Stochastic Approximation Approach for Decentralized Optimization on Time Varying Random Networks

Chung-Yiu Yau[1] **Haoming Liu**[2]

Hoi-To Wai[3]

[1]University of Minnesota [2]Beijing University of Technology [3]The Chinese University of Hong Kong
[1]cyau@umn.edu [2]liuhaoming@bjut.edu.cn [3]htwai@cuhk.edu.hk

Abstract

This paper initiates the study of a stochastic approximation approach with the Fully Stochastic Primal Dual Algorithm (FSPDA) framework for decentralized optimization on random and time varying topologies. Our framework relies on a novel observation that randomnesses in time varying topology can be incorporated into a stochastic equality constrained optimization formulation. We derived two new algorithms supporting sparsified communication on time varying topologies — FSPDA-SA allows agents to execute multiple local gradient steps to accelerate convergence, and FSPDA-STORM further incorporates variance reduction to improve sample complexity. For problems with smooth (possibly non-convex) objective function, within T iterations, FSPDA-SA (resp. FSPDA-STORM) finds an $\mathcal{O}(1/\sqrt{T})$-stationary (resp. $\mathcal{O}(1/T^{2/3})$) solution. Our methods are the first to establish finite-time stationarity guarantees in smooth nonconvex decentralized stochastic optimization over i.i.d. random subgraphs (such as disconnected subgraphs at every iteration) of a fixed connected base graph without bounded local function heterogeneity.

1 Introduction

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

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

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

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

where $\mathbb{P}_i$ represents the i-th data distribution and we assume $f_i \neq f_j$, i.e., the heterogeneous data setting. Problem (1) is relevant to the distributed learning problem in the decentralized case where a central server is absent. Prior works [Nedic and Ozdaglar, 2009, Lian et al., 2017, Nedic et al., 2017, Qu and Li, 2017] demonstrated that *decentralized* algorithms tackle (1) efficiently through message exchanges among the neighbors and local stochastic gradient updates.

An important direction towards practical decentralized optimization is to consider a *time varying graph topology* setting where the *active edge set* in $\mathcal{G}$ changes over time. This is a generic setting

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Prior Works	SG	TV	w/o BH	Post-transient Rate
Prox-GPDA [Hong et al., 2017]	✗	✗	✓	Asympt.
NEXT [Lorenzo and Scutari, 2016]	✗	✓	✓	Asympt.
DSGD [Koloskova et al., 2020]	✓	✓	✗	$\mathcal{O}(\sqrt{L}\sigma/\sqrt{nT})$
Swarm-SGD [Nadiradze et al., 2021]	✓	✓	✗	$\mathcal{O}(\sigma^2/\sqrt{T})$
CHOCO-SGD [Koloskova et al., 2019a]	✓	✗$^\sharp$	✗	$\mathcal{O}(\sqrt{L}\sigma/\sqrt{nT})$
ProxSkip [Mishchenko et al., 2022, Guo et al., 2023]	✓	✗†	✓	$\mathcal{O}(\sqrt{L}\sigma/\sqrt{nT})$
Local-GT [Liu et al., 2024]	✓	✗†	✓	$\mathcal{O}(\sqrt{L}\sigma/\sqrt{nT})$
LED [Alghunaim, 2024]	✓	✗†	✓	$\mathcal{O}(\sqrt{L}\sigma/\sqrt{nT})$
Push-ASGD [Chen et al., 2024]	✓	✗‡	✓	$\mathcal{O}(L_s^{2/3}\sigma^{2/3}/T^{2/3})$
FSPDA-SA (**This Work**)	✓	✓	✓	$\mathcal{O}(L\sigma/\sqrt{nT})$
FSPDA-STORM (**This Work**)	✓	✓	✓	$\mathcal{O}(L_s^{4/3}\sigma^{2/3}/T^{2/3})$

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 'Post-transient Rate' is the dominant term of expected squared gradient norm $\mathbb{E}[\|\nabla F(\bar{\mathbf{x}})\|^2]$ after $T \gg 1$ iterations. Note that σ^2 is the variance of stochastic gradient. $^\sharp$CHOCO-SGD incorporates broadcast gossip as a special case of compression. †ProxSkip, Local-GT, LED consider local updates with periodic communication. ‡Push-ASGD considers connected instantaneous networks.

