control (CDC)*, pages 3397–3402. IEEE, 2018.
- 414 Xinlei Yi, Shengjun Zhang, Tao Yang, Tianyou Chai, and Karl H Johansson. Linear convergence  
415 of first-and zeroth-order primal–dual algorithms for distributed nonconvex optimization. *IEEE*  
416 *Transactions on Automatic Control*, 67(8):4194–4201, 2021.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]



Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Due to limited computing resources and time constraints, we are unable to perform multiple runs of our algorithms and report the error bars. We will produce the error bar statistics if time permits.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

## 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

## 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

### 16. Declaration of LLM usage



726 Question: Does the paper describe the usage of LLMs if it is an important, original, or  
727 non-standard component of the core methods in this research? Note that if the LLM is used  
728 only for writing, editing, or formatting purposes and does not impact the core methodology,  
729 scientific rigorousness, or originality of the research, declaration is not required.

730 Answer: [NA]

731 Justification: [NA]

732 Guidelines:

- 733 • The answer NA means that the core method development in this research does not  
734 involve LLMs as any important, original, or non-standard components.
- 735 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)  
736 for what should or should not be described.

## 737 A Asynchronous Implementation of FSPDA

738 We show that FSPDA algorithms can be implemented in an asynchronous manner. For illustration  
 739 purpose, we concentrate on FSPDA-SA and a scenario where any sparsified communication round is  
 740 almost delay-free and only one edge is active at a time, also known as the pairwise gossip setting, while  
 741 the stochastic gradient computation constitutes the major synchronization overhead in the algorithm.  
 742 We describe the implementation of (8a), (8b) by the pseudo-code in Algorithm 1 from the perspective  
 743 of the  $i$ -th agent. Notice that the threads `communication_thread()` and `computation_thread()`  
 744 are persistently running *in parallel* at each agent. Moreover, the local variables  $\mathcal{B}_i$ ,  $g_i$ ,  $t_i$  are held by  
 745 agent  $i$  and updated by both threads.

746 The thread `communication_thread()` aims at preparing the message  $\sum_{j \in \mathcal{N}_i(\xi_a^t)} \mathbf{C}_{ij}(\xi_a^t)(\mathbf{x}_j^t - \mathbf{x}_i^t)$   
 747 needed at (8a), (8b) using a gossip-like step. At each run by agent  $i$ , the agent checks if s/he is  
 748 connected to any active neighbor on the current graph  $\mathcal{G}^t = \mathcal{G}(\xi_a^t) := (\mathcal{V}, \mathcal{E}(\xi_a^t))$ . If the pair of  
 749 agents  $(i, j)$  are connected, they will exchange the sparsified decision variable  $\{(\mathbf{x}_{i,k}^{t_i}, \mathbf{x}_{j,k}^{t_j})\}_{k \in \mathcal{I}_{ij}(\xi_a^t)}$   
 750 and make preparation for the computation thread. The neighbor  $j$  whose communicated with agent  
 751  $i$  will be added to the buffer  $\mathcal{B}_i$ . Notice that the protocol is designed such that the asynchronous  
 752 implementation of FSPDA-SA aligns with (8a), (8b). For instance, our asynchronous gradient model  
 753 with  $b_i(\xi^t) = 0$  allows agent  $i$  to skip gradient computation and the sparse extended graph model  
 754 (6) allows agent  $i$  to skip communication, as implemented in [L7, Alg. 2] which serves as an update  
 755 after the *idle* state of agent  $i$  to compensate for the missed iterations led by neighbor  $j \in \mathcal{B}_i$  using the  
 756 locally stored  $\hat{\lambda}_i^{t_i}$ .

757 The thread `computation_thread()` aims at executing the primal-dual steps (8a), (8b) with local  
 758 updates, i.e., updating using local stochastic gradient before the next round of communication. Note  
 759 that in federated learning, local update has been used extensively which led to significant performance  
 760 improvement; see [Kairouz et al., 2021, Qin et al., 2021]. As mentioned, the stochastic gradient (SG)  
 761 computation in [L2, Alg. 3] is the major bottleneck of the algorithm. Upon the completion of [L2,  
 762 Alg. 3], the two cases in [L3, Alg. 3] & [L5, Alg. 3] essentially implement (8a), (8b). For the case  
 763 of [L5, Alg. 3] where the communication buffer  $\mathcal{B}_i$  is non-empty, we perform a sparse gossip with  
 764 SG update. Upon completing this round of primal-dual update, the communication buffer  $\mathcal{B}_i$  will be  
 765 cleared.

766 Overall, by running the two threads persistently at each agent, the decentralized system effectively  
 767 implements FSPDA-SA in (8a), (8b) as an asynchronous algorithm. The same asynchronous imple-  
 768 mentation can be easily extended to FSPDA-STORM.

---

### Algorithm 1 FSPDA from Agent $i$ 's Perspective

---

- 1: **input:** Iteration number  $T$ .
  - 2: **local variable (initialize):** Communication buffer  $\mathcal{B}_i$ ; gradient counter  $g_i = 0$ ; iteration counter  $t_i = 0$ .
  - 3: **while**  $\max_{j \in [n]} t_j < T$  **in parallel do**
  - 4:   `communication_thread()`; see Algorithm 2.
  - 5:   `computation_thread()`; see Algorithm 3.
  - 6: **end while**
-

---

**Algorithm 2** *communication\_thread()* of Agent  $i$ 

---

```
1: local variable: buffer  $\mathcal{B}_i$ ; counters  $g_i, t_i$ .
2: for  $(i, j) \in \mathcal{E}(\xi_a^t)$  do
3:   if  $\mathcal{B}_i = \emptyset$  and  $\mathcal{B}_j = \emptyset$  then
4:     Agents  $i, j$  exchanges  $t_i, t_j$ .
5:   if  $t_i < t_j$  then
6:     Interrupt [L2, Alg. 3] of computation_thread() and run [L4, Alg. 3] to consume (any) SG
       buffer.
7:      $\Rightarrow$  Local non-SG step: evaluate

$$\mathbf{x}_i^{t_j} = \mathbf{x}_i^{t_i} - (t_j - t_i)\eta\hat{\lambda}_i^{t_i}, \quad (31)$$

       and set  $\hat{\lambda}_i^{t_j} = \hat{\lambda}_i^{t_i}, t_i \leftarrow t_j$ .
8:   end if
9:   /* begin streaming index-value pairs */
10:  for  $k \in \mathcal{I}_{ij}(\xi_a^t)$  do
11:    Agent  $j$  receives  $(k, \mathbf{x}_{i,k}^{t_i})$  from agent  $i$ .
12:    Agent  $i$  receives  $\mathbf{x}_{j,k}^{t_j}$  from agent  $j$ .
13:  end for
14:  /* end streaming upon time-out or interruption */
15:  if gossip is successful then
16:    Update  $\mathcal{B}_i \leftarrow \{j\}, \mathcal{B}_j \leftarrow \{i\}$ .
17:  end if
18: end if
19: end for
```

---

---

**Algorithm 3** *computation\_thread()* of Agent  $i$ 

---

```
1: local variable: buffer  $\mathcal{B}_i$ ; counters  $g_i, t_i$ .
2: /* begin compute SG */
   Compute  $\nabla f_i(\mathbf{x}_i^{t_i}; \xi_i^{t_i})$  and increment  $g_i \leftarrow g_i + 1$ .
   /* end compute SG upon completion or interruption */
3: if communication buffer  $\mathcal{B}_i = \emptyset$  then
4:    $\Rightarrow$  Local SG step: with  $\hat{c}_i = g_i/(t_i + 1)$ , evaluate

$$\mathbf{x}_i^{t_i+1} = \mathbf{x}_i^{t_i} - \eta\hat{\lambda}_i^{t_i} - \alpha\hat{c}_i\nabla f_i(\mathbf{x}_i^{t_i}; \xi_i^{t_i}), \quad (32)$$

       and set  $\hat{\lambda}_i^{t_i+1} = \hat{\lambda}_i^{t_i}, t_i \leftarrow t_i + 1$ .
5: else
6:   Identify the quantities:

$$t'_i = \max\{t_i, \max_{j \in \mathcal{B}_i} t_j\}, \quad d_i = 1 + t'_i - t_i, \quad \mathbf{C}_{ij}(\xi^{t'_i}) = \text{BinDiag}(\mathcal{I}_{ij}(\xi^{t'_i})), \quad (33)$$


$$\hat{c}_i = \begin{cases} g_i/(t'_i + 1) & \text{if } \nabla f_i(\mathbf{x}_i^{t_i}; \xi_i^{t'_i}) \text{ is ready,} \\ 0 & \text{otherwise.} \end{cases} \quad (34)$$

7:    $\Rightarrow$  Gossip with SG step: evaluate

$$\mathbf{x}_i^{t'_i+1} = \mathbf{x}_i^{t_i} - \gamma \sum_{j \in \mathcal{B}_i} \mathbf{C}_{ij}(\xi^{t'_i})(\mathbf{x}_i^{t_i} - \mathbf{x}_j^{t_j}) - d_i\eta\hat{\lambda}_i^{t_i} - \alpha\hat{c}_i\nabla f_i(\mathbf{x}_i^{t_i}; \xi_i^{t'_i}), \quad (35)$$


$$\hat{\lambda}_i^{t'_i+1} = \hat{\lambda}_i^{t_i} + \beta \sum_{j \in \mathcal{B}_i} \mathbf{C}_{ij}(\xi^{t'_i})(\mathbf{x}_i^{t_i} - \mathbf{x}_j^{t_j}), \quad (36)$$

       and set  $t_i \leftarrow t'_i + 1, \mathcal{B}_i \leftarrow \emptyset$ .
8: end if
```

---

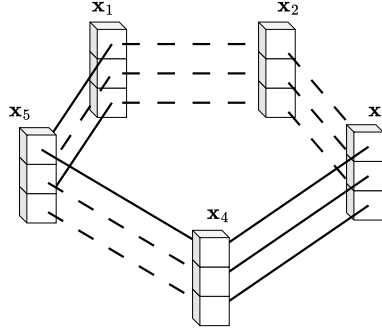


Figure 3: Illustration of a (time-varying) random graph  $\mathcal{G}(\xi)$  for primal variable of dimension  $d = 3$  on a ring network of  $n = 5$  nodes. Solid lines represent active edges while dashed lines represent disconnected edges. In this example, node 2 is considered as idle in an asynchronous environment.  $\mathbf{C}_{15}(\xi)$  is a diagonal matrix such that  $\text{diag}(\mathbf{C}_{15}(\xi)) = (1, 0, 1)$ .

## B Detailed Convergence Rate Analysis for FSPDA-SA

Under the parameter choices of  $\mathbf{a} = \frac{80\delta_1 L^2 n}{\sqrt{T\bar{\sigma}^2}\gamma_\infty \rho_{\min}}$ ,  $\alpha = \sqrt{n/(T\bar{\sigma}^2)}$ , we list several remarks on the convergence rate analysis of FSPDA-SA as follows.

- In addition to the crude bound in (19), we can derive a fine grained characterization for the convergence rate of FSPDA-SA. Here, an interesting aspect is bounding the *transient time* of the number of iterations for the decentralized algorithm to achieve the rate on par with CSGD in (19), independent of the network topology [Pu et al., 2021]. Using the definition of  $\mathbf{C}_\sigma$  in (70), it can be shown that FSPDA-SA has a transient time of

$$T_{\text{trans}} = \Omega\left(\frac{\sigma_A^4}{\rho_{\min}^4} \cdot \max\left\{n^6 \rho_{\max}^2, \min\left\{\frac{\bar{\rho}_{\max}^4 \rho_{\max}^6}{n\bar{\sigma}^2 \rho_{\min}^3}, \frac{n^{5/2} \bar{\rho}_{\max}^2 \rho_{\max}^4}{\bar{\sigma}^2 \rho_{\min}^2}\right\}\right\}\right) \quad (37)$$

where we have hidden the dependence of  $L$ ,  $F(\bar{\mathbf{x}}^0) - f_*$ ,  $\|\mathbf{x}^0\|_{\mathbf{K}}^2$  in the  $\Omega(\cdot)$  notation.

- Theorem 3.5 also guarantees that  $\mathbb{E}[\|\mathbf{x}^T\|_{\mathbf{K}}^2] \rightarrow 0$  and  $\mathbb{E}[\|\mathbf{v}^T\|_{\mathbf{K}}^2] \rightarrow 0$  as  $T \rightarrow \infty$ . The latter ensures that the gradient tracking error [Lu et al., 2019] variable converges to zero and thus FSPDA-SA is stable at a global stationary solution (see more discussion around the definition of  $\mathbf{v}$  in (30)). Notice that at each iteration, FSPDA-SA only communicates the local parameters  $\mathbf{x}^t$  once, while the gradient tracking algorithm [Lu et al., 2019] communicates twice for the gradient tracking vectors and the local parameters.
- For sufficiently large  $T$ , the effect of noisy network only remains dominant in  $\mathbb{E}[\|\mathbf{x}^T\|_{\mathbf{K}}^2] = \mathcal{O}(n^2 \sigma_A^2 \rho_{\max} / (T \rho_{\min}^2))$ , keeping the rate in (19) unaffected. When comparing against synchronous DSGD [Lian et al., 2017], with  $b_i(\xi) = 1$  for all  $i \in [n]$ , FSPDA-SA will converge faster in terms of stationarity during the post-transient stage due to the shorter iteration time under coordinate sparsification and random graph.
- When  $\bar{b}_i \neq \bar{b}_j$ , i.e., when the local update rates are non-uniform, the asynchronous gradient induces an error that will be reflected on the convergence error  $\mathcal{O}(\sigma/\sqrt{nT})$  through  $\sigma$ . We remark that under a fully controlled environment, FSPDA-SA can accelerate through asynchronous gradient only when the iteration time speedup out-weights the induced variance error.

## C Proof of Theorem 3.5

As our goal is to develop bounds for  $\|\nabla F(\bar{\mathbf{x}}^t)\|^2$ , a natural idea is to consider the descent of the primal objective value  $F(\bar{\mathbf{x}}^t)$ . This yields:

796 **Lemma C.1.** Under Assumption 3.1 and 3.3, and the step size condition  $\alpha \leq \frac{1}{4L}$ ,

$$\mathbb{E}_t [F(\bar{\mathbf{x}}^{t+1})] \leq F(\bar{\mathbf{x}}^t) - \frac{\alpha}{4} \|\nabla F(\bar{\mathbf{x}}^t)\|^2 + \frac{3\alpha L^2}{4n} \|\mathbf{x}^t\|_{\mathbf{K}}^2 + \frac{\alpha^2 L}{2n} \bar{\sigma}^2. \quad (38)$$

797 See Appendix C.1 for the proof. It can be seen from the above lemma that if the consensus error  
798 satisfies  $\|\mathbf{x}^t\|_{\mathbf{K}}^2 = o(1)$ , then setting  $\alpha = 1/\sqrt{T}$  suffices to yield

$$(1/T) \sum_{t=1}^T \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] = \mathcal{O}(1/\sqrt{T}). \quad (39)$$

799 That said, the evolution of  $\|\mathbf{x}^t\|_{\mathbf{K}}^2$  tends to be complicated as the latter co-evolves with the dual  
800 variable  $\lambda^t$  and the gradient. To simplify, we impose the following preliminary condition on the step  
801 sizes

$$\gamma \leq \min \left\{ \frac{\rho_{\min}}{\rho_{\max}^2}, \frac{\rho_{\min}}{2\sigma_A^2 \rho_{\max}} \right\}, \quad \alpha \leq 1, \quad \eta \leq 1. \quad (40)$$

802 **Lemma C.2.** Under Assumptions 3.1, 3.2, 3.3, 3.4 and the step size condition (40). The consensus  
803 error follows the recursive inequality

$$\mathbb{E} [\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2] \leq \left[ 1 - \frac{\gamma}{2} \rho_{\min} + \alpha(1 + 3L^2) \right] \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] + 2\eta^2 \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] \quad (41)$$

$$- 2\eta \mathbb{E} [\langle \mathbf{x}^t | \mathbf{v}^t \rangle_{\mathbf{K} - \gamma \mathbf{A}^\top \mathbf{R} \mathbf{A}}] + \alpha^2 n \bar{\sigma}^2. \quad (42)$$

804 See Appendix C.2 for the proof.

805 The above lemma shows an intricate structure for the consensus error as the latter also depends on  
806 the violation of (29), i.e.,  $\|\mathbf{v}^t\|_{\mathbf{K}}^2$ , and the *weighted* inner product between  $\mathbf{x}^t, \mathbf{v}^t$ . Naturally, we can  
807 further control the above terms:

808 **Lemma C.3.** Under Assumption 3.1, 3.3. Let  $\alpha \leq 1$ , then for any constant  $c > 0$ , the dual error  
809 satisfies

$$\begin{aligned} \mathbb{E} [\|\mathbf{v}^{t+1}\|_{\mathbf{Q} + c\mathbf{K}}^2] &\leq \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{Q} + c\mathbf{K}}^2] + 2\alpha(\rho_{\min}^{-1} + c) \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] + 2\beta \mathbb{E} [\langle \mathbf{v}^t | \mathbf{x}^t \rangle_{\mathbf{K} + c\mathbf{A}^\top \mathbf{R} \mathbf{A}}] \\ &+ \left( 2\beta^2 \rho_{\max}^2 + \frac{10\alpha^3 L^4}{\eta^2} \right) (\rho_{\min}^{-1} + c) \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] + \frac{5\alpha^3}{\eta^2} L^2 (\rho_{\min}^{-1} + c) \{ \bar{\sigma}^2 + 2n \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] \} \end{aligned} \quad (43)$$

810 See Appendix C.3 for the proof. Notice that we have considered the weighted norm  $\|\mathbf{v}^t\|_{\mathbf{Q} + c\mathbf{K}}^2$  to  
811 induce a favorable inner product term for  $\mathbf{x}^t, \mathbf{v}^t$  below:

812 **Lemma C.4.** Under Assumption 3.1, 3.3, and the step size condition (40), then

$$\begin{aligned} \mathbb{E} [\langle \mathbf{x}^{t+1} | \mathbf{v}^{t+1} \rangle_{\mathbf{K}}] &\leq \frac{\alpha - \eta}{2} \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] + \mathbb{E} [\langle \mathbf{x}^t | \mathbf{v}^t \rangle_{\mathbf{K} - \gamma \mathbf{A}^\top \mathbf{R} \mathbf{A} - \eta \beta \mathbf{A}^\top \mathbf{R} \mathbf{A}}] \\ &+ \left\{ \beta \rho_{\max} + \frac{\alpha + 3\alpha^2 L^2 + \alpha \gamma^2 \sigma_A^2 \rho_{\max} + 20\alpha^3 L^4}{2\eta} + \frac{\eta \beta^2 \rho_{\max}^2}{2} \right\} \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] \\ &+ \frac{10\alpha^3 L^2 + \alpha^3 n}{2\eta} \bar{\sigma}^2 + \frac{10\alpha^3 L^2 n}{\eta} \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2]. \end{aligned}$$

813 See Appendix C.4 for the proof.

814 The above lemma shows that the recursion for the inner product  $\langle \mathbf{x}^t | \mathbf{v}^t \rangle_{\mathbf{K}}$  also depends on  $\|\mathbf{x}^t\|_{\mathbf{K}}^2$ ,  
815  $\|\mathbf{v}^t\|_{\mathbf{K}}^2$ , etc., and can be similarly controlled. Our final step is to construct a potential function  $F_t$   
816 whose recursive relation can guide us towards the convergence of FSPDA:



**Theorem C.5.** For some constants  $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d} > 0$ , we define the potential function

$$F_t = \mathbb{E} [F(\bar{\mathbf{x}}^t) + \mathbf{a} \|\mathbf{x}^t\|_{\mathbf{K}}^2 + \mathbf{b} \|\mathbf{v}^t\|_{\mathbf{Q} + \mathbf{c}\mathbf{K}}^2 + \mathbf{d} \langle \mathbf{x}^t \mid \mathbf{v}^t \rangle_{\mathbf{K}}]. \quad (44)$$

Then, by the following choice of hyperparameters

$$\mathbf{b} = \mathbf{a} \cdot \frac{\eta}{\beta}, \quad \mathbf{c} = \frac{(\eta\beta + \gamma)\mathbf{d} - 2\eta\gamma\mathbf{a}}{2\beta\mathbf{b}}, \quad \mathbf{d} = \delta_1\eta\mathbf{a}, \quad (45)$$

$$\alpha \leq \alpha_\infty, \quad \eta \leq \eta_\infty, \quad \gamma \leq \gamma_\infty, \quad \beta = 1, \quad \delta_1 \geq 8, \quad (46)$$

where

$$\gamma_\infty := \frac{\rho_{\min}}{\rho_{\max}^2} \min \left\{ 1, \frac{\rho_{\max}}{2\sigma_A^2} \right\}, \quad \eta_\infty := \frac{\rho_{\min}^2}{64\delta_1^2\bar{\rho}_{\max}^2\rho_{\max}^2} \gamma_\infty, \quad (47)$$

$$\alpha_\infty := \frac{\gamma_\infty\rho_{\min}}{80\delta_1\sqrt{n}} \min \left\{ \frac{\mathbf{a}}{L^2}, \quad \eta_\infty\rho_{\min}, \sqrt{\frac{\eta_\infty\rho_{\min}}{L^2\mathbf{a}}} \right\}, \quad (48)$$

it holds that  $F_t \geq F(\bar{\mathbf{x}}^t) \geq f_\star > -\infty$  for any  $t \geq 0$ , and the potential function follows the inequality

$$F_{t+1} \leq F_t - \frac{\alpha}{8} \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] - \frac{\mathbf{a}\gamma\rho_{\min}}{8} \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] - \mathbf{a}\eta^2 \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] + \mathbb{C}_\sigma \alpha^2 \bar{\sigma}^2 / n \quad (49)$$

such that

$$\mathbb{C}_\sigma \leq \frac{L}{2} + \mathbf{a}n^2 + \mathbf{a}n\alpha L^2 \left( \frac{5}{\eta\beta\rho_{\min}} + \frac{5\delta_1\gamma}{\eta\beta} + 5\delta_1 \right) + \frac{\mathbf{a}\delta_1 n^2 \alpha}{2} \quad (50)$$

See Appendix C.5 for the proof.

817

818 By summing up (49) from  $t = 0$  to  $t = T - 1$  and rearranging terms, we obtain the bounds that lead  
819 to the theorem.  $\square$

## 820 C.1 Proof of Lemma C.1

821 By Assumption 3.1, it is straightforward to derive that

$$\mathbb{E}_t [F(\bar{\mathbf{x}}^{t+1})] \leq F(\bar{\mathbf{x}}^t) - \frac{\alpha}{n} \langle \nabla F(\bar{\mathbf{x}}^t) \mid \mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^t) \rangle + \frac{\alpha^2 L}{2n^2} \mathbb{E}_t [\|\mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^t; \xi^t)\|^2] \quad (51)$$

822 where we have used the shorthand notation  $\mathbf{1}_\otimes^\top := \mathbf{1}^\top \otimes \mathbf{I}$ . The second term of (51) can be bounded  
823 as

$$\frac{\alpha}{n} \langle \nabla F(\bar{\mathbf{x}}^t) \mid \mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^t) \rangle \geq \frac{\alpha}{2} \|\nabla F(\bar{\mathbf{x}}^t)\|^2 - \frac{\alpha L^2}{2n} \|\mathbf{x}^t\|_{\mathbf{K}}^2 \quad (52)$$

824 The third term of (51) can be bounded as

$$\mathbb{E}_t [\|\mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^t; \xi^t)\|^2] = \mathbb{E}_t [\|\mathbf{1}_\otimes^\top (\nabla \mathbf{f}(\mathbf{x}^t; \xi^t) - \nabla \mathbf{f}(\mathbf{x}^t))\|^2] + \|\mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^t)\|^2$$

825 Notice that due to the independence of gradient noise, we have

$$826 \mathbb{E}_t [\|\mathbf{1}_\otimes^\top (\nabla \mathbf{f}(\mathbf{x}^t; \xi^t) - \nabla \mathbf{f}(\mathbf{x}^t))\|^2] \leq n\bar{\sigma}^2. \text{ Furthermore,}$$

$$\|\mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^t)\|^2 \leq 2nL^2 \|\mathbf{x}^t\|_{\mathbf{K}}^2 + 2n^2 \|\nabla F(\bar{\mathbf{x}}^t)\|^2. \quad (53)$$

827 Substituting the above into (51) and setting the step size  $\alpha \leq 1/(4L)$  concludes the proof of the  
828 lemma.  $\square$

## 829 C.2 Proof of Lemma C.2

830 Let  $\mathbf{1}_\otimes := \mathbf{1} \otimes \mathbf{I}$  and  $\mathbf{L} := \mathbf{A}^\top \mathbf{R} \mathbf{A}$ , we introduce the following quantities to facilitate the analysis:

$$\begin{aligned} \mathbf{e}_s^t &:= \alpha (\nabla \mathbf{f}(\mathbf{x}^t) - \nabla \mathbf{f}(\mathbf{x}^t; \xi^t)) + \gamma (\mathbf{L} - \mathbf{A}(\xi^t)^\top \mathbf{A}) \mathbf{x}^t \\ \mathbf{e}_g^t &:= \alpha (\nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^t) - \nabla \mathbf{f}(\mathbf{x}^t)) \end{aligned}$$

831 We can simplify the primal update as

$$\begin{aligned}\mathbf{x}^{t+1} &= (\mathbf{I} - \gamma\mathbf{L})\mathbf{x}^t - \eta\mathbf{A}^\top\boldsymbol{\lambda}^t - \alpha\nabla\mathbf{f}(\mathbf{1}_\otimes\bar{\mathbf{x}}^t) + \mathbf{e}_s^t + \mathbf{e}_g^t \\ &= (\mathbf{I} - \gamma\mathbf{L})\mathbf{x}^t - \eta\mathbf{v}^t + \mathbf{e}_s^t + \mathbf{e}_g^t\end{aligned}\quad (54)$$

832 By (54), the consensus error can be measured by

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2] &= \mathbb{E}[\|(\mathbf{I} - \gamma\mathbf{L})\mathbf{x}^t - \eta\mathbf{v}^t + \mathbf{e}_s^t + \mathbf{e}_g^t\|_{\mathbf{K}}^2] \\ &\stackrel{(i)}{=} \mathbb{E}[\|(\mathbf{I} - \gamma\mathbf{L})\mathbf{x}^t - \eta\mathbf{v}^t + \mathbf{e}_g^t\|_{\mathbf{K}}^2] + \mathbb{E}[\|\mathbf{e}_s^t\|_{\mathbf{K}}^2] \\ &= \mathbb{E}[\|(\mathbf{I} - \gamma\mathbf{L})\mathbf{x}^t\|_{\mathbf{K}}^2] + \mathbb{E}[\|\eta\mathbf{v}^t - \mathbf{e}_g^t\|_{\mathbf{K}}^2] - 2\mathbb{E}[\langle(\mathbf{I} - \gamma\mathbf{L})\mathbf{x}^t \mid \eta\mathbf{v}^t - \mathbf{e}_g^t\rangle_{\mathbf{K}}] + \mathbb{E}[\|\mathbf{e}_s^t\|_{\mathbf{K}}^2]\end{aligned}\quad (55)$$

833 where (i) uses the independence of random variables  $\mathbb{E}[\mathbf{e}_s^t|\mathbf{x}^t, \boldsymbol{\lambda}^t] = \mathbf{0}$ . Let  $\gamma \leq \rho_{\min}/\rho_{\max}^2$ , by  
834 Assumption 3.2, the first term of (55) can be bounded by

$$\mathbb{E}[\|(\mathbf{I} - \gamma\mathbf{L})\mathbf{x}^t\|_{\mathbf{K}}^2] \leq (1 - \gamma\rho_{\min})\mathbb{E}[\|\mathbf{x}^t\|_{\mathbf{K}}^2]. \quad (56)$$

835 The second term of (55) can be bounded by

$$\mathbb{E}[\|\eta\mathbf{v}^t - \mathbf{e}_g^t\|_{\mathbf{K}}^2] \leq 2\eta^2\mathbb{E}[\|\mathbf{v}^t\|_{\mathbf{K}}^2] + 2\mathbb{E}[\|\mathbf{e}_g^t\|_{\mathbf{K}}^2], \quad (57)$$

836 and the third term of (55) can be bounded by

$$\begin{aligned}&-2\mathbb{E}[\langle(\mathbf{I} - \gamma\mathbf{L})\mathbf{x}^t \mid \eta\mathbf{v}^t - \mathbf{e}_g^t\rangle_{\mathbf{K}}] \\ &\leq -2\eta\mathbb{E}[\langle\mathbf{x}^t \mid \mathbf{v}^t\rangle_{(\mathbf{I} - \gamma\mathbf{L})\mathbf{K}}] + \alpha\mathbb{E}[\|(\mathbf{I} - \gamma\mathbf{L})\mathbf{x}^t\|_{\mathbf{K}}^2] + \frac{1}{\alpha}\mathbb{E}[\|\mathbf{e}_g^t\|_{\mathbf{K}}^2]\end{aligned}\quad (58)$$

837 Gathering the above inequalities, we obtain the following recursion on the consensus error:

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2] &\leq \mathbb{E}[(1 + \alpha)(1 - \gamma\rho_{\min})\|\mathbf{x}^t\|_{\mathbf{K}}^2 + 2\eta^2\|\mathbf{v}^t\|_{\mathbf{K}}^2] \\ &\quad + \mathbb{E}[\|\mathbf{e}_s^t\|_{\mathbf{K}}^2 + (2 + \frac{1}{\alpha})\|\mathbf{e}_g^t\|_{\mathbf{K}}^2 - 2\eta\langle\mathbf{x}^t \mid \mathbf{v}^t\rangle_{\mathbf{K} - \gamma\mathbf{L}}].\end{aligned}$$

838 Now we tackle the error term  $\|\mathbf{e}_s^t\|^2$ , by the independence between  $\xi_a$  and  $\xi_1, \dots, \xi_n$ , it yields

$$\mathbb{E}[\|\mathbf{e}_s^t\|_{\mathbf{K}}^2] \leq \alpha^2 \sum_{i=1}^n \sigma_i^2 + \gamma^2 \sigma_A^2 \rho_{\max} \mathbb{E}[\|\mathbf{x}^t\|_{\mathbf{K}}^2], \quad (59)$$

839 where we have applied Assumptions 3.2, 3.3, 3.4 to obtain the above property. Moreover,

$$\mathbb{E}[\|\mathbf{e}_g^t\|_{\mathbf{K}}^2] \leq \alpha^2 L^2 \mathbb{E}[\|\mathbf{x}^t\|_{\mathbf{K}}^2]. \quad (60)$$

840 To simplify expression, we assume  $\alpha \leq 1$ . Combining the upper bounds of  $\mathbb{E}[\|\mathbf{e}_s^t\|_{\mathbf{K}}^2]$  and  $\mathbb{E}[\|\mathbf{e}_g^t\|_{\mathbf{K}}^2]$   
841 gives us

$$\begin{aligned}\mathbb{E}[\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2] &\leq [(1 + \alpha)(1 - \gamma\rho_{\min}) + \gamma^2 \sigma_A^2 \rho_{\max} + 3\alpha L^2] \mathbb{E}[\|\mathbf{x}^t\|_{\mathbf{K}}^2] \\ &\quad + 2\eta^2\mathbb{E}[\|\mathbf{v}^t\|_{\mathbf{K}}^2] - 2\eta\mathbb{E}[\langle\mathbf{x}^t \mid \mathbf{v}^t\rangle_{\mathbf{K} - \gamma\mathbf{L}}] + \alpha^2 \sum_{i=1}^n \sigma_i^2\end{aligned}$$

842 Using the step size condition (40) to simplify the first term completes the proof.  $\square$

### 843 C.3 Proof of Lemma C.3

844 Let  $\mathbf{1}_\otimes := \mathbf{1} \otimes \mathbf{I}$  and  $\mathbf{L} := \mathbf{A}^\top \mathbf{R} \mathbf{A}$ , we observe that  $\mathbf{v}^{t+1}$  is updated through the recursion:

$$\begin{aligned}\mathbf{v}^{t+1} &= \mathbf{A}^\top \boldsymbol{\lambda}^{t+1} + \frac{\alpha}{\eta} \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^{t+1}) \\ &= \mathbf{v}^t + \underbrace{\beta \mathbf{A}^\top \mathbf{A}(\xi^t) \mathbf{x}^t + \frac{\alpha}{\eta} \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^{t+1}) - \frac{\alpha}{\eta} \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^t)}_{=:\Delta \mathbf{v}^t}\end{aligned}\quad (61)$$

845 Therefore,

$$\begin{aligned} \mathbb{E} \left[ \|\mathbf{v}^{t+1}\|_{\mathbf{Q}+\mathbf{cK}}^2 \right] &\leq \mathbb{E} \left[ \|\mathbf{v}^t\|_{\mathbf{Q}+\mathbf{cK}}^2 \right] + 2\mathbb{E} \left[ \langle \mathbf{v}^t \mid \Delta \mathbf{v}^t \rangle_{\mathbf{Q}+\mathbf{cK}} \right] + 2\beta^2 \mathbb{E} \left[ \|\mathbf{A}^\top \mathbf{A}(\xi^t) \mathbf{x}^t\|_{\mathbf{Q}+\mathbf{cK}}^2 \right] \\ &\quad + \frac{2\alpha^2}{\eta^2} \mathbb{E} \left[ \|\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)\|_{\mathbf{Q}+\mathbf{cK}}^2 \right] \end{aligned} \quad (62)$$

846 The second term of (62) can be simplified as

$$\begin{aligned} \mathbb{E} \left[ \langle \mathbf{v}^t \mid \Delta \mathbf{v}^t \rangle_{\mathbf{Q}+\mathbf{cK}} \right] &\leq \beta \mathbb{E} \left[ \langle \mathbf{v}^t \mid \mathbf{x}^t \rangle_{\mathbf{K}+\mathbf{cL}} \right] + \alpha \mathbb{E} \left[ \|\mathbf{v}^t\|_{\mathbf{Q}+\mathbf{cK}}^2 \right] \\ &\quad + \frac{\alpha}{4\eta^2} \mathbb{E} \left[ \|\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)\|_{\mathbf{Q}+\mathbf{cK}}^2 \right] \end{aligned} \quad (63)$$

847 where we used  $\mathbb{E}[(\mathbf{Q} + \mathbf{cK})\mathbf{A}^\top \mathbf{A}(\xi^t)] = \mathbf{K} + \mathbf{cL}$  and note that  $\mathbb{E} \left[ \|\mathbf{v}^t\|_{\mathbf{Q}+\mathbf{cK}}^2 \right] \leq (\rho_{\min}^{-1} +$