covering cases when the communication links are unreliable, or the agents choose not to communicate in a certain round (a.k.a. local updates) [Koloskova et al., 2019a, Nadiradze et al., 2021]. 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 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. These works focused on DSGD that exhibits slow convergence (in transient stage) and its performance is limited by the degree of data heterogeneity, i.e., $\varsigma := \sup_{\mathbf{x}\in\mathbb{R}^d} \|\nabla f_i(\mathbf{x}) - \nabla f_j(\mathbf{x})\|$. The prior restrictions can be relaxed using sophisticated algorithms such as gradient tracking (GT) [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, Zhou and Pu, 2026] for similar results and [Nguyen et al., 2025] for a recent development on finite-time consensus graphs. Note that such algorithms require extra synchronization overhead. In addition, [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 $F(\mathbf{x})$; [Chen et al., 2024] considered time varying directed graphs that are strongly connected at all realizations; [Liang et al., 2025] utilized a double-loop design for time varying directed graphs. We provide a non-exhaustive list summarizing the existing works in Table 1.

The above highlights a gap in existing literature —

Is there any algorithm that achieves fast convergence on time varying (random) topology under heterogeneous data distribution?

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 sophisticated 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 time varying topology. Applying stochastic approximation (SA) 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 treats the random time varying topology as a part of randomness.
- We show that after T iterations, FSPDA-SA (resp. FSPDA-STORM) finds in expectation a solution with squared gradient norm as $\mathcal{O}(1/\sqrt{T})$ (resp. $\mathcal{O}(1/T^{2/3})$). To our best knowledge, FSPDA-STORM achieves the first near-optimal sample complexity for time-varying graph decentralized optimization. Interestingly, using momentum in dual updates benefits the consensus error convergence.
- We demonstrate that 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.

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 $\langle \mathbf{a} \mid \mathbf{b}\rangle_{\mathbf{W}} := \mathbf{a}^\top \mathbf{W}\mathbf{b}$. Similarly, the $\mathbf{W}$-weighted (semi) norm is denoted by $\|\mathbf{a}\|^2_{\mathbf{W}} := \langle \mathbf{a} \mid \mathbf{a}\rangle_{\mathbf{W}}$. The subscript notation is omitted for $\mathbf{I}$-weighted inner products. For square matrix $\mathbf{X}$, $(\mathbf{X})^\dagger$ denotes its Moore-Penrose 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 $\widetilde{\mathbf{A}} \in \{-1, 0, 1\}^{|\mathcal{E}|\times n}$ be an incidence matrix of $\mathcal{G}$. Let $\mathbf{A} = \widetilde{\mathbf{A}} \otimes \mathbf{I}_d \in \{-1, 0, 1\}^{|\mathcal{E}|d\times nd}$, 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 ξ_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, $\mathbf{I}(\xi_a)$ may select a random subset of d coordinates. This enables us to simultaneously achieve random sparsification for communication compression.

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

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

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

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) := \mathbb{E}_\xi[\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}; \xi)], \ \ \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}; \xi) := \textstyle\sum_{i=1}^n f_i(\mathbf{x}_i;\xi_i) + \tilde{\eta}\left\langle \boldsymbol{\lambda} \,\middle|\, \mathbf{R}^{-1}\mathbf{A}(\xi_a)\mathbf{x}\right\rangle + \frac{\tilde{\gamma}}{2}\|\mathbf{R}^{-1}\mathbf{A}(\xi_a)\mathbf{x}\|^2_{\mathbf{R}^2}, \tag{4}$$

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 ξ whenever the notation is clear from the context.

FSPDA applies stochastic approximation (SA) to seek a saddle point of (4). Recognizing $\mathbf{A}(\xi)^\top\mathbf{A}(\xi) = \mathbf{A}^\top\mathbf{A}(\xi)$, we consider the stochastic gradients[1]:

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

where $\nabla\mathbf{f}(\mathbf{x};\xi) = [\nabla f_1(\mathbf{x}_1;\xi_1); \ldots; \nabla f_n(\mathbf{x}_n;\xi_n)] \in \mathbb{R}^{nd}$. Notice that we have taken a deterministic $\mathbf{A}$ for the term related to $\boldsymbol{\lambda}$ in $\nabla_{\mathbf{x}}\mathcal{L}$, allowing a memory-efficient implementation for the ith agent to store the ith block of $[\mathbf{A}^\top\boldsymbol{\lambda}]_i \in \mathbb{R}^d$ only, as opposed to a stochastic dual coupling of $\mathbf{A}(\xi)^\top\boldsymbol{\lambda}$ which requires the ith agent to store a set of duals in the neighborhood $\mathcal{N}_i$ ($\{[\boldsymbol{\lambda}]_k \in \mathbb{R}^d : k \in \mathcal{N}_i\}$, a total