848  $\mathbf{c})\mathbb{E} \left[ \|\mathbf{v}^t\|_{\mathbf{K}}^2 \right]$  due to Assumption 3.2. The third term of (62) can be simplified as

$$\begin{aligned} \mathbb{E} \left[ \|\mathbf{A}^\top \mathbf{A}(\xi^t) \mathbf{x}^t\|_{\mathbf{Q}+\mathbf{cK}}^2 \right] &\leq (\rho_{\min}^{-1} + \mathbf{c}) \mathbb{E} \left[ \|\mathbf{A}^\top \mathbf{A}(\xi^t) \mathbf{x}^t\|_{\mathbf{K}}^2 \right] \\ &= (\rho_{\min}^{-1} + \mathbf{c}) \mathbb{E} \left[ \|\mathbf{x}^t\|_{\mathbf{A}(\xi^t)^\top \mathbf{A} \mathbf{A}^\top \mathbf{A}(\xi^t)}^2 \right] \\ &\leq \rho(\mathbf{A} \mathbf{A}^\top) \bar{\rho}_{\max} (\rho_{\min}^{-1} + \mathbf{c}) \mathbb{E} \left[ \|\mathbf{x}^t\|_{\mathbf{K}}^2 \right] \\ &= \bar{\rho}_{\max}^2 (\rho_{\min}^{-1} + \mathbf{c}) \mathbb{E} \left[ \|\mathbf{x}^t\|_{\mathbf{K}}^2 \right]. \end{aligned}$$

849 The fourth term of (62) can be simplified using Lemma C.6 as

$$\begin{aligned} &\mathbb{E} \left[ \|\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)\|_{\mathbf{Q}+\mathbf{cK}}^2 \right] \\ &\leq \frac{4\alpha^2 n L^2}{(\rho_{\min}^{-1} + \mathbf{c})^{-1}} \left\{ \frac{1}{2n^2} \sum_{i=1}^n \sigma_i^2 + \mathbb{E} \left[ \|\nabla F(\bar{\mathbf{x}}^t)\|^2 \right] + \frac{L^2}{n} \mathbb{E} \left[ \|\mathbf{x}^t\|_{\mathbf{K}}^2 \right] \right\}. \end{aligned} \quad (64)$$

850 Combining the above inequalities and using the condition  $\alpha \leq 1$  to simplify constants yields the  
851 lemma.  $\square$

#### 852 C.4 Proof of Lemma C.4

853 Let  $\mathbf{1}_{\otimes} := \mathbf{1} \otimes \mathbf{I}$ ,  $\mathbf{L} := \mathbf{A}^\top \mathbf{R} \mathbf{A}$ , and denote:

$$\Delta \mathbf{v}^t := \beta \mathbf{L} \mathbf{x}^t + \frac{\alpha}{\eta} (\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)),$$

854 then by the recursions in (54) and (61),

$$\begin{aligned} &\mathbb{E} \left[ \langle \mathbf{x}^{t+1} \mid \mathbf{v}^{t+1} \rangle_{\mathbf{K}} \right] \\ &= \mathbb{E} \left[ \langle \mathbf{x}^t \mid \mathbf{v}^t \rangle_{\mathbf{K} - (\gamma + \eta\beta)\mathbf{L}} \right] + \mathbb{E} \left[ \|\mathbf{x}^t\|_{\beta(\mathbf{I} - \gamma\mathbf{L})\mathbf{L}}^2 \right] - \eta \mathbb{E} \left[ \|\mathbf{v}^t\|_{\mathbf{K}}^2 \right] \\ &\quad + \frac{\alpha}{\eta} \mathbb{E} \left[ \langle \mathbf{x}^t \mid \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t) \rangle_{(\mathbf{I} - \gamma\mathbf{L})\mathbf{K}} \right] - \alpha \mathbb{E} \left[ \langle \mathbf{v}^t \mid \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t) \rangle_{\mathbf{K}} \right] \\ &\quad + \mathbb{E} \left[ \langle \mathbf{e}_g^t \mid \mathbf{v}^t + \Delta \mathbf{v}^t \rangle_{\mathbf{K}} \right] + \frac{\alpha}{\eta} \mathbb{E} \left[ \langle \mathbf{e}_s^t \mid \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t) \rangle_{\mathbf{K}} \right], \end{aligned} \quad (65)$$

855 where we have used the fact  $\mathbb{E}[\mathbf{e}_s | \mathbf{x}^t] = \mathbf{0}$ . We first note that  $\mathbb{E} \left[ \|\mathbf{x}^t\|_{\beta(\mathbf{I} - \gamma\mathbf{L})\mathbf{L}}^2 \right] \leq \beta(\rho_{\max} -$   
856  $\gamma\rho_{\min}^2)\mathbb{E} \left[ \|\mathbf{x}^t\|_{\mathbf{K}}^2 \right]$ . By the Young's inequality, we get

$$\begin{aligned} &\mathbb{E} \left[ \langle \mathbf{x}^t \mid \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t) \rangle_{(\mathbf{I} - \gamma\mathbf{L})\mathbf{K}} \right] \\ &\leq \frac{1}{2} (1 - \gamma\rho_{\min}) \mathbb{E} \left[ \|\mathbf{x}^t\|_{\mathbf{K}}^2 \right] + \frac{1}{2} \mathbb{E} \left[ \|\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)\|^2 \right]. \end{aligned} \quad (66)$$

857 Moreover,

$$\begin{aligned} &\mathbb{E} \left[ \langle \mathbf{v}^t \mid \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t) \rangle_{\mathbf{K}} \right] \\ &\leq \frac{1}{2} \mathbb{E} \left[ \|\mathbf{v}^t\|_{\mathbf{K}}^2 \right] + \frac{1}{2} \mathbb{E} \left[ \|\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)\|^2 \right]. \end{aligned}$$

858 Next, it is derived in (60) that  $\mathbb{E} [\|\mathbf{e}_g^t\|_{\mathbf{K}}^2] \leq \alpha^2 L^2 \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2]$ . Thus,

$$\begin{aligned}
& \mathbb{E} [\langle \mathbf{e}_g^t \mid \mathbf{v}^t + \Delta \mathbf{v}^t \rangle_{\mathbf{K}}] \\
& \leq \frac{1}{2} \left( \frac{3}{\eta} \right) \mathbb{E} [\|\mathbf{e}_g^t\|_{\mathbf{K}}^2] + \frac{1}{2} \left( \frac{\eta}{3} \right) \mathbb{E} [\|\mathbf{v}^t + \Delta \mathbf{v}^t\|_{\mathbf{K}}^2] \\
& \leq \frac{3}{2\eta} \mathbb{E} [\|\mathbf{e}_g^t\|_{\mathbf{K}}^2] + \frac{\eta}{2} \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] + \frac{\eta\beta^2}{2} \mathbb{E} [\|\mathbf{L}\mathbf{x}^t\|_{\mathbf{K}}^2] + \frac{\eta}{2} \cdot \frac{\alpha^2}{\eta^2} \mathbb{E} [\|\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)\|_{\mathbf{K}}^2] \\
& \leq \left( \frac{3\alpha^2 L^2}{2\eta} + \frac{\eta\beta^2 \rho_{\max}^2}{2} \right) \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] + \frac{\eta}{2} \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] + \frac{\eta}{2} \cdot \frac{\alpha^2}{\eta^2} \mathbb{E} [\|\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)\|_{\mathbf{K}}^2].
\end{aligned} \tag{67}$$

859 Similarly, it is derived in (59) that  $\mathbb{E} [\|\mathbf{e}_s^t\|_{\mathbf{K}}^2] \leq \alpha^2 \sum_{i=1}^n \sigma_i^2 + \gamma^2 \sigma_A^2 \rho_{\max} \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2]$ . Thus,

$$\begin{aligned}
& \mathbb{E} [\langle \mathbf{e}_s^t \mid \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t) \rangle_{\mathbf{K}}] \\
& \leq \frac{1}{2} \mathbb{E} [\|\mathbf{e}_s^t\|_{\mathbf{K}}^2] + \frac{1}{2} \mathbb{E} [\|\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)\|_{\mathbf{K}}^2] \\
& \leq \frac{\alpha^2}{2} \sum_{i=1}^n \sigma_i^2 + \frac{\gamma^2 \sigma_A^2 \rho_{\max}}{2} \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] + \frac{1}{2} \mathbb{E} [\|\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)\|_{\mathbf{K}}^2].
\end{aligned} \tag{68}$$

860 Therefore, using the condition  $\eta \leq 1$  and simplifying terms yield

$$\begin{aligned}
\mathbb{E} [\langle \mathbf{x}^{t+1} \mid \mathbf{v}^{t+1} \rangle_{\mathbf{K}}] & \leq \mathbb{E} [\langle \mathbf{x}^t \mid \mathbf{v}^t \rangle_{\mathbf{K} - (\gamma + \eta\beta)\mathbf{L}}] + \frac{\alpha - \eta}{2} \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] + \frac{\alpha^3}{2\eta} \sum_{i=1}^n \sigma_i^2 \\
& \quad + \left\{ \beta \rho_{\max} + \frac{\alpha}{2\eta} + \frac{3\alpha^2 L^2}{2\eta} + \frac{\eta\beta^2 \rho_{\max}^2}{2} + \frac{\alpha\gamma^2 \sigma_A^2 \rho_{\max}}{2\eta} \right\} \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] \\
& \quad + \left( \frac{2\alpha}{\eta} + \frac{\alpha^2}{2\eta} \right) \mathbb{E} [\|\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)\|_{\mathbf{K}}^2].
\end{aligned} \tag{69}$$

861 The proof is concluded by applying Lemma C.6 to bound the last term.  $\square$

## 862 C.5 Proof of Lemma C.5

863 By combining the results of Lemma C.1, C.2, C.3, C.4, we obtain

$$F_{t+1} \leq F_t - \mathbb{C}_{\nabla F} \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] + \mathbb{C}_{\sigma} \alpha^2 \bar{\sigma}^2 / n + \mathbb{E} [\mathbb{C}_{\mathbf{x}} \|\mathbf{x}^t\|_{\mathbf{K}}^2 + \mathbb{C}_{\mathbf{v}} \|\mathbf{v}^t\|_{\mathbf{K}}^2 + \langle \mathbf{x}^t \mid \mathbf{v}^t \rangle_{\mathbf{C}_{xv}}] \tag{70}$$

864 where

$$\mathbb{C}_{\nabla F} := \frac{\alpha}{4} - \mathbf{b} \frac{10\alpha^3 L^2 n}{\eta^2} (\rho_{\min}^{-1} + \mathbf{c}) - \mathbf{d} \frac{10\alpha^3 L^2 n}{\eta} \tag{71}$$

$$\mathbb{C}_{\sigma} := \frac{L}{2} + \mathbf{a} n^2 + \mathbf{b} \frac{5n\alpha L^2}{\eta^2} (\rho_{\min}^{-1} + \mathbf{c}) + \mathbf{d} \left( \frac{5n\alpha L^2}{\eta} + \frac{\alpha n^2}{2\eta} \right) \tag{72}$$

$$\begin{aligned}
\mathbb{C}_{\mathbf{x}} &:= \frac{3\alpha L^2}{4n} + \mathbf{a} \left( -\frac{\gamma}{2} \rho_{\min} + \alpha(1 + 3L^2) \right) + \mathbf{b} \left( 2\beta^2 \bar{\rho}_{\max}^2 + \frac{10\alpha^3 L^4}{\eta^2} \right) (\rho_{\min}^{-1} + \mathbf{c}) \\
& \quad + \mathbf{d} \left( \beta \rho_{\max} + \frac{\alpha}{2\eta} + \frac{3\alpha^2 L^2}{2\eta} + \frac{\alpha\gamma^2 \sigma_A^2 \rho_{\max}}{2\eta} \right) + \mathbf{d} \left( \frac{10\alpha^3 L^4}{\eta} + \frac{\eta\beta^2 \rho_{\max}^2}{2} \right)
\end{aligned} \tag{73}$$

$$\mathbb{C}_{\mathbf{v}} := \mathbf{a}(2\eta^2) + \mathbf{b}(2\alpha(\rho_{\min}^{-1} + \mathbf{c})) + \mathbf{d} \frac{\alpha - \eta}{2} \tag{74}$$

$$\mathbf{C}_{xv} := (2\gamma\eta\mathbf{a} + 2\beta\mathbf{b}\mathbf{c} - \mathbf{d}(\gamma + \eta\beta)) \mathbf{A}^{\top} \mathbf{R} \mathbf{A} + (-2\eta\mathbf{a} + 2\beta\mathbf{b}) \mathbf{K}. \tag{75}$$

865 The equality  $\mathbf{C}_{xv} = \mathbf{0}$  is ensured by the parameter choices in (45). We then observe that for any  
 866  $\delta_0 > 0$ , it holds

$$\begin{aligned} F_t &\geq F(\bar{\mathbf{x}}^t) + \left( \mathbf{a} - \frac{\mathbf{d}}{2\delta_0} \right) \|\mathbf{x}^t\|_{\mathbf{K}}^2 + \|\mathbf{v}^t\|_{\mathbf{b}\mathbf{Q} + (\mathbf{bc} - \frac{\mathbf{d}\delta_0}{2})\mathbf{K}}^2 \\ &\geq \frac{1}{n} \sum_{i=1}^n f_i(\bar{\mathbf{x}}^t) + \|\mathbf{v}^t\|_{(\mathbf{b}\rho_{\max}^{-1} + \mathbf{bc} - \frac{\mathbf{d}^2}{4\mathbf{a}})\mathbf{K}}^2, \end{aligned}$$

867 where the second inequality is achieved by setting  $\delta = \frac{\mathbf{d}}{2\mathbf{a}}$  and using Assumption 3.4. Together with

$$\mathbf{b}\rho_{\max}^{-1} + \mathbf{bc} - \frac{\mathbf{d}^2}{4\mathbf{a}} = \left( \frac{\eta\mathbf{a}}{\beta\rho_{\max}} - \frac{\mathbf{d}^2}{4\mathbf{a}} \right) + \left( \frac{\gamma\mathbf{d}}{2\beta} - \frac{\eta\gamma\mathbf{a}}{\beta} \right) + \frac{\eta\mathbf{d}}{2} \geq 0$$

868 which is due to the step size choices

$$\mathbf{d} = \delta_1\eta\mathbf{a}, \quad \delta_1 \geq 8, \quad \eta\beta \leq \frac{4}{\delta_1^2\rho_{\max}}. \quad (76)$$

869 We obtain that  $F_t \geq F(\bar{\mathbf{x}}^t) > -\infty$ . Second, to upper bound  $\mathbb{C}_{\mathbf{v}}$ , under the additional condition  
 870  $\eta\beta \leq \gamma$ , and provided that

$$\begin{cases} \frac{2\alpha\eta}{\beta\rho_{\min}} \leq \frac{\eta^2}{3} & \Leftrightarrow \alpha \leq \frac{\eta\beta\rho_{\min}}{6} \\ \frac{2\delta_1\alpha\eta\gamma}{\beta} \leq \frac{\eta^2}{3} & \Leftrightarrow \alpha \leq \frac{\eta\beta}{6\delta_1\gamma} \\ \frac{\delta_1\alpha\eta}{2} \leq \frac{\eta^2}{3} & \Leftrightarrow \alpha \leq \frac{2\eta}{3\delta_1} \end{cases} \quad (77)$$

871 We obtain that

$$\mathbb{C}_{\mathbf{v}} \leq \mathbf{a}\eta \left( 2\eta + \frac{2\alpha}{\beta\rho_{\min}} + \frac{2\delta_1\alpha\gamma}{\beta} + \frac{\delta_1\alpha}{2} - \frac{\delta_1\eta}{2} \right) \leq -\mathbf{a}\eta^2.$$

872 Third, to upper bound  $\mathbb{C}_{\mathbf{x}}$ , under the conditions  $\alpha \leq \frac{1}{2L^2}$ ,  $\eta\beta \leq \gamma$ ,  $\eta\beta \leq \rho_{\max}^{-1}$ ,  $\gamma \leq \sigma_A^{-1}\rho_{\max}^{-1/2}$ , we  
 873 have

$$\begin{aligned} \mathbb{C}_{\mathbf{x}} &\leq \frac{3\alpha L^2}{4n} + \left( -\frac{\gamma}{2}\rho_{\min} + \alpha(1 + 3L^2 + 4\delta_1) \right. \\ &\quad \left. + (2\eta\beta\bar{\rho}_{\max}^2 + \frac{5\alpha}{2\eta\beta})(\rho_{\min}^{-1} + \delta_1\gamma) + \frac{3}{2}\delta_1\eta\beta\rho_{\max} \right) \mathbf{a} \end{aligned}$$

874 With the step size conditions:

$$\begin{cases} \alpha(1 + 3L^2 + 4\delta_1) \leq \frac{\gamma}{16}\rho_{\min} & \Leftrightarrow \alpha \leq \frac{\gamma}{16(1+3L^2+4\delta_1)}\rho_{\min} \\ 2\eta\beta\bar{\rho}_{\max}^2\rho_{\min}^{-1} \leq \frac{\gamma}{32}\rho_{\min} & \Leftrightarrow \eta\beta \leq \frac{\gamma}{64} \cdot \frac{\rho_{\min}^2}{\bar{\rho}_{\max}^2} \\ 2\delta_1\gamma\eta\beta\bar{\rho}_{\max}^2\rho_{\min}^{-1} \leq \frac{\gamma}{32}\rho_{\min} & \Leftrightarrow \eta\beta \leq \frac{1}{64\delta_1} \cdot \frac{\rho_{\min}}{\bar{\rho}_{\max}^2} \\ \frac{5\alpha}{2\eta\beta\rho_{\min}} \leq \frac{\gamma}{32}\rho_{\min} & \Leftrightarrow \alpha \leq \frac{\gamma\eta\beta}{80}\rho_{\min}^2 \\ \frac{5\delta_1\gamma\alpha}{2\eta\beta} \leq \frac{\gamma}{32}\rho_{\min} & \Leftrightarrow \alpha \leq \frac{\eta\beta}{80\delta_1}\rho_{\min} \\ \frac{3}{2}\delta_1\eta\beta\rho_{\max} \leq \frac{\gamma}{16}\rho_{\min} & \Leftrightarrow \eta\beta \leq \frac{\gamma}{24\delta_1} \cdot \frac{\rho_{\min}}{\rho_{\max}} \end{cases} \quad (78)$$

875 and  $\alpha \leq \frac{\mathbf{a}\gamma n\rho_{\min}}{6L^2}$ , it is guaranteed that

$$\mathbb{C}_{\mathbf{x}} \leq 3\alpha L^2/(4n) - \mathbf{a}\gamma\rho_{\min}/4 \leq -\mathbf{a}\gamma\rho_{\min}/8. \quad (79)$$

876 Fourth, to lower bound  $\mathbb{C}_{\nabla F}$ , under the condition  $\eta\beta \leq \gamma$  and the step size conditions

$$\begin{cases} \mathbf{a} \frac{10\alpha^3 L^2 n}{\eta\beta\rho_{\min}} \leq \frac{\alpha}{24} & \Leftrightarrow \alpha \leq \sqrt{\frac{\eta\beta\rho_{\min}}{240L^2 n \mathbf{a}}} \\ \mathbf{a} \frac{10\delta_1\alpha^3 \gamma L^2 n}{\eta\beta} \leq \frac{\alpha}{24} & \Leftrightarrow \alpha \leq \sqrt{\frac{\eta\beta}{240\delta_1\gamma L^2 n \mathbf{a}}} \\ \mathbf{a} \cdot 10\delta_1\alpha^3 L^2 n \leq \frac{\alpha}{24} & \Leftrightarrow \alpha \leq \sqrt{\frac{1}{240\delta_1 L^2 n \mathbf{a}}} \end{cases} \quad (80)$$

877 we have

$$\mathbb{C}_{\nabla F} \geq \frac{\alpha}{4} - \mathbf{a} \left( \frac{10\alpha^3 L^2 n}{\eta\beta\rho_{\min}} + \frac{5\alpha^3 L^2 n}{\eta^2\beta} (2\delta_1\gamma\eta) + 10\delta_1\alpha^3 L^2 n \right) \geq \frac{\alpha}{8}.$$

878 Finally, the condition  $\eta\beta \leq \gamma$  guarantees:

$$\mathbb{C}_{\sigma} \leq \frac{L}{2} + \mathbf{a}n^2 + \mathbf{a}n\alpha L^2 \left( \frac{5}{\eta\beta\rho_{\min}} + \frac{5\delta_1\gamma}{\eta\beta} + 5\delta_1 \right) + \frac{\mathbf{a}\delta_1 n^2 \alpha}{2}. \quad (81)$$

879 This shows the desired properties regarding  $\mathbb{C}_{\mathbf{x}}, \mathbb{C}_{\nabla F}, \mathbb{C}_{\sigma}, \mathbb{C}_{\mathbf{v}}$ . Gathering the step size conditions,  
 880 i.e., (76), (77), (78), (80) and simplifying yields (46).  $\square$

### 881 C.6 Proof of Corollary 3.7

882 The proof can be seen as a direct extension of (49) with Assumption 3.6. For instance, from (49) we  
883 get

$$\begin{aligned}
F_{t+1} - f_\star &\leq F_t - f_\star - \frac{\alpha}{8} \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] + \mathbb{C}_\sigma \alpha^2 \bar{\sigma}^2 - \frac{\mathbf{a}\gamma\rho_{\min}}{8} \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] - \mathbf{a}\eta^2 \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] \\
&= (1 - \delta)(F_t - f_\star) + \delta(\mathbb{E} [F(\bar{\mathbf{x}}^t)] - f_\star) - \frac{\alpha}{8} \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] + \mathbb{C}_\sigma \alpha^2 \bar{\sigma}^2 \\
&\quad + (\delta\mathbf{a} - \frac{\mathbf{a}\gamma\rho_{\min}}{8}) \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] + \delta\mathbf{b} \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{Q}}^2] + (\delta\mathbf{b}\mathbf{c} - \mathbf{a}\eta^2) \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] \\
&\quad + \delta\mathbf{d} \mathbb{E} [\langle \mathbf{x}^t \mid \mathbf{v}^t \rangle_{\mathbf{K}}],
\end{aligned} \tag{82}$$

884 for any  $\delta > 0$ . Applying Assumption 3.6,  $\mathbf{Q} \preceq \rho_{\min}^{-1} \mathbf{K}$  and  $\langle \mathbf{x}^t \mid \mathbf{v}^t \rangle_{\mathbf{K}} \leq \frac{1}{2} \|\mathbf{x}^t\|_{\mathbf{K}}^2 + \frac{1}{2} \|\mathbf{v}^t\|_{\mathbf{K}}^2$  gives  
885 us

$$\begin{aligned}
F_{t+1} - f_\star &\leq (1 - \delta)(F_t - f_\star) + \left(\frac{\delta}{2\mu} - \frac{\alpha}{8}\right) \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] + \mathbb{C}_\sigma \alpha^2 \bar{\sigma}^2 \\
&\quad + \left[\delta(\mathbf{a} + \frac{\mathbf{d}}{2}) - \frac{\mathbf{a}\gamma\rho_{\min}}{8}\right] \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] + \left[\delta(\frac{\mathbf{b}}{\rho_{\min}} + \mathbf{b}\mathbf{c} + \frac{\mathbf{d}}{2}) - \mathbf{a}\eta^2\right] \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2].
\end{aligned} \tag{83}$$

886 Then choosing  $\delta > 0$  such that  $\delta \leq \min\{\alpha\mu/4, \gamma\rho_{\min}/16, \eta\beta/(3\rho_{\min}), \eta/12\}$  enforces the  
887 coefficients of excessive terms to be non-positive, thus give rise to (20).  $\square$   
888

### 889 C.7 Auxiliary Lemma

890 **Lemma C.6.** Under Assumption 3.1 and 3.3,

$$\mathbb{E} [\|\nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^t)\|^2] \leq 4\alpha^2 n L^2 \left\{ \frac{1}{2n^2} \sum_{i=1}^n \sigma_i^2 + \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] + \frac{L^2}{n} \|\mathbf{x}^t\|_{\mathbf{K}}^2 \right\},$$

891 where we have denoted  $\mathbf{1}_\otimes := \mathbf{1} \otimes \mathbf{I}$ .

892 *Proof of Lemma C.6.* By the Lipschitz gradient assumption on each local objective function  $f_i$ ,

$$\mathbb{E} [\|\nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^t)\|^2] \leq n L^2 \mathbb{E} [\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2] = \frac{\alpha^2 L^2}{n} \mathbb{E} [\|\mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^t; \xi^t)\|^2].$$

893 The latter can be further expanded as

$$\begin{aligned}
&\frac{\alpha^2 L^2}{n} \mathbb{E} [\|\mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^t; \xi^t) - \mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^t) + \mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^t)\|^2] \\
&\leq \frac{2\alpha^2 L^2}{n} \left\{ \sum_{i=1}^n \sigma_i^2 + \mathbb{E} [\|\mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^t)\|^2] \right\}.
\end{aligned}$$

894 Lastly, we note that

$$\begin{aligned}
&\mathbb{E} [\|\mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^t)\|^2] \\
&\leq 2n^2 \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] + 2n \sum_{i=1}^n \mathbb{E} [\|\nabla f_i(\mathbf{x}_i^t) - \nabla f_i(\bar{\mathbf{x}}^t)\|^2] \\
&\leq 2n^2 \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] + 2n L^2 \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2].
\end{aligned}$$

895 This completes the proof.  $\square$   
896

### 897 D Proof of Theorem 3.9

898 The proof structure of Theorem 3.9 is similar to that of Theorem 3.5. Below we summarize the  
899 Lemmas that control the error quantities governing the convergence of FSPDA-STORM.

900 **Notations.** We introduce following shorthand notations to indicate the additional error factors.

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}; \xi) := \nabla \mathbf{f}(\mathbf{x}; \xi) + \frac{\eta}{\alpha} \mathbf{A}^\top \boldsymbol{\lambda} + \frac{\gamma}{\alpha} \mathbf{A}^\top \mathbf{A}(\xi) \mathbf{x} \quad (84)$$

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) := \nabla \mathbf{f}(\mathbf{x}) + \frac{\eta}{\alpha} \mathbf{A}^\top \boldsymbol{\lambda} + \frac{\gamma}{\alpha} \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x} \quad (85)$$

$$\bar{\mathbf{m}}_x^t := \frac{1}{n} \mathbf{1}_{\otimes}^\top \mathbf{m}_x^t \quad (86)$$

901 We recall that FSPDA-STORM can be described by the following system:

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha \mathbf{m}_x^t \quad (87)$$

$$\boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \beta \mathbf{m}_\lambda^t \quad (88)$$

$$\mathbf{m}_x^{t+1} = \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}; \xi^{t+1}) + (1 - a_x)(\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t; \xi^{t+1})) \quad (89)$$

$$\mathbf{m}_\lambda^{t+1} = \mathbf{A}(\xi^{t+1}) \mathbf{x}^{t+1} + (1 - a_\lambda)(\mathbf{m}_\lambda^t - \mathbf{A}(\xi^{t+1}) \mathbf{x}^t) \quad (90)$$

902 **Lemma D.1.** Under Assumption 3.1 and the step size condition  $\alpha \leq 1/(2L)$ ,

$$\mathbb{E}_t [F(\bar{\mathbf{x}}^{t+1})] \leq F(\bar{\mathbf{x}}^t) - \frac{\alpha}{4} \|\bar{\mathbf{m}}_x^t\|^2 - \frac{\alpha}{2} \|\nabla F(\bar{\mathbf{x}}^t)\|^2 + \frac{\alpha L^2}{n} \|\mathbf{x}^t\|_{\mathbf{K}}^2 + \alpha \left\| \bar{\mathbf{m}}_x^t - \frac{1}{n} \mathbf{1}_{\otimes}^\top \nabla \mathbf{f}(\mathbf{x}^t) \right\|^2 \quad (91)$$

903 See Appendix D.1 for the proof.

904 Notice that as  $\mathbf{1}_{\otimes}^\top \mathbf{A} = \mathbf{0}$ , the network average primal STORM estimator  $\bar{\mathbf{m}}_x^t$  evolve as

$$\bar{\mathbf{m}}_x^{t+1} = \frac{1}{n} \mathbf{1}_{\otimes}^\top \nabla \mathbf{f}(\mathbf{x}^{t+1}; \xi^{t+1}) + (1 - a_x)(\bar{\mathbf{m}}_x^t - \frac{1}{n} \mathbf{1}_{\otimes}^\top \nabla \mathbf{f}(\mathbf{x}^t; \xi^{t+1})), \quad (92)$$

905 i.e.,  $\bar{\mathbf{m}}_x^t$  tracks the global objective function gradient  $(1/n) \mathbf{1}_{\otimes}^\top \nabla \mathbf{f}(\mathbf{x}^t)$ . Meanwhile, the local estimator  
906  $\mathbf{m}_x^t$  tracks the Lagrangian gradient  $\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \mathbb{E} [\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}; \xi)]$ .

907 With Assumption 3.8, we are able to construct momentum estimation error bound where  $a_x$  controls  
908 the variance noise as in Lemma D.2 and D.3.

909 **Lemma D.2.** Under Assumption 3.1, 3.2, 3.3, 3.8 and the step size condition  $0 < a_x \leq 1$ ,

$$\mathbb{E} \left[ \left\| \bar{\mathbf{m}}_x^{t+1} - \frac{1}{n} \mathbf{1}_{\otimes}^\top \nabla \mathbf{f}(\mathbf{x}^{t+1}) \right\|^2 \right] \leq (1 - a_x)^2 \mathbb{E} \left[ \left\| \bar{\mathbf{m}}_x^t - \frac{1}{n} \mathbf{1}_{\otimes}^\top \nabla \mathbf{f}(\mathbf{x}^t) \right\|^2 \right] \quad (93)$$

$$\begin{aligned} &+ \frac{8(L_s^2 + L^2)\alpha^2}{n} \mathbb{E} [\|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2] + 32(L_s^2 + L^2)\alpha^2 \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] \\ &+ \frac{32(L_s^2 + L^2)\eta^2}{n} \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] + \frac{32(L_s^2 + L^2)(\alpha^2 L^2 + \gamma^2 \rho_{\max}^2)}{n} \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] + 2a_x^2 \bar{\sigma}^2 \end{aligned} \quad (94)$$

910 See Appendix D.2 for the proof.

911 **Lemma D.3.** Under Assumption 3.1, 3.2, 3.3, 3.4, 3.8 and the step size conditions  $\alpha \leq$   
912  $\sqrt{a_x/(64(L^2 + L_s^2))}$ ,  $\gamma \leq \sqrt{a_x/(64(\rho_{\max}^2 + \bar{\rho}_{\max}^2))}$ ,  $0 < a_x \leq 1$ ,  $\gamma \leq 1/(8\sqrt{a_x}\sigma_A^2)$ ,

$$\mathbb{E} [\|\mathbf{m}_x^{t+1} - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1})\|^2] \leq (1 - \frac{a_x}{4}) \mathbb{E} [\|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2] \quad (95)$$

$$+ 4a_x^2 n \bar{\sigma}^2 + 64\alpha^2 n G \cdot \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] + \left( \frac{16a_x^2 \eta^2 \gamma^2 \sigma_A^2}{\alpha^2} + 64\eta^2 G \right) \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] \quad (96)$$

$$+ \left( \frac{20a_x^2 \gamma^2 \sigma_A^2}{\alpha^2} + 64(\alpha^2 L^2 + \gamma^2 \rho_{\max}^2) G \right) \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] \quad (97)$$

913 where  $G := L_s^2 + L^2 + (\gamma^2/\alpha^2)\bar{\rho}_{\max}^2 + (\gamma^2/\alpha^2)\rho_{\max}^2$ .

914 See Appendix D.3 for the proof.



915 **Lemma D.4.** Under Assumption 3.1, 3.4 and the step size condition  $\alpha \leq$   
 916  $\gamma\rho_{\min}\sqrt{1-\gamma\rho_{\min}}/(\sqrt{32}L)$ ,

$$\mathbb{E} [\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2] \leq (1 - \frac{\gamma\rho_{\min}}{8})\mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] + \frac{\eta^2}{(1 - \gamma\rho_{\min})(1 - \gamma\rho_{\min}/2)}\mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] \quad (98)$$

$$- \frac{2\eta(1 - \gamma\rho_{\min}/4)}{(1 - \gamma\rho_{\min})} \langle \mathbf{x}^t \mid \mathbf{v}^t \rangle_{\mathbf{K} - \gamma\mathbf{A}^\top \mathbf{R} \mathbf{A}} + \frac{2\alpha^2}{\gamma\rho_{\min}} \mathbb{E} [\|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2] \quad (99)$$

917 See Appendix D.4 for the proof.

918 **Lemma D.5.** Under Assumption 3.1, 3.2 and the step size conditions  $\alpha \leq \min\{1/\sqrt{12}, 1/(2L)\}$ ,  
 919  $\eta\beta \leq 1/12$ ,  $\beta \leq 1$ ,

$$\mathbb{E} [\langle \mathbf{v}^{t+1} \mid \mathbf{x}^{t+1} \rangle_{\mathbf{K}}] \leq \mathbb{E} [\langle \mathbf{v}^t \mid \mathbf{x}^t \rangle_{\mathbf{K} - (\gamma + \eta\beta)\mathbf{A}^\top \mathbf{R} \mathbf{A}}] - \frac{\eta}{2} \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] \quad (100)$$

$$+ \left( \frac{3\alpha^2 L^2}{\eta} + 2\beta\rho_{\max}^2 - \gamma\beta\rho_{\min}^2 + \frac{5\beta}{2} + \frac{5\alpha^2}{2\eta} \right) \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] \quad (101)$$

$$+ \left( \left( \frac{3}{\eta} + \frac{1}{2} \right) \alpha^2 + 2\alpha^2\beta + \frac{2\alpha^4}{\eta} \right) \mathbb{E} [\|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2] \quad (102)$$

$$+ \frac{\beta}{2} \mathbb{E} [\|\mathbf{A}^\top \mathbf{m}_\lambda^t - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t\|^2] + \frac{\alpha^2 L^2 \eta}{2\eta} \mathbb{E} [\|\bar{\mathbf{m}}_x^t\|^2] \quad (103)$$

920 See Appendix D.5 for the proof.

921 **Lemma D.6.** Under Assumption 3.1, 3.4 and the step size conditions  $\alpha \leq \eta$ ,  $\beta \leq 1$ , for any constant  
 922  $\mathbf{b}, \mathbf{c} > 0$ ,

$$\mathbb{E} [\|\mathbf{v}^{t+1}\|_{\mathbf{bQ} + \mathbf{cK}}^2] \leq (1 + \beta + \frac{\alpha}{\eta}) \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{bQ} + \mathbf{cK}}^2] + 2\beta \mathbb{E} [\langle \mathbf{v}^t \mid \mathbf{x}^t \rangle_{\mathbf{bK} + \mathbf{cA}^\top \mathbf{R} \mathbf{A}}] \quad (104)$$

$$+ 3\beta^2 \rho_{\max}^2 (\mathbf{b}\rho_{\min}^{-1} + \mathbf{c}) \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] + 4\beta (\mathbf{b}\rho_{\min}^{-1} + \mathbf{c}) \mathbb{E} [\|\mathbf{A}^\top \mathbf{m}_\lambda^t - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t\|^2] \quad (105)$$

$$+ \frac{4n\alpha^3 L^2}{\eta} (\mathbf{b}\rho_{\min}^{-1} + \mathbf{c}) \mathbb{E} [\|\bar{\mathbf{m}}_x^t\|^2] \quad (106)$$

923 See Appendix D.6 for the proof.

924 **Lemma D.7.** Under Assumption 3.1, 3.2, 3.4 and the step size condition  $a_\lambda \leq \sigma_A^{-1}$ ,

$$\mathbb{E} [\|\mathbf{A}^\top \mathbf{m}_\lambda^{t+1} - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^{t+1}\|^2] \leq (1 - a_\lambda)^2 \mathbb{E} [\|\mathbf{A}^\top \mathbf{m}_\lambda^t - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t\|^2] \quad (107)$$

$$+ (10a_\lambda^2 \sigma_A^2 + 32(\rho_{\max}^2 + \bar{\rho}_{\max}^2)(\alpha^2 L^2 + \gamma^2 \rho_{\max}^2)) \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] + 16\eta^2 \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] \quad (108)$$

$$+ 8\alpha^2 (1 + \rho_{\max}^2 + \bar{\rho}_{\max}^2) \mathbb{E} [\|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2] + 32\alpha^2 n (\rho_{\max}^2 + \bar{\rho}_{\max}^2) \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2]$$

925 See Appendix D.7 for the proof.<sup>1</sup>

926 With the above Lemmas, we are ready to construct a potential function  $F_t$  that balances the interaction  
 927 between the error quantities:

<sup>1</sup>Note that the proof can be easily extended to the case when  $a_\lambda = 1$ , i.e., FSPDA-STORM without dual momentum, by ignoring the result of Lemma D.7 and applying  $\mathbf{m}_\lambda^t = \mathbf{A}^\top \mathbf{A}(\xi^t) \mathbf{x}^t$  together with Assumption 3.4.