[1] Some readers might be concerned that the expectation of (5) does not recover the gradients of the augmented Lagrangian in (4). We note that the dual gradient in (5) can be seen as unbiased stochastic gradient to the augmented Lagrangian of an equivalent problem with the stochastic equality constraint $\mathbb{E}_{\xi_a\sim\mathbb{P}_a}[\mathbf{A}(\xi)]\mathbf{x} = \mathbf{0}$. Meanwhile, our analysis does not depend on the unbiasedness of the stochastic approximation of random incidence matrix to the posed Lagrangian formulation, as the Lagrangian formulation is only used to construct the algorithm while our proof of convergence relies on the construction of a new Lyapunov function system.

of $\mathbb{R}^{|\mathcal{N}_i|d}$) for decoding the per-iteration random dual coupling. Now observe the ith 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 ith agent as:

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

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 ξ. Eq. (6) *only* relies on $\mathbf{x}_j$ from neighbor j that is connected on the time varying topology $\mathcal{G}(\xi)$. 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 step sizes,

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

Substituting the variables with $\widehat{\boldsymbol{\lambda}} := \mathbf{A}^\top\boldsymbol{\lambda}$ yields:

FSPDA-SA: with $\widehat{\boldsymbol{\lambda}}^0 = \mathbf{0}$ and $\mathbf{x}^0 \in \mathbb{R}^{nd}$, for any $t \geq 0$ and any $i \in [n]$,

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

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

FSPDA-STORM **Algorithm.** The second variant of FSPDA uses the recursive momentum variance reduction technique [Cutkosky and Orabona, 2019] to accelerate convergence. We 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,

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

The aim of $\mathbf{m}_x^{t+1}$ is to estimate $\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1})$. Instead of $\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1}, \boldsymbol{\lambda}^{t+1}; \xi^{t+1})$, we now include an extra zero-mean term $\mathbf{m}_x^t - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t; \xi^{t+1})$. The latter is a control variate that is computed recursively. Particularly, it reduces variance with a momentum parameter a_x when the stochastic gradient map satisfies a mean-square Lipschitz condition. We summarize the algorithm as:

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

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t - \alpha\mathbf{m}_{x,i}^t, \;\; \widehat{\boldsymbol{\lambda}}_i^{t+1} = \widehat{\boldsymbol{\lambda}}_i^t + \beta\mathbf{m}_{\lambda,i}^t, \tag{10a}$$

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

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

Note that in theory (see later in Sec. 3), $\mathbf{x}^0, \widehat{\boldsymbol{\lambda}}^0, \mathbf{m}_x^0, \mathbf{m}_\lambda^0$ shall be initialized as $\mathbf{x}_i^0 = \bar{\mathbf{x}}^0$, $\widehat{\boldsymbol{\lambda}}_i^0 = (\alpha/\eta)(\nabla F(\bar{\mathbf{x}}^0) - \nabla f_i(\bar{\mathbf{x}}^0))$, $\mathbf{m}_{x,i}^0 = \nabla F(\bar{\mathbf{x}}^0)$, $\mathbf{m}_{\lambda,i}^0 = \mathbf{0}$ according to the initialization conditions above (24). We remark that a simple choice $\widehat{\boldsymbol{\lambda}}^0 = \mathbf{m}_{x,i}^0 = \mathbf{m}_{\lambda,i}^0 = \mathbf{0}$ works well in practice, suggesting that the theoretical initial error quantities are insignificant compared to the dominant optimization error and stochastic error.

Both FSPDA-SA and FSPDA-STORM can be implemented on random time varying topology, and support random sparisification for communication compression. The key is to observe that in (8),

(10), the only information required for agent i is to obtain $\sum_{j\in\mathcal{N}_i(\xi_a^t)}\mathbf{C}_{ij}(\xi_a^t)(\mathbf{x}_i^t-\mathbf{x}_j^t)$, and in addition $\sum_{j\in\mathcal{N}_i(\xi_a^t)}\mathbf{C}_{ij}(\xi_a^t)(\mathbf{x}_i^{t-1}-\mathbf{x}_j^{t-1})$ for FSPDA-STORM, at iteration t. Note that this step is analogous to the popular random gossip algorithm for average consensus [Boyd et al., 2006].