**Theorem D.8.** For some constants  $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}, \mathbf{e}, \mathbf{f}, \mathbf{g} > 0$ , we define the potential function

$$F_t = \mathbb{E} \left[ F(\bar{\mathbf{x}}^t) + \mathbf{a} \|\mathbf{x}^t\|_{\mathbf{K}}^2 + \|\mathbf{v}^t\|_{\mathbf{b}\mathbf{Q} + \mathbf{c}\mathbf{K}}^2 + \mathbf{d} \langle \mathbf{x}^t \mid \mathbf{v}^t \rangle_{\mathbf{K}} + \mathbf{e} \|\bar{\mathbf{m}}_x^t - \frac{1}{n} \mathbf{1}_{\otimes}^\top \nabla \mathbf{f}(\mathbf{x}^t)\|^2 \right] \quad (109)$$

$$+ \mathbf{f} \|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2 + \mathbf{g} \|\mathbf{A}^\top \mathbf{m}_\lambda^t - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t\|^2 \quad (110)$$

Then, by the following choice of hyperparameters

$$\mathbf{a} = n^{-1}, \quad \mathbf{b} = \mathbf{a} \cdot \frac{\eta(1 - \gamma\rho_{\min}/4)}{\beta(1 - \gamma\rho_{\min})} = \mathcal{O}(n^{-1}T^{2/3}), \quad (111)$$

$$\mathbf{c} = \frac{\mathbf{d}}{2} \cdot \left( \frac{\gamma}{\beta} + \eta \right) - \mathbf{a} \cdot \frac{\eta\gamma}{\beta} (1 - \gamma\rho_{\min}/4)(1 - \gamma\rho_{\min})^{-1} = \mathcal{O}(T^{2/3}), \quad (112)$$

$$\mathbf{d} = \mathcal{O}(T^{1/3}), \quad \mathbf{e} = \mathcal{O}(a_x^{-1/2}) = \mathcal{O}(\bar{\sigma}^{2/3}T^{1/3}), \quad (113)$$

$$\mathbf{f} = \frac{\mathbf{a}}{\gamma} = \mathcal{O}(n^{-1}T^{1/3}), \quad \mathbf{g} = \mathcal{O}(n^{-1}T^{1/3}), \quad (114)$$

$$\alpha = \mathcal{O}(\bar{\sigma}^{-2/3}T^{-1/3}), \quad \eta = \mathcal{O}(n), \quad (115)$$

$$\gamma = \mathcal{O}(T^{-1/3}), \quad \beta = \mathcal{O}\left(\frac{\mathbf{a}}{\mathbf{d}} \cdot \gamma\right) = \mathcal{O}(n^{-1}T^{-2/3}), \quad (116)$$

$$a_x = \mathcal{O}(\bar{\sigma}^{-4/3}T^{-2/3}), \quad a_\lambda = \mathcal{O}(T^{-1/3}), \quad (117)$$

the potential function follows the inequality

$$F_{t+1} \leq F_t - \frac{\alpha}{4} \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] - \frac{\alpha}{8} \mathbb{E} [\|\bar{\mathbf{m}}_x^t\|^2] + (\mathbf{e} \cdot 2a_x^2 + \mathbf{f} \cdot 4a_x^2n)\bar{\sigma}^2 \quad (118)$$

$$- \mathbf{a} \cdot \frac{\gamma\rho_{\min}}{8} \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] - \mathbf{d} \cdot \frac{\eta}{4} \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] - \mathbf{e} \cdot \frac{a_x}{2} \mathbb{E} \left[ \left\| \bar{\mathbf{m}}_x^t - \frac{1}{n} \mathbf{1}_{\otimes}^\top \nabla \mathbf{f}(\mathbf{x}^t) \right\|^2 \right] \quad (119)$$

$$- \mathbf{f} \cdot \frac{a_x}{8} \mathbb{E} [\|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2] - \mathbf{g} \cdot \frac{a_\lambda}{2} \mathbb{E} [\|\mathbf{A}^\top \mathbf{m}_\lambda^t - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t\|^2] \quad (120)$$

See Appendix D.8 for the proof.

Finally, summing up (120) from  $t = 0$  to  $t = T - 1$  give us the convergence bound for the network average iterate and consensus error as

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] \leq \frac{F_0 - F_T}{T\alpha/4} + \frac{(\mathbf{e} \cdot 2a_x^2 + \mathbf{f} \cdot 4a_x^2n)\bar{\sigma}^2}{\alpha/4} = \mathcal{O}\left(\frac{F_0 - F_T + \bar{\sigma}^{2/3}}{T^{2/3}}\right) \quad (121)$$

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] \leq \frac{F_0 - F_T}{T\mathbf{a}\gamma\rho_{\min}/8} + \frac{(\mathbf{e} \cdot 2a_x^2 + \mathbf{f} \cdot 4a_x^2n)\bar{\sigma}^2}{\mathbf{a}\gamma\rho_{\min}/8} = \mathcal{O}\left(\frac{F_0 - F_T + \bar{\sigma}^{2/3}}{n^{-1}\rho_{\min}T^{2/3}}\right) \quad (122)$$

for large enough  $T$ .

### D.1 Proof of Lemma D.1

By Assumption 3.1,

$$\mathbb{E}_t[F(\bar{\mathbf{x}}^{t+1})] \quad (123)$$

$$= F(\bar{\mathbf{x}}^t) - \alpha \langle \nabla F(\bar{\mathbf{x}}^t) \mid \bar{\mathbf{m}}_x^t \rangle + \frac{\alpha^2 L}{2} \|\bar{\mathbf{m}}_x^t\|^2 \quad (124)$$

$$= F(\bar{\mathbf{x}}^t) - \left( \frac{\alpha}{2} + \frac{\alpha^2 L}{2} \right) \|\bar{\mathbf{m}}_x^t\|^2 - \frac{\alpha}{2} \|\nabla F(\bar{\mathbf{x}}^t)\|^2 + \frac{\alpha}{2} \|\bar{\mathbf{m}}_x^t - \nabla F(\bar{\mathbf{x}}^t)\|^2 \quad (125)$$

$$\stackrel{(i)}{\leq} F(\bar{\mathbf{x}}^t) - \frac{\alpha}{4} \|\bar{\mathbf{m}}_x^t\|^2 - \frac{\alpha}{2} \|\nabla F(\bar{\mathbf{x}}^t)\|^2 + \alpha \left\| \bar{\mathbf{m}}_x^t - \frac{1}{n} \mathbf{1}_{\otimes}^\top \nabla \mathbf{f}(\mathbf{x}^t) \right\|^2 + \alpha \left\| \frac{1}{n} \mathbf{1}_{\otimes}^\top \nabla \mathbf{f}(\mathbf{x}^t) - \nabla F(\bar{\mathbf{x}}^t) \right\|^2$$

where (i) uses the step size condition  $\alpha \leq 1/(2L)$ , and applying Assumption 3.1 completes the proof.

□

937 **D.2 Proof of Lemma D.2**

938 By the network average momentum update in (92),

$$\mathbb{E} \left[ \left\| \bar{\mathbf{m}}_x^{t+1} - \frac{1}{n} \mathbf{1}_{\otimes}^{\top} \nabla \mathbf{f}(\mathbf{x}^{t+1}) \right\|^2 \right] \quad (126)$$

$$= \mathbb{E} \left[ \left\| (1 - a_x) \left( \bar{\mathbf{m}}_x^t - \frac{1}{n} \mathbf{1}_{\otimes}^{\top} \nabla \mathbf{f}(\mathbf{x}^t) \right) + a_x \frac{1}{n} \mathbf{1}_{\otimes}^{\top} (\nabla \mathbf{f}(\mathbf{x}^{t+1}; \xi^{t+1}) - \nabla \mathbf{f}(\mathbf{x}^{t+1})) \right. \right. \quad (127)$$

$$\left. + (1 - a_x) \frac{1}{n} \mathbf{1}_{\otimes}^{\top} (\nabla \mathbf{f}(\mathbf{x}^t) - \nabla \mathbf{f}(\mathbf{x}^t; \xi^{t+1}) - (\nabla \mathbf{f}(\mathbf{x}^{t+1}) - \nabla \mathbf{f}(\mathbf{x}^{t+1}; \xi^{t+1}))) \right\|^2 \right] \quad (128)$$

$$\leq (1 - a_x)^2 \mathbb{E} \left[ \left\| \bar{\mathbf{m}}_x^t - \frac{1}{n} \mathbf{1}_{\otimes}^{\top} \nabla \mathbf{f}(\mathbf{x}^t) \right\|^2 \right] + 2a_x^2 \mathbb{E} \left[ \left\| \frac{1}{n} \mathbf{1}_{\otimes}^{\top} (\nabla \mathbf{f}(\mathbf{x}^{t+1}; \xi^{t+1}) - \nabla \mathbf{f}(\mathbf{x}^{t+1})) \right\|^2 \right] \quad (129)$$

$$+ 2(1 - a_x)^2 \mathbb{E} \left[ \left\| \frac{1}{n} \mathbf{1}_{\otimes}^{\top} (\nabla \mathbf{f}(\mathbf{x}^t) - \nabla \mathbf{f}(\mathbf{x}^{t+1})) - (\nabla \mathbf{f}(\mathbf{x}^t; \xi^{t+1}) - \nabla \mathbf{f}(\mathbf{x}^{t+1}; \xi^{t+1})) \right\|^2 \right]$$

939 Now observe that

$$2a_x^2 \mathbb{E} \left[ \left\| \frac{1}{n} \mathbf{1}_{\otimes}^{\top} (\nabla \mathbf{f}(\mathbf{x}^{t+1}; \xi^{t+1}) - \nabla \mathbf{f}(\mathbf{x}^{t+1})) \right\|^2 \right] \leq 2a_x^2 \bar{\sigma}^2, \quad (130)$$

940 and

$$2(1 - a_x)^2 \mathbb{E} \left[ \left\| \frac{1}{n} \mathbf{1}_{\otimes}^{\top} (\nabla \mathbf{f}(\mathbf{x}^t) - \nabla \mathbf{f}(\mathbf{x}^{t+1})) - (\nabla \mathbf{f}(\mathbf{x}^t; \xi^{t+1}) - \nabla \mathbf{f}(\mathbf{x}^{t+1}; \xi^{t+1})) \right\|^2 \right] \quad (131)$$

$$\leq 4(1 - a_x)^2 \mathbb{E} \left[ \left\| \frac{1}{n} \mathbf{1}_{\otimes}^{\top} (\nabla \mathbf{f}(\mathbf{x}^t) - \nabla \mathbf{f}(\mathbf{x}^{t+1})) \right\|^2 \right] \quad (132)$$

$$+ 4(1 - a_x)^2 \mathbb{E} \left[ \left\| \frac{1}{n} \mathbf{1}_{\otimes}^{\top} (\nabla \mathbf{f}(\mathbf{x}^t; \xi^{t+1}) - \nabla \mathbf{f}(\mathbf{x}^{t+1}; \xi^{t+1})) \right\|^2 \right] \quad (133)$$

$$\leq \frac{4(1 - a_x)^2}{n} (L^2 + L_s^2) \mathbb{E} [\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2] \quad (134)$$

941 Finally, applying Lemma D.9 and utilizing the step size condition  $0 < a_x \leq 1$  to simplify the  
 942 coefficients will complete the proof.  $\square$

943 **D.3 Proof of Lemma D.3**

$$\mathbb{E} \left[ \left\| \mathbf{m}_x^{t+1} - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}) \right\|^2 \right] \quad (135)$$

$$= \mathbb{E} \left[ \left\| (1 - a_x) \left( \mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t) \right) + a_x \left( \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}; \xi^{t+1}) - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}) \right) \right. \right. \quad (136)$$

$$\left. + (1 - a_x) \left( \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t) - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t; \xi^{t+1}) \right) \right\|^2 \right] \quad (137)$$

$$+ (1 - a_x) \left\| \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}; \xi^{t+1}) - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}) \right\|^2 \right] \quad (138)$$

$$\leq (1 - a_x)^2 \mathbb{E} [\left\| \mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t) \right\|^2] \quad (139)$$

$$+ 2a_x^2 \mathbb{E} [\left\| \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}; \xi^{t+1}) - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}) \right\|^2] \quad (140)$$

$$+ 2(1 - a_x)^2 \mathbb{E} [\left\| \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}; \xi^{t+1}) - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t; \xi^{t+1}) \right. \quad (141)$$

$$\left. - (\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}) - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)) \right\|^2] \quad (142)$$

944 Now observe that

$$2a_x^2 \mathbb{E} [\|\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}; \xi^{t+1}) - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1})\|^2] \quad (143)$$

$$= 2a_x^2 \mathbb{E} \left[ \left\| \nabla \mathbf{f}(\mathbf{x}^{t+1}; \xi^{t+1}) - \nabla \mathbf{f}(\mathbf{x}^{t+1}) + \frac{\gamma}{\alpha} \mathbf{A}^\top \mathbf{A}(\xi^{t+1}) \mathbf{x}^{t+1} - \frac{\gamma}{\alpha} \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^{t+1} \right\|^2 \right] \quad (144)$$

$$\leq 4a_x^2 n \bar{\sigma}^2 + \frac{4a_x^2 \gamma^2 \sigma_A^2}{\alpha^2} \mathbb{E} [\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2] \quad (145)$$

$$\leq 4a_x^2 n \bar{\sigma}^2 + \frac{20a_x^2 \gamma^2 \sigma_A^2}{\alpha^2} \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] + \frac{16a_x^2 \eta^2 \gamma^2 \sigma_A^2}{\alpha^2} \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] \quad (146)$$

$$+ 16a_x^2 \gamma^2 \sigma_A^2 \mathbb{E} [\|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2] \quad (147)$$

945 where the last inequality is expanded from the primal update rule  $\mathbf{x}^{t+1} = (1 - \gamma \mathbf{A}^\top \mathbf{R} \mathbf{A}) \mathbf{x}^t - \eta \mathbf{v}^t -$   
 946  $\alpha(\nabla \mathbf{f}(\mathbf{x}^t) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)) - \alpha(\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t))$  and the step size condition  $\alpha \leq 1/(2L)$ . On the  
 947 other hand,

$$\mathbb{E} [\|\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}; \xi^{t+1}) - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t; \xi^{t+1}) - (\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}) - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t))\|^2] \\ = \mathbb{E} [\|\nabla \mathbf{f}(\mathbf{x}^{t+1}; \xi^{t+1}) - \nabla \mathbf{f}(\mathbf{x}^t; \xi^{t+1}) - (\nabla \mathbf{f}(\mathbf{x}^{t+1}) - \nabla \mathbf{f}(\mathbf{x}^t))\|^2] \quad (148)$$

$$+ \frac{\gamma}{\alpha} (\mathbf{A}^\top \mathbf{A}(\xi^{t+1})(\mathbf{x}^{t+1} - \mathbf{x}^t) - \mathbf{A}^\top \mathbf{R} \mathbf{A}(\mathbf{x}^{t+1} - \mathbf{x}^t))\|^2] \quad (149)$$

$$\leq 4 \left( L_s^2 + L^2 + \frac{\gamma^2}{\alpha^2} \bar{\rho}_{\max}^2 + \frac{\gamma^2}{\alpha^2} \rho_{\max}^2 \right) \mathbb{E} [\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2] \quad (150)$$

948 Finally, applying Lemma D.9 and utilizing the step size conditions  $\alpha \leq \sqrt{a_x/(64(L^2 + L_s^2))}$ ,  
 949  $\gamma \leq \sqrt{a_x/(64(\rho_{\max}^2 + \bar{\rho}_{\max}^2))}$ ,  $0 < a_x \leq 1$ ,  $\gamma \leq 1/(8\sqrt{a_x} \sigma_A^2)$  to simplify the coefficients will  
 950 complete the proof.  $\square$

#### 951 D.4 Proof of Lemma D.4

$$\mathbb{E} [\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2] = \mathbb{E} [\|\mathbf{x}^t - \alpha \mathbf{m}_x^t\|_{\mathbf{K}}^2] \quad (151)$$

$$= \mathbb{E} [\|\mathbf{x}^t - \alpha \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t) - \alpha(\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t))\|_{\mathbf{K}}^2] \quad (152)$$

$$\leq (1 + z_1) \mathbb{E} [\|\mathbf{x}^t - \alpha \nabla \mathbf{f}(\mathbf{x}^t) - \eta \mathbf{A}^\top \boldsymbol{\lambda}^t - \gamma \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t\|_{\mathbf{K}}^2] \quad (153)$$

$$+ (1 + z_1^{-1}) \alpha^2 \mathbb{E} [\|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2] \quad \text{for any } z_1 > 0, \quad (154)$$

952 Now observe that

$$\mathbb{E} [\|\mathbf{x}^t - \alpha \nabla \mathbf{f}(\mathbf{x}^t) - \eta \mathbf{A}^\top \boldsymbol{\lambda}^t - \gamma \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t\|_{\mathbf{K}}^2] \quad (155)$$

$$= \mathbb{E} [\|(\mathbf{I} - \gamma \mathbf{A}^\top \mathbf{R} \mathbf{A}) \mathbf{x}^t - \eta \mathbf{v}^t - \alpha(\nabla \mathbf{f}(\mathbf{x}^t) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t))\|_{\mathbf{K}}^2] \quad (156)$$

$$\leq (1 + z_2) \mathbb{E} [\|(\mathbf{I} - \gamma \mathbf{A}^\top \mathbf{R} \mathbf{A}) \mathbf{x}^t - \eta \mathbf{v}^t\|_{\mathbf{K}}^2] \quad (157)$$

$$+ (1 + z_2^{-1}) \alpha^2 \mathbb{E} [\|\nabla \mathbf{f}(\mathbf{x}^t) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)\|^2] \quad \text{for any } z_2 > 0, \quad (158)$$

$$\leq (1 + z_2)(1 - \gamma \rho_{\min}) \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] + (1 + z_2) \eta^2 \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] \quad (159)$$

$$- 2(1 + z_2) \eta \langle \mathbf{x}^t | \mathbf{v}^t \rangle_{\mathbf{K} - \gamma \mathbf{A}^\top \mathbf{R} \mathbf{A}} + (1 + z_2^{-1}) \alpha^2 L^2 \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] \quad (160)$$

953 Then, we choose  $z_1 = \frac{\gamma \rho_{\min}/2}{1 - \gamma \rho_{\min}}$  so that  $(1 + z_1)(1 - \gamma \rho_{\min}) = 1 - \gamma \rho_{\min}/2$  and similarly  $z_2 =$   
 954  $\frac{\gamma \rho_{\min}/4}{1 - \gamma \rho_{\min}/2}$  so that  $(1 + z_1)(1 + z_2)(1 - \gamma \rho_{\min}) = 1 - \gamma \rho_{\min}/4$ . Finally, observing that

$$(1 + z_1)(1 + z_2^{-1}) \alpha^2 L^2 \leq \frac{4\alpha^2 L^2}{\gamma \rho_{\min}(1 - \gamma \rho_{\min})} \leq \frac{\gamma \rho_{\min}}{8} \quad (161)$$

955 when imposing  $\alpha \leq \gamma \rho_{\min} \sqrt{1 - \gamma \rho_{\min}} / (\sqrt{32} L)$  completes the proof.  $\square$

#### 956 D.5 Proof of Lemma D.5

957 By the update rules, we have

$$\mathbf{v}^{t+1} = \mathbf{v}^t + \beta \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t + \beta \mathbf{A}^\top (\mathbf{m}_\lambda^t - \mathbf{R} \mathbf{A} \mathbf{x}^t) + \frac{\alpha}{\eta} (\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)) \quad (162)$$

$$\mathbf{x}^{t+1} = (1 - \gamma \mathbf{A}^\top \mathbf{R} \mathbf{A}) \mathbf{x}^t - \eta \mathbf{v}^t - \alpha(\nabla \mathbf{f}(\mathbf{x}^t) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)) - \alpha(\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)) \quad (163)$$

Therefore, for any constants  $z_1, z_2 > 0$ ,

$$\mathbb{E} [\langle \mathbf{v}^{t+1} \mid \mathbf{x}^{t+1} \rangle_{\mathbf{K}}] \leq \mathbb{E} [\langle \mathbf{v}^t \mid \mathbf{x}^t \rangle_{\mathbf{K} - (\gamma + \eta\beta)\mathbf{A}^\top \mathbf{R} \mathbf{A}}] + (-1 + \frac{z_1}{2} + \frac{z_2}{2})\eta \cdot \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] \quad (164)$$

$$+ \left( \frac{\alpha^2 L^2}{2z_1 \eta} + \beta \rho_{\max}^2 + \frac{\beta^2}{2} \rho_{\max}^2 + \frac{\alpha^2 L^2}{2} - \gamma \beta \rho_{\min}^2 + \frac{\beta^2}{2} \rho_{\max}^2 \right) \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] \quad (165)$$

$$+ \left( \frac{\alpha^2}{2z_2 \eta} + \frac{\alpha^2}{2} \right) \mathbb{E} [\|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2] + \frac{\beta}{2} \mathbb{E} [\|\mathbf{A}^\top \mathbf{m}_\lambda^t - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t\|_{\mathbf{K}}^2] \quad (166)$$

$$+ \left( \frac{\beta}{2} + \frac{\alpha^2}{2\eta} \right) \mathbb{E} [\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2] + \frac{L^2 n}{2\eta} \mathbb{E} [\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2] \quad (167)$$

Note that by (163), we have

$$\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2 \leq 4(1 - \gamma \rho_{\min})^2 \|\mathbf{x}^t\|_{\mathbf{K}}^2 + 4\eta^2 \|\mathbf{v}^t\|_{\mathbf{K}}^2 + 4\alpha^2 L^2 \|\mathbf{x}^t\|_{\mathbf{K}}^2 + 4\alpha^2 \|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2 \quad (168)$$

Now choose  $z_1 = z_2 = 1/6$ , then applying  $\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2 = \alpha^2 \|\bar{\mathbf{m}}_x^t\|^2$  and utilizing the step size conditions  $\alpha \leq \min\{1/\sqrt{12}, 1/(2L)\}$ ,  $\eta\beta \leq 1/12$ ,  $\beta \leq 1$  to simplify the coefficients will complete the proof.  $\square$

## D.6 Proof of Lemma D.6

By the dual update rule, we have

$$\mathbb{E} [\|\mathbf{v}^{t+1}\|_{\mathbf{bQ} + \mathbf{cK}}^2] \quad (169)$$

$$= \mathbb{E} \left[ \|\mathbf{v}^t + \beta \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t + \beta \mathbf{A}^\top (\mathbf{m}_\lambda^t - \mathbf{R} \mathbf{A} \mathbf{x}^t) + \frac{\alpha}{\eta} (\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t))\|_{\mathbf{bQ} + \mathbf{cK}}^2 \right] \\ \leq \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{bQ} + \mathbf{cK}}^2] + \mathbb{E} [\|\mathbf{v}^{t+1} - \mathbf{v}^t\|_{\mathbf{bQ} + \mathbf{cK}}^2] \quad (170)$$

$$+ 2\mathbb{E} \left[ \left\langle \mathbf{v}^t \mid \beta \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t + \beta \mathbf{A}^\top (\mathbf{m}_\lambda^t - \mathbf{R} \mathbf{A} \mathbf{x}^t) + \frac{\alpha}{\eta} (\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)) \right\rangle_{\mathbf{bQ} + \mathbf{cK}} \right] \\ \leq (1 + \beta + \frac{\alpha}{\eta}) \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{bQ} + \mathbf{cK}}^2] + 2\beta \mathbb{E} [\langle \mathbf{v}^t \mid \mathbf{x}^t \rangle_{\mathbf{bK} + \mathbf{cA}^\top \mathbf{R} \mathbf{A}}] \quad (171)$$

$$+ (\frac{\alpha}{\eta} + \frac{3\alpha^2}{\eta^2}) \mathbb{E} [\|\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)\|_{\mathbf{bQ} + \mathbf{cK}}^2] + 3\beta^2 \rho_{\max}^2 \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{bQ} + \mathbf{cK}}^2] \quad (172)$$

$$+ (\beta + 3\beta^2) \mathbb{E} [\|\mathbf{A}^\top \mathbf{m}_\lambda^t - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t\|_{\mathbf{bQ} + \mathbf{cK}}^2] \quad (173)$$

$$\leq (1 + \beta + \frac{\alpha}{\eta}) \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{bQ} + \mathbf{cK}}^2] + 2\beta \mathbb{E} [\langle \mathbf{v}^t \mid \mathbf{x}^t \rangle_{\mathbf{bK} + \mathbf{cA}^\top \mathbf{R} \mathbf{A}}] \quad (174)$$

$$+ \frac{4n\alpha L^2}{\eta} (\mathbf{b}\rho_{\min}^{-1} + \mathbf{c}) \mathbb{E} [\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2] + 3\beta^2 \rho_{\max}^2 (\mathbf{b}\rho_{\min}^{-1} + \mathbf{c}) \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] \quad (175)$$

$$+ 4\beta (\mathbf{b}\rho_{\min}^{-1} + \mathbf{c}) \mathbb{E} [\|\mathbf{A}^\top \mathbf{m}_\lambda^t - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t\|^2] \quad (176)$$

Then, applying  $\|\bar{\mathbf{x}}^{t+1} - \bar{\mathbf{x}}^t\|^2 = \alpha^2 \|\bar{\mathbf{m}}_x^t\|^2$  and utilizing the step size conditions  $\alpha \leq \eta$ ,  $\beta \leq 1$  to simplify the coefficients will complete the proof.  $\square$

## D.7 Proof of Lemma D.7

$$\mathbb{E} [\|\mathbf{A}^\top \mathbf{m}_\lambda^{t+1} - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^{t+1}\|^2] \quad (177)$$

$$= \mathbb{E} [\|(1 - a_\lambda)(\mathbf{A}^\top \mathbf{m}_\lambda^t - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t) + a_\lambda(\mathbf{A}^\top \mathbf{A}(\xi^{t+1})\mathbf{x}^{t+1} - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^{t+1}) \\ + (1 - a_\lambda)(\mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t - \mathbf{A}^\top \mathbf{A}(\xi^{t+1})\mathbf{x}^t - (\mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^{t+1} - \mathbf{A}^\top \mathbf{A}(\xi^{t+1})\mathbf{x}^{t+1}))\|^2] \quad (178)$$

$$\leq (1 - a_\lambda)^2 \mathbb{E} [\|\mathbf{A}^\top \mathbf{m}_\lambda^t - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t\|^2] + 2a_\lambda^2 \mathbb{E} [\|\mathbf{A}^\top \mathbf{A}(\xi^{t+1})\mathbf{x}^{t+1} - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^{t+1}\|^2] \quad (179)$$

$$+ 2(1 - a_\lambda)^2 \mathbb{E} [\|\mathbf{A}^\top \mathbf{R} \mathbf{A}(\mathbf{x}^t - \mathbf{x}^{t+1}) - \mathbf{A}^\top \mathbf{A}(\xi^{t+1})(\mathbf{x}^t - \mathbf{x}^{t+1})\|^2] \quad (180)$$

$$\leq (1 - a_\lambda)^2 \mathbb{E} [\|\mathbf{A}^\top \mathbf{m}_\lambda^t - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t\|^2] + 2a_\lambda^2 \sigma_A^2 \mathbb{E} [\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2] \quad (181)$$

$$+ 4(1 - a_\lambda)^2 (\rho_{\max}^2 + \bar{\rho}_{\max}^2) \mathbb{E} [\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2] \quad (182)$$

968 Now note that by the primal update rule  $\mathbf{x}^{t+1} = (1 - \gamma \mathbf{A}^\top \mathbf{R} \mathbf{A}) \mathbf{x}^t - \eta \mathbf{v}^t - \alpha (\nabla \mathbf{f}(\mathbf{x}^t) - \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^t)) -$   
 969  $\alpha (\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t))$ , we have

$$\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2 \leq 4(1 - \gamma \rho_{\min})^2 \|\mathbf{x}^t\|_{\mathbf{K}}^2 + 4\eta^2 \|\mathbf{v}^t\|_{\mathbf{K}}^2 + 4\alpha^2 L^2 \|\mathbf{x}^t\|_{\mathbf{K}}^2 + 4\alpha^2 \|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2 \quad (183)$$

970 Finally, applying Lemma D.9 and the step size condition  $a_\lambda \leq \sigma_A^{-1}$  to simplify the coefficients will  
 971 complete the proof.  $\square$

## 972 D.8 Proof of Theorem D.8

973 Combining the results of Lemma D.1, D.4, D.6, D.5, D.2, D.3, D.7, when the step sizes satisfy

$$\alpha \leq \min \left\{ \frac{1}{2L}, \sqrt{\frac{a_x}{64(L^2 + L_s^2)}}, \frac{\gamma \rho_{\min} \sqrt{1 - \gamma \rho_{\min}}}{\sqrt{32}L}, \frac{1}{\sqrt{12}}, \frac{1}{2L}, \eta \right\}, \quad (184)$$

$$\gamma \leq \min \left\{ \sqrt{\frac{a_x}{64(\rho_{\max}^2 + \bar{\rho}_{\max}^2)}}, \frac{1}{8\sqrt{a_x} \sigma_A^2} \right\}, \quad (185)$$

$$\eta \beta \leq \frac{1}{12}, \quad \beta \leq 1, \quad 0 < a_x \leq 1, \quad a_\lambda \leq \frac{1}{\sigma_A}, \quad (186)$$

974 we obtain

$$F_{t+1} \leq F_t + \mathbb{C}_{\nabla F} \mathbb{E} [\|\nabla F(\bar{\mathbf{x}}^t)\|^2] + \mathbb{C}_\sigma \bar{\sigma}^2 + \mathbb{C}_{\mathbf{x}} \mathbb{E} [\|\mathbf{x}^t\|_{\mathbf{K}}^2] + \mathbb{C}_{\mathbf{v}} \mathbb{E} [\|\mathbf{v}^t\|_{\mathbf{K}}^2] \quad (187)$$

$$+ \langle \mathbf{x}^t \mid \mathbf{v}^t \rangle_{\mathbf{C}_{xv}} + \mathbb{C}_{\Delta \bar{\mathbf{m}}_x} \mathbb{E} \left[ \left\| \bar{\mathbf{m}}_x^t - \frac{1}{n} \mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^t) \right\|^2 \right] \quad (188)$$

$$+ \mathbb{C}_{\Delta \mathbf{m}_x} \mathbb{E} [\|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2] + \mathbb{C}_{\Delta \mathbf{m}_\lambda} \mathbb{E} [\|\mathbf{A}^\top \mathbf{m}_\lambda^t - \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t\|^2] \quad (189)$$

$$+ \mathbb{C}_{\bar{\mathbf{m}}_x} \mathbb{E} [\|\bar{\mathbf{m}}_x^t\|^2] \quad (190)$$