Local & Asynchronous Updates. The *local update* scheme where each agent i is allowed to update its own local variables $\mathbf{x}_i, \boldsymbol{\lambda}_i$ 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 FSPDA where the time varying topology $\mathcal{E}^{(t)}$ alternates between $\mathcal{E}^{(t)}=\mathcal{E}$ and $\mathcal{E}^{(t)}=\emptyset$.

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

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

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

This reduces the FSPDA-SA recursion into a primal-only sequence. 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 $\mathbf{W}=\mathbf{I}-\gamma\mathrm{Diag}(\tilde{\mathbf{W}}\mathbf{1})+\gamma\tilde{\mathbf{W}}$ where $\tilde{\mathbf{W}}$ is the 0-1 adjacency matrix of $\mathcal{G}$. Here, with an appropriate choice of γ, $\mathbf{W}$ will be doubly stochastic and satisfies the convergence requirement in [Shi et al., 2015]. For stochastic optimization on random networks using FSPDA-SA, (11) suggests each agent to keep the current and previous iterates received from neighbors in the corresponding time varying topology.

Similarly, by expanding (10a)–(10c), we observe that the FSPDA-STORM sequence is equivalent to the following recursion when $a_x=a_\lambda$, $\mathbf{x}_i^0=\bar{\mathbf{x}}^0$, $\mathbf{m}_x^0=\nabla\mathbf{f}(\mathbf{x}^0)$ and $\mathbf{m}_\lambda^0=\boldsymbol{\lambda}^0=\mathbf{0}$:

$$\begin{aligned}\mathbf{x}^{t+2}\approx\;&2\mathbf{x}^{t+1}-\mathbf{x}^t-\alpha\left(\nabla\mathbf{f}(\mathbf{x}^{t+1};\xi^{t+1})-\nabla\mathbf{f}(\mathbf{x}^t;\xi^t)\right)\\&-\gamma\mathbf{A}^\top\mathbf{A}(\xi^{t+1})\mathbf{x}^{t+1}+(\gamma-\eta\beta)\mathbf{A}^\top\mathbf{A}(\xi^t)\mathbf{x}^t\\&-\textstyle\sum_{k=1}^{t-1}(1-a_x)^{t-k}\Big[\alpha\Big(\nabla\mathbf{f}(\mathbf{x}^{k+1};\xi^{k+1})-\nabla\mathbf{f}(\mathbf{x}^{k+1};\xi^{k+2})+\nabla\mathbf{f}(\mathbf{x}^k;\xi^{k+1})-\nabla\mathbf{f}(\mathbf{x}^k;\xi^k)\Big)\\&+\gamma\left(\mathbf{A}^\top\mathbf{A}(\xi^{k+1})\mathbf{x}^{k+1}-\mathbf{A}^\top\mathbf{A}(\xi^{k+2})\mathbf{x}^{k+1}\right)+(\gamma-\eta\beta)\left(\mathbf{A}^\top\mathbf{A}(\xi^{k+1})\mathbf{x}^k-\mathbf{A}^\top\mathbf{A}(\xi^k)\mathbf{x}^k\right)\Big]\end{aligned}\tag{12}$$

where we have omitted a zero-mean term of order $\mathcal{O}((1-a_x)^t)$ in the expression. Note that the first part of (12) is equivalent to (11), while the remainder parts are zero-mean. We observe that FSPDA-STORM can be regarded as an extension of EXTRA/GT to random time varying topology, yet with control variate that helps to reduce variance in the stochastic updates.