975 where

$$\mathbb{C}_{\nabla F} = -\frac{\alpha}{2} + \mathbf{e} \cdot 32(L_s^2 + L^2)\alpha^2 + \mathbf{f} \cdot 64\alpha^2 n G + \mathbf{g} \cdot 32\alpha^2 n (\rho_{\max}^2 + \bar{\rho}_{\max}^2) \quad (191)$$

$$\leq -\frac{\alpha}{4} \quad \text{when} \quad \begin{cases} \alpha \leq \min \left\{ \mathbf{e}^{-1} \cdot \frac{1}{512(L_s^2 + L^2)}, \mathbf{g}^{-1} \cdot \frac{1}{512n(\rho_{\max}^2 + \bar{\rho}_{\max}^2)} \right\} \\ \mathbf{f} \cdot 64\alpha^2 n G \leq \frac{\alpha}{24} \Leftarrow \begin{cases} \alpha \leq \mathbf{f}^{-1} \cdot \frac{1}{1024n(L_s^2 + L^2)} \\ \gamma^2 \leq \mathbf{f}^{-1} \cdot \frac{\alpha}{1024n(\rho_{\max}^2 + \bar{\rho}_{\max}^2)} \end{cases} \end{cases} \quad (192)$$

$$\mathbb{C}_\sigma = \mathbf{e} \cdot 2a_x^2 + \mathbf{f} \cdot 4a_x^2 n \quad (193)$$

$$\mathbb{C}_{\mathbf{x}} = -\mathbf{a} \cdot \frac{\gamma \rho_{\min}}{8} + \frac{\alpha L^2}{n} + 3\beta^2 \rho_{\max}^2 (\mathbf{b} \rho_{\min}^{-1} + \mathbf{c}) \quad (194)$$

$$+ \mathbf{d} \cdot \left( \frac{3\alpha^2 L^2}{\eta} + 2\beta \rho_{\max}^2 - \gamma \beta \rho_{\min}^2 + \frac{5\beta}{2} + \frac{5\alpha^2}{2\eta} \right) \quad (195)$$

$$+ \mathbf{e} \cdot 32(L_s^2 + L^2)(\alpha^2 L^2 + \gamma^2 \rho_{\max}^2)/n \quad (196)$$

$$+ \mathbf{f} \cdot \left( \frac{20a_x^2 \gamma^2 \sigma_A^2}{\alpha^2} + 64(\alpha^2 L^2 + \gamma^2 \rho_{\max}^2) G \right) \quad (197)$$

$$+ \mathbf{g} \cdot (10a_\lambda^2 \sigma_A^2 + 32(\rho_{\max}^2 + \bar{\rho}_{\max}^2)(\alpha^2 L^2 + \gamma^2 \rho_{\max}^2)) \quad (198)$$

$$\leq -\mathbf{a} \cdot \frac{\gamma \rho_{\min}}{16} \quad \text{when} \quad \left\{ \begin{array}{l} \alpha \leq \mathbf{a} \cdot \frac{\gamma \rho_{\min} n}{256 L^2} \\ \alpha^2 \leq \min \left\{ \frac{\mathbf{a}}{\mathbf{d}} \cdot \frac{\eta \gamma \rho_{\min}}{768 L^2}, \quad \frac{\mathbf{a}}{\mathbf{d}} \cdot \frac{\eta \gamma \rho_{\min}}{640} \right\} \\ \beta^2 \leq \min \left\{ \frac{\mathbf{a}}{\mathbf{b}} \cdot \frac{\gamma \rho_{\min}^2}{1536 \rho_{\max}^2}, \quad \frac{\mathbf{a}}{\mathbf{c}} \cdot \frac{\gamma \rho_{\min}}{1536 \rho_{\max}^2} \right\} \\ \alpha^3 \leq \frac{\mathbf{a}}{\mathbf{d}} \cdot \frac{\eta \gamma \rho_{\min}}{768 L^4} \\ \beta \leq \min \left\{ \frac{\mathbf{a}}{\mathbf{d}} \cdot \frac{\gamma \rho_{\min}}{512 \rho_{\max}^2}, \quad \frac{\mathbf{a}}{\mathbf{d}} \cdot \frac{\gamma \rho_{\min}}{640} \right\} \\ \alpha^2 \leq \min \left\{ \frac{\mathbf{a}}{\mathbf{e}} \cdot \frac{\gamma \rho_{\min} n}{8192 L^2 (L_s^2 + L^2)}, \quad \frac{\mathbf{a}}{\mathbf{g}} \cdot \frac{\gamma \rho_{\min}}{8192 L^2 (\bar{\rho}_{\max}^2 + \rho_{\max}^2)} \right\} \\ \gamma \leq \min \left\{ \frac{\mathbf{a}}{\mathbf{e}} \cdot \frac{\rho_{\min} n}{8192 \rho_{\max}^2 (L_s^2 + L^2)}, \quad \frac{\mathbf{a}}{\mathbf{g}} \cdot \frac{\rho_{\min}}{8192 \rho_{\max}^2 (\bar{\rho}_{\max}^2 + \rho_{\max}^2)} \right\} \\ a_x^2 \leq \frac{\mathbf{a}}{\mathbf{f}} \cdot \frac{\alpha^2 \rho_{\min}}{5120 \gamma \sigma_A^2} \\ \mathbf{f} \cdot 64 \alpha^2 L^2 G \leq \mathbf{a} \cdot \frac{\gamma \rho_{\min}}{256} \Leftrightarrow \left\{ \begin{array}{l} \alpha^2 \leq \frac{\mathbf{a}}{\mathbf{f}} \cdot \frac{\gamma \rho_{\min}}{32768 L^2 (L_s^2 + L^2)} \\ \gamma \leq \frac{\mathbf{a}}{\mathbf{f}} \cdot \frac{\rho_{\min}}{32768 L^2 (\bar{\rho}_{\max}^2 + \rho_{\max}^2)} \end{array} \right. \\ \mathbf{f} \cdot 64 \gamma^2 \rho_{\max}^2 G \leq \mathbf{a} \cdot \frac{\gamma \rho_{\min}}{256} \Leftrightarrow \left\{ \begin{array}{l} \gamma \leq \frac{\mathbf{a}}{\mathbf{f}} \cdot \frac{\rho_{\min}}{32768 \rho_{\max}^2 (L_s^2 + L^2)} \\ \gamma^3 \leq \frac{\mathbf{a}}{\mathbf{f}} \cdot \frac{\alpha^2 \rho_{\min}}{32768 \rho_{\max}^2 (\bar{\rho}_{\max}^2 + \rho_{\max}^2)} \end{array} \right. \\ a_\lambda^2 \leq \frac{\mathbf{a}}{\mathbf{g}} \cdot \frac{\gamma \rho_{\min}}{2560 \sigma_A^2} \end{array} \right. \quad (199)$$

$$\mathbb{C}_{\mathbf{v}} = -\mathbf{d} \cdot \frac{\eta}{2} + \mathbf{a} \cdot \frac{\eta^2}{(1 - \gamma \rho_{\min})(1 - \gamma \rho_{\min}/2)} + (\beta + \frac{\alpha}{\eta})(\mathbf{b} \rho_{\min}^{-1} + \mathbf{c}) \quad (200)$$

$$+ \mathbf{e} \cdot \frac{32(L_s^2 + L^2)\eta^2}{n} + \mathbf{f} \cdot \left( \frac{16a_x^2 \eta^2 \gamma^2 \sigma_A^2}{\alpha^2} + 64\eta^2 G \right) + \mathbf{g} \cdot 16\eta^2 \quad (201)$$

$$\leq -\mathbf{d} \cdot \frac{\eta}{4} \quad \text{when} \quad \left\{ \begin{array}{l} \eta \leq \frac{\mathbf{d}}{\mathbf{a}} \cdot (28(1 - \gamma \rho_{\min})(1 - \gamma \rho_{\min}/2))^{-1} \\ \beta + \frac{\alpha}{\eta} \leq \min \left\{ \frac{\mathbf{d}}{\mathbf{b}} \cdot \frac{\eta \rho_{\min}}{28}, \quad \frac{\mathbf{d}}{\mathbf{c}} \cdot \frac{\eta}{28} \right\} \\ \eta \leq \min \left\{ \frac{\mathbf{d}}{\mathbf{e}} \cdot \frac{n}{896(L_s^2 + L^2)}, \quad \frac{\mathbf{d}}{\mathbf{g}} \cdot (1/448) \right\} \\ a_x^2 \eta \gamma^2 \leq \frac{\mathbf{d}}{\mathbf{f}} \cdot \frac{\alpha^2}{448 \sigma_A^2} \\ \mathbf{f} \cdot 64 \eta^2 G \leq \mathbf{d} \cdot \frac{\eta}{28} \Leftrightarrow \left\{ \begin{array}{l} \eta \leq \frac{\mathbf{d}}{\mathbf{f}} \cdot (3584(L_s^2 + L^2))^{-1} \\ \eta \gamma^2 \leq \frac{\mathbf{d}}{\mathbf{f}} \cdot \frac{\alpha^2}{3584(\bar{\rho}_{\max}^2 + \rho_{\max}^2)} \end{array} \right. \end{array} \right. \quad (202)$$

$$\mathbf{C}_{xv} = \left( -\mathbf{a} \cdot \frac{2\eta(1 - \gamma \rho_{\min}/4)}{1 - \gamma \rho_{\min}} + \mathbf{b} \cdot 2\beta \right) \mathbf{K} \quad (203)$$

$$+ \left( \mathbf{a} \cdot \frac{2\eta \gamma (1 - \gamma \rho_{\min}/4)}{1 - \gamma \rho_{\min}} + \mathbf{c} \cdot 2\beta - \mathbf{d} \cdot (\gamma + \eta \beta) \right) \mathbf{A}^\top \mathbf{R} \mathbf{A} \quad (204)$$

$$= \mathbf{0} \quad \text{when} \quad \left\{ \begin{array}{l} \mathbf{b} = \frac{\eta(1 - \gamma \rho_{\min}/4)}{\beta(1 - \gamma \rho_{\min})} \mathbf{a}, \\ \mathbf{c} = \frac{\mathbf{d}(\gamma + \eta \beta) - \mathbf{a} \cdot 2\eta \gamma (1 - \gamma \rho_{\min}/4)(1 - \gamma \rho_{\min})^{-1}}{2\beta} \end{array} \right. \quad (205)$$

$$\mathbb{C}_{\Delta \bar{\mathbf{m}}_x} = -\mathbf{e} \cdot a_x + \alpha \quad (206)$$

$$\leq -\mathbf{e} \cdot \frac{a_x}{2} \quad \text{when} \quad \alpha \leq \mathbf{e} \cdot \frac{a_x}{2} \quad (207)$$

$$\mathbb{C}_{\Delta \mathbf{m}_x} = -\mathbf{f} \cdot \frac{a_x}{4} + \mathbf{a} \cdot \frac{2\alpha^2}{\gamma \rho_{\min}} + \mathbf{d} \cdot \left( \left( \frac{3}{\eta} + \frac{1}{2} \right) \alpha^2 + 2\alpha^2 \beta + \frac{2\alpha^4}{\eta} \right) \quad (208)$$

$$+ \mathbf{e} \cdot \frac{8(L_s^2 + L^2)\alpha^2}{n} + \mathbf{g} \cdot 8\alpha^2(1 + \rho_{\max}^2 + \bar{\rho}_{\max}^2) \quad (209)$$



$$\leq -\mathbf{f} \cdot \frac{a_x}{8} \quad \text{when} \quad \begin{cases} \alpha^2 \leq \min \left\{ \frac{\mathbf{f}}{\mathbf{a}} \cdot \frac{a_x \gamma \rho_{\min}}{64}, \quad \frac{\mathbf{f}}{\mathbf{d}} \cdot \frac{a_x \eta}{384}, \quad \frac{\mathbf{f}}{\mathbf{d}} \cdot \frac{a_x}{64} \right\} \\ \alpha^2 \beta \leq \frac{\mathbf{f}}{\mathbf{d}} \cdot \frac{a_x}{256} \\ \alpha^4 \leq \frac{\mathbf{f}}{\mathbf{d}} \cdot \frac{a_x \eta}{256} \\ \alpha^2 \leq \min \left\{ \frac{\mathbf{f}}{\mathbf{e}} \cdot \frac{a_x n}{256(L_s^2 + L^2)}, \quad \frac{\mathbf{f}}{\mathbf{g}} \cdot \frac{a_x}{256(1 + \rho_{\max}^2 + \bar{\rho}_{\max}^2)} \right\} \end{cases} \quad (210)$$

$$\mathbb{C}_{\Delta \mathbf{m}_\lambda} = -\mathbf{g} \cdot a_\lambda + 4\beta(\mathbf{b}\rho_{\min}^{-1} + \mathbf{c}) + \mathbf{d} \cdot \frac{\beta}{2} \quad (211)$$

$$\leq -\mathbf{g} \cdot \frac{a_\lambda}{2} \quad \text{when} \quad \beta \leq \min \left\{ \frac{\mathbf{g}}{\mathbf{b}} \cdot \frac{a_\lambda \rho_{\min}}{24}, \quad \frac{\mathbf{g}}{\mathbf{c}} \cdot \frac{a_\lambda}{24}, \quad \frac{\mathbf{g}}{\mathbf{d}} \cdot \frac{a_\lambda}{3} \right\} \quad (212)$$

$$\mathbb{C}_{\bar{\mathbf{m}}_x} = -\frac{\alpha}{4} + \frac{4n\alpha^3 L^2}{\eta}(\mathbf{b}\rho_{\min}^{-1} + \mathbf{c}) + \mathbf{d} \cdot \frac{\alpha^2 L^2 n}{2\eta} \quad (213)$$

$$\leq -\frac{\alpha}{8} \quad \text{when} \quad \begin{cases} \alpha^2 \leq \min \left\{ \mathbf{b}^{-1} \cdot \frac{\eta \rho_{\min}}{96nL^2}, \quad \mathbf{c}^{-1} \cdot \frac{\eta}{96nL^2} \right\} \\ \alpha \leq \mathbf{d}^{-1} \frac{\eta}{12nL^2} \end{cases} \quad (214)$$

976 Notice that the hyperparameter choices in (111) - (117) will satisfy all of the above conditions,  
977 therefore completes the proof to (120).  $\square$

## 978 D.9 Auxiliary Lemma

979 **Lemma D.9.** Under Assumption 3.1, 3.2,

$$\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 \leq 2\alpha^2 \|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2 + 8\alpha^2 n \|\nabla F(\bar{\mathbf{x}}^t)\|^2 \quad (215)$$

$$+ 8\eta^2 \|\mathbf{v}^t\|_{\mathbf{K}}^2 + (8\alpha^2 L^2 + 8\gamma^2 \rho_{\max}^2) \|\mathbf{x}^t\|_{\mathbf{K}}^2 \quad (216)$$

980 *Proof of Lemma D.9.*

$$\|\mathbf{x}^{t+1} - \mathbf{x}^t\|^2 = \alpha^2 \|\mathbf{m}_x^t\|^2 \leq 2\alpha^2 \|\mathbf{m}_x^t - \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2 + 2\alpha^2 \|\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2 \quad (217)$$

981 Upon noting that  $\|\mathbf{v}^t\|_{\mathbf{K}}^2 = \|\mathbf{A}^\top \boldsymbol{\lambda}^t - (\alpha/\eta)(n^{-1} \mathbf{1}\mathbf{1}^\top - \mathbf{I}_n) \otimes \mathbf{I}_d \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)\|^2$ , we expand

$$\|\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2 = \left\| \nabla \mathbf{f}(\mathbf{x}^t) + \frac{\eta}{\alpha} \mathbf{A}^\top \boldsymbol{\lambda}^t + \frac{\gamma}{\alpha} \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t \right\|^2 \quad (218)$$

$$= \left\| \mathbf{1}_{\otimes} \nabla F(\bar{\mathbf{x}}^t) + \frac{\eta}{\alpha} \mathbf{A}^\top \boldsymbol{\lambda}^t - (\mathbf{1}_{\otimes} \nabla F(\bar{\mathbf{x}}^t) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)) - (\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t) - \nabla \mathbf{f}(\mathbf{x}^t)) + \frac{\gamma}{\alpha} \mathbf{A}^\top \mathbf{R} \mathbf{A} \mathbf{x}^t \right\|^2$$

$$\leq 4\|\mathbf{1}_{\otimes} \nabla F(\bar{\mathbf{x}}^t)\|^2 + \frac{4\eta^2}{\alpha^2} \left\| \mathbf{A}^\top \boldsymbol{\lambda}^t - \frac{\alpha}{\eta} (\mathbf{1}_{\otimes} \nabla F(\bar{\mathbf{x}}^t) - \nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t)) \right\|^2 + 4\|\nabla \mathbf{f}(\mathbf{1}_{\otimes} \bar{\mathbf{x}}^t) - \nabla \mathbf{f}(\mathbf{x}^t)\|^2$$

$$+ \frac{4\gamma^2 \rho_{\max}^2}{\alpha^2} \|\mathbf{x}^t\|_{\mathbf{K}}^2 \quad (219)$$

$$\leq 4n\|\nabla F(\bar{\mathbf{x}}^t)\|^2 + \frac{4\eta^2}{\alpha^2} \|\mathbf{v}^t\|_{\mathbf{K}}^2 + (4L^2 + \frac{4\gamma^2 \rho_{\max}^2}{\alpha^2}) \|\mathbf{x}^t\|_{\mathbf{K}}^2 \quad (220)$$

982 Combining the above inequalities completes the proof.  $\square$

## 983 E Ablation Study

984 In this section, we investigate how FSPDA behaves under different problem configurations, for  
985 instance, the different levels of data heterogeneity, random graph sparsity, graph topology, gradient  
986 noise and dual momentum. Unless specified explicitly, we assume the experiment adopts  $\mathcal{G}$  as the  
987 fully connected (complete) graph topology.

### 988 E.1 Data Heterogeneity

989 To study the effect of heterogeneity of data distribution across agents, we experiment with two types  
990 of data splitting for MNIST: (i) the dataset is split into  $n = 10$  disjoint sets according to the class  
991 labels, or (ii) the dataset is split into  $n$  evenly distributed disjoint sets by shuffling. Setup (i) creates a

large discrepancy across local objective functions, i.e., a larger data heterogeneity. Figure 4 compares the performance of FSPDA and the benchmark algorithms under the above setup which demonstrates the robustness of FSPDA under heterogeneous data distribution. For instance, the performance of CHOCO-SGD and DSGD hugely degrades in the heterogeneous setup (i), while that of FSPDA-SA is only affected by a small margin and FSPDA-STORM is able to converge to the same low error in both setup (i) and (ii). We list the hyperparameters used in Figure 4 by Table 4 in Appendix F.

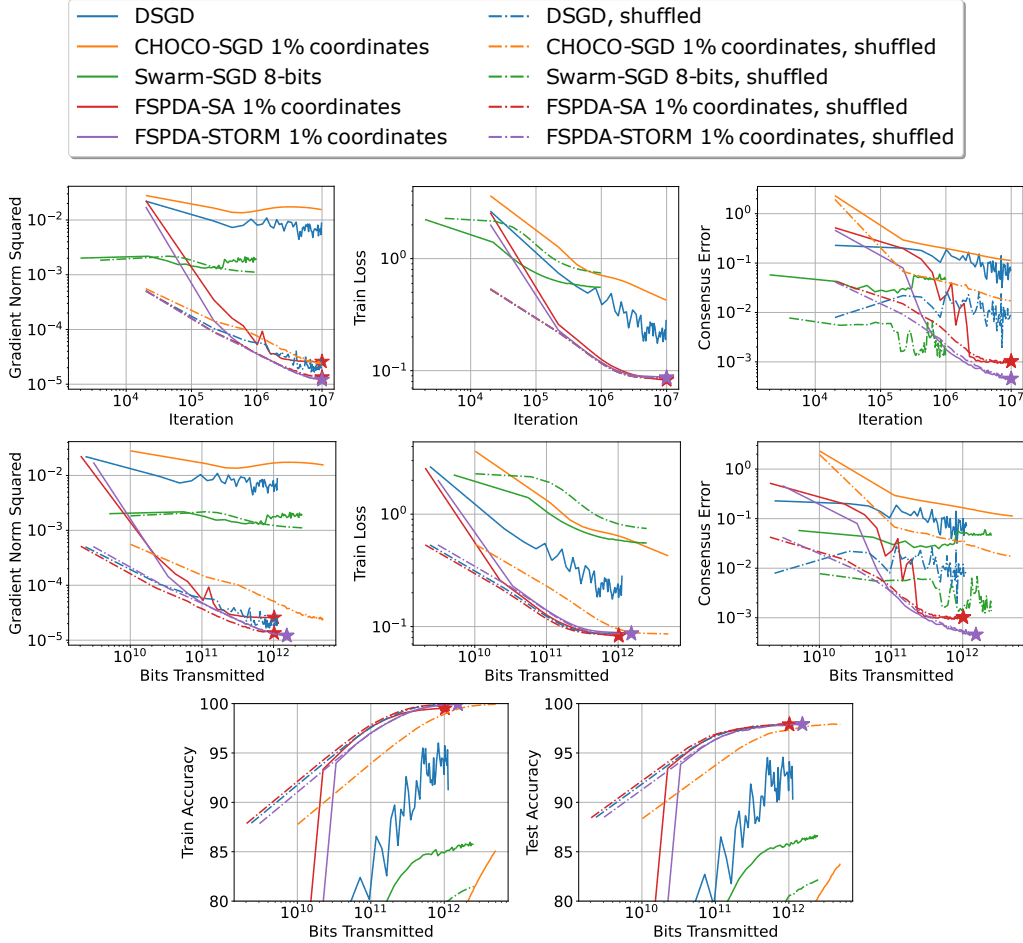


Figure 4: Feed-forward neural network classification training on MNIST with two levels of data heterogeneity.

## E.2 Random Graph Sparsity

Next, we study the effects of random graph sparsity by the experiments shown in Figure 5. As the random graph sparsity decreases, the random graph variance  $\sigma_A^2$  in Assumption 3.4 decreases which will significantly improve the consensus error. We observe from the figure that despite the different levels of consensus error, 4 out of 5 configurations eventually converge to the same stationarity. In the extreme case of 0.01% sparsity with one-edge random graph, we see the dominance of sparsity error which leads to a longer transient time as expected from Theorem 3.5. We conclude that for a fixed number of iteration, a certain amount of communication sparsity can be tolerated in FSPDA without degradation in optimization error. For Figure 5, we tuned FSPDA-SA to use the step sizes  $\alpha = 10^{-4}, \eta = 10^{-6}, \gamma = 0.5, \beta = 1$  for all configurations.

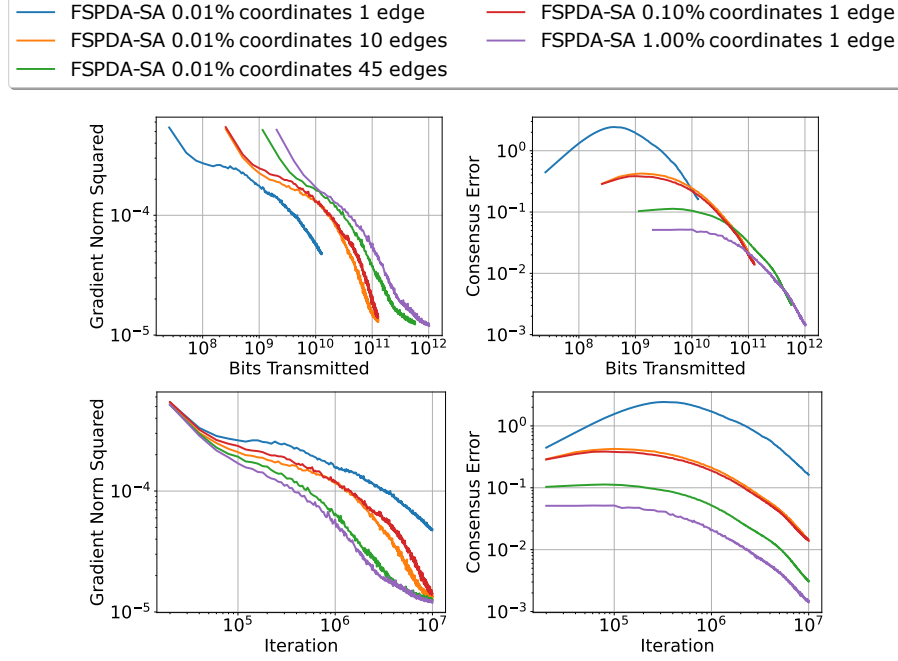


Figure 5: Feed-forward neural network classification training on shuffled MNIST. Random graph of  $k$  edges in expectation ( $k \in \{1, 10, 45\}$ ) is drawn from a complete topology per iteration.

### 1008 E.3 Graph Topology

1009 We then study the effects of network topology  $\mathcal{G}$  in FSPDA-SA by drawing one-edge random graphs  
 1010 from a complete graph, an ER graph with probability  $p = 0.5$  and a ring graph in Figure 6. Note  
 1011 that the communication cost per iteration is the same across different topologies due to the use of  
 1012 one-edge random graph. The result indicates the transient effect of topology where it only slow down  
 1013 the convergence of consensus error while converging to the same level of stationarity. For Figure 6,  
 1014 we tuned FSPDA-SA to use the step sizes  $\alpha = 10^{-4}$ ,  $\eta = 10^{-6}$ ,  $\gamma = 0.5$ ,  $\beta = 1$  for all configurations.

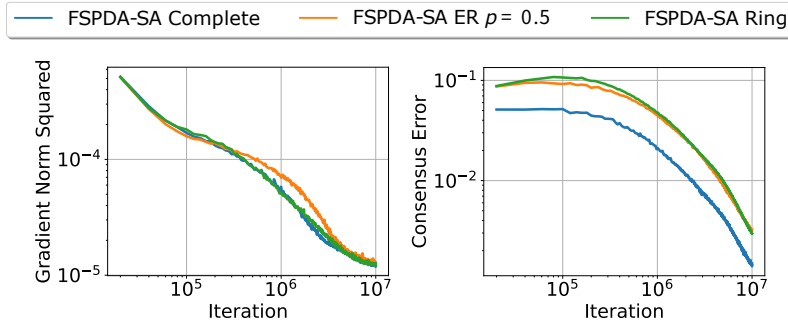


Figure 6: Feed-forward neural network classification training on shuffled MNIST with different graph topology  $\mathcal{G}$ . Only one edge is activated per iteration, exchanging 1% coordinates of model parameters.

### 1015 E.4 Deterministic Gradient

1016 For the case when the gradient estimate is exact, i.e.,  $\bar{\sigma}^2 = 0$ , we compare the performance of  
 1017 FSPDA-SA against deterministic gradient algorithm DIGing [Nedic et al., 2017] in Figure 7. Despite  
 1018 FSPDA-SA only performs model parameter gossip while DIGing performs an extra step of gradient

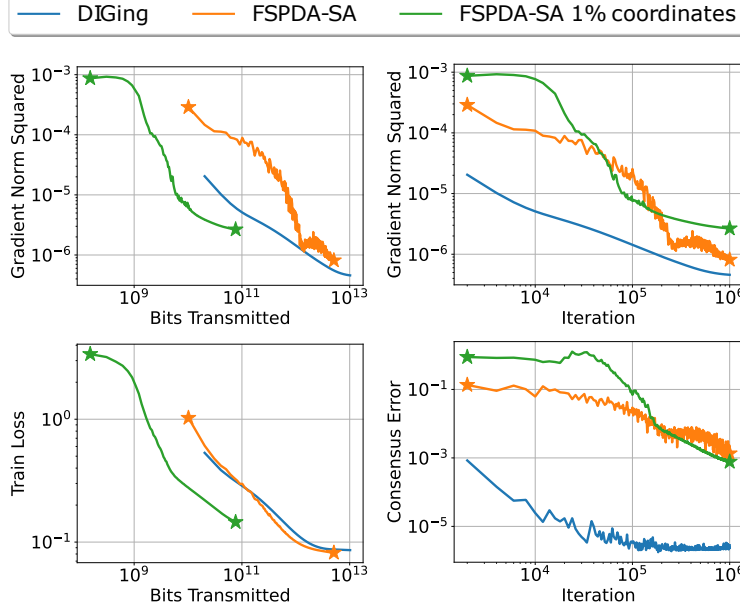


Figure 7: Feed-forward neural network classification training on class separated MNIST using exact local gradient, i.e.,  $\sigma_i = 0 \forall i$ . One-edge random graph is drawn from a complete topology per iteration.

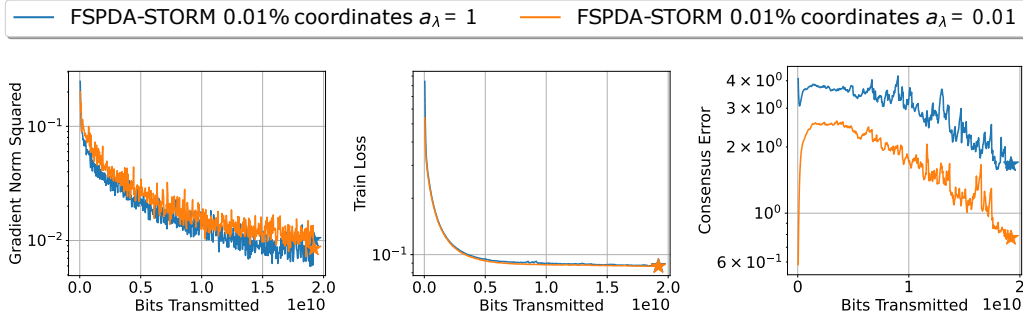


Figure 8: Feed-forward neural network classification training on shuffled MNIST using  $10^7$  iterations. One-edge random graph is drawn from a complete topology per iteration.

1019 tracker gossip, FSPDA-SA shows comparative performance and reduced the communication cost by  
1020 half.

1021 Also, in the absence of stochastic gradient noise, notice that the effect of parameter sparsification  
1022 immediately transfer to a slower optimization convergence. This is in line with our theorem by  
1023 observing the convergence rate in Theorem 3.5, where  $\sigma_A^4$  remains dominant in the convergence  
1024 bound. We list the hyperparameters used in Figure 7 by Table 5 in Appendix F.

## 1025 E.5 Dual Momentum

1026 To investigate the benefits of dual momentum in FSPDA-STORM, we construct a case where both  
1027 the primal and dual stochastic gradients carry large variance error. In Figure 8, the local objective  
1028 function gradient is estimated by batch size 1, and one-edge random graphs with 0.01% coordinate  
1029 sparsification is adopted for communication. We can observe that applying dual momentum ( $a_\lambda =$   
1030 0.01) outperforms not applying dual momentum ( $a_\lambda = 1$ ) in terms of consensus error convergence.  
1031 We list the hyperparameters used in Figure 8 by Table 6 in Appendix F.

1032 **F Experiment Hyperparameters**

FSPDA	$\alpha$	$\eta$	$\gamma$	$\beta$	$a_x$	$a_\lambda$
-SA (10% sparse coor.)	$10^{-4}$	$10^{-5}$	0.5	1	-	-
-STORM (6.7% sparse coor.)	$10^{-3}$	$10^{-2}$	0.5	0.1	$10^{-2}$	$10^{-2}$
K-GT	<b>Opt. S.S.</b>	<b>Local Steps</b>	$\eta_s$	-	-	-
	$10^{-4}$	150	1	-	-	-
LED	<b>Opt. S.S.</b>	<b>Local Steps</b>	-	-	-	-
	$10^{-4}$	75	-	-	-	-
Decen-Scaffnew	<b>Opt. S.S.</b>	$\tau$	$p$	-	-	-
	$10^{-4}$	130	0.013	-	-	-
DSGD	<b>Opt. S.S.</b>	<b>Edge Prob.</b>	-	-	-	-
	$10^{-4}$	0.013	-	-	-	-
Swarm-SGD	<b>Opt. S.S.</b>	<b>Quant. Side Length</b>	-	-	-	-
	$5 \times 10^{-5}$	$10^{-4}$	-	-	-	-
CHOCO-SGD	<b>Opt. S.S.</b>	<b>Consensus S.S.</b>	<b>Active Prob.</b>	-	-	-
	$10^{-4}$	$10^{-3}$	0.03	-	-	-

Table 2: Hyperparameter values used in Figure 1.

FSPDA-SA	$\max_t \alpha_t$	$\max_t \eta_t$	$\gamma$	$\beta$
10% coordinates	0.1	$5 \times 10^{-9}$	0.5	1
1% coordinates	0.1	$10^{-9}$	0.5	1
0.1% coordinates	0.05	$5 \times 10^{-10}$	0.5	1
CHOCO-SGD	<b>Max. Opt. S.S.</b>	<b>Consensus S.S.</b>	<b>Active Prob.</b>	-
10% coordinates	0.1	0.05	0.1	-
1% coordinates	0.1	0.005	0.1	-
Swarm-SGD	<b>Max. Opt. S.S.</b>	<b>Quant. Side Length</b>	-	-
8-bits quantization	$10^{-3}$	$3 \times 10^{-5}$	-	-

Table 3: Hyperparameter values used in Figure 2.

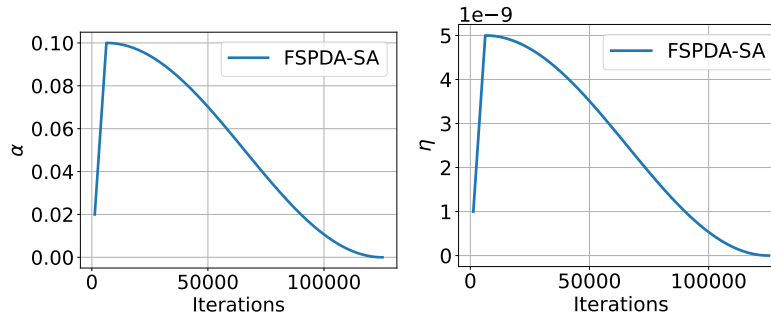


Figure 9: Illustration of step size cosine scheduling used in Figure 2 for FSPDA-SA with 10% sparse coordinates.

DSGD	<b>Opt. S.S.</b>	<b>Edge Prob.</b>	-	-	-	-
(i) hetero.	$10^{-4}$	$5 \times 10^{-4}$	-	-	-	-
(ii) homo.	$10^{-4}$	$5 \times 10^{-4}$	-	-	-	-
CHOCO-SGD	<b>Opt. S.S.</b>	<b>Consensus S.S.</b>	<b>Active Prob.</b>	-	-	-
(i) hetero.	$10^{-4}$	$10^{-3}$	0.1	-	-	-
(ii) homo.	$10^{-4}$	$10^{-3}$	0.1	-	-	-
Swarm-SGD	<b>Opt. S.S.</b>	<b>Quant. Side Length</b>	-	-	-	-
(i) hetero.	$5 \times 10^{-5}$	$10^{-4}$	-	-	-	-
(ii) homo.	$5 \times 10^{-5}$	$10^{-4}$	-	-	-	-
FSPDA-SA	$\alpha$	$\eta$	$\gamma$	$\beta$	-	-
(i) hetero.	$10^{-4}$	$10^{-4}$	0.5	1	-	-
(ii) homo.	$10^{-4}$	$10^{-5}$	0.5	1	-	-
FSPDA-STORM	$\alpha$	$\eta$	$\gamma$	$\beta$	$a_x$	$a_\lambda$
(i) hetero.	$10^{-3}$	$10^{-3}$	0.5	0.1	0.1	0.1
(ii) homo.	$10^{-3}$	$10^{-4}$	0.5	0.1	0.1	0.1

Table 4: Hyperparameter values used in Figure 4.

FSPDA-SA	$\alpha$	$\eta$	$\gamma$	$\beta$
no sparse	$10^{-3}$	$5 \times 10^{-6}$	0.5	1
1% coordinates	$10^{-4}$	$5 \times 10^{-4}$	0.5	1
DIGing	<b>Opt. S.S.</b>	-	-	-
no sparse	$10^{-3}$	-	-	-

Table 5: Hyperparameter values used in Figure 7.

	$\alpha$	$\eta$	$\gamma$	$\beta$	$a_x$	$a_\lambda$
FSPDA-STORM	$10^{-3}$	$5 \times 10^{-6}$	0.5	1	$10^{-3}$	1
FSPDA-STORM	$10^{-3}$	$5 \times 10^{-6}$	0.5	1	$10^{-3}$	$10^{-2}$

Table 6: Hyperparameter values used in Figure 8.

	Time	Peak Memory	Machine
Figure 1	27 hours	339 MB	Intel Xeon Gold 6148 CPU
Figure 2	38 hours	(GPU) 112 GB (CPU) 6650 MB	8× NVIDIA V100 GPU
Figure 4	197 hours	372 MB	Intel Xeon Gold 6148 CPU
Figure 5	201 hours	360 MB	Intel Xeon Gold 6148 CPU
Figure 6	72 hours	1146 MB	Intel Xeon Gold 6148 CPU
Figure 7	354 hours	1066 MB	Intel Xeon Gold 6148 CPU
Figure 8	46 hours	350 MB	Intel Xeon Gold 6148 CPU

Table 7: Statistics of experiment compute time (per algorithm run) and compute instance.