3 Convergence Analysis of FSPDA

This section presents the convergence analysis of FSPDA. Unless otherwise specified, we focus on the case with smooth but possibly non-convex objective function. Specifically,

Assumption 3.1. Each f_i is L-smooth and is lower bounded, i.e., for $i=1,\ldots,n$,

$$\|\nabla f_i(\mathbf{x})-\nabla f_i(\mathbf{y})\|\le L\|\mathbf{x}-\mathbf{y}\|,\;\; f_i(\mathbf{x})\ge f_\star>-\infty,\;\forall\,\mathbf{x},\mathbf{y}\in\mathbb{R}^d.\tag{13}$$

We consider the random graph realizations $\mathcal{G}(\xi_a)$ as subgraphs of the fixed network topology $\mathcal{G}$. Let $\mathbf{R}=\mathbb{E}\left[\mathbf{I}(\xi_a)\right]$. We require $\mathcal{G}$ to be connected and each edge of $\mathcal{G}$ is selected with a non-zero probability, i.e., $\mathrm{diag}(\mathbf{R})>\mathbf{0}$. This leads to the following property concerned with the *expected graph Laplacian matrix* $\mathbf{A}^\top\mathbf{R}\mathbf{A}=\mathbb{E}\left[\mathbf{A}(\xi_a)^\top\mathbf{A}\right]$.

Define the projection matrix onto the orthogonal subspace of consensual vectors $\mathbf{K}:=(\mathbf{I}_n-\mathbf{1}\mathbf{1}^\top/n)\otimes\mathbf{I}_d$, our analysis uses a condition similar to [Koloskova et al., 2020, Assumption 4]:

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

$$\rho_{\min}\mathbf{K}\preceq\mathbf{A}^\top\mathbf{R}\mathbf{A}\preceq\rho_{\max}\mathbf{K},\;\;\bar{\rho}_{\min}\mathbf{K}\preceq\mathbf{A}^\top\mathbf{A}\preceq\bar{\rho}_{\max}\mathbf{K}.\tag{14}$$

It also holds $\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{K} = \mathbf{A}^\top\mathbf{R}\mathbf{A} = \mathbf{K}\mathbf{A}^\top\mathbf{R}\mathbf{A}$. Note if we define $\mathbf{Q} := (\mathbf{A}^\top\mathbf{R}\mathbf{A})^\dagger$ such that it holds $\mathbf{Q}\mathbf{A}^\top\mathbf{R}\mathbf{A} = \mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{Q} = \mathbf{K}$, we also have $\rho_{\max}^{-1}\mathbf{K} \preceq \mathbf{Q} \preceq \rho_{\min}^{-1}\mathbf{K}$.

As mentioned, the existence of $\infty > \rho_{\max} \geq \rho_{\min} > 0$ is guaranteed if $\mathrm{diag}(\mathbf{R}) > \mathbf{0}$. Note the choice of $\mathbf{I}(\xi_a)$ and thus the random subgraphs can be arbitrary. In the extreme case, our setup allows $\mathcal{G}(\xi_a) = (\mathcal{V}, \{(i,j)\})$ where only one random edge $(i,j) \in \mathcal{E}$ is drawn in each random graph. Moreover, if $\gamma \leq \rho_{\min}/\rho_{\max}^2$, we observe that

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

The expected operator $(\mathbf{I} - \gamma\mathbf{A}^\top\mathbf{R}\mathbf{A}) = \mathbb{E}[\mathbf{I} - \gamma\mathbf{A}^\top\mathbf{A}(\xi_a)]$ thus serves as a 'mixing matrix' in average consensus and $\rho_{\min}$ quantifies the rate of mixing, playing a similar role as the expected contraction in [Koloskova et al., 2020, Eq. (12)].

We consider the following condition on stochastic gradients:

Assumption 3.3. There exists $\sigma_i \geq 0$ such that for every $\mathbf{x}_i \in \mathbb{R}^d$, $i \in [n]$,

$$\mathbb{E}_{\xi_i \sim \mathbb{P}_i}[\|\nabla f_i(\mathbf{x}_i; \xi_i) - \nabla f_i(\mathbf{x}_i)\|^2] \leq \sigma_i^2. \tag{15}$$

We also define $\bar{\sigma}^2 := (1/n)\sum_{i=1}^n \sigma_i^2$. While a direct extension to the stochastic error growth condition (i.e., replacing the R.H.S. of (15) by $\sigma_i^2(1 + \|\nabla f_i(\mathbf{x}_i)\|^2)$ is possible, we will stick with the above standard assumption for brevity. On the other hand,

Lemma 3.4. *Under Assumption 3.2, there exists $\sigma_A \geq 0$ such that for every $\mathbf{x} \in \mathbb{R}^{nd}$,*

$$\mathbb{E}_{\xi_a \sim \mathbb{P}_a}[\|\mathbf{A}(\xi_a)^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}\|^2] \leq \sigma_A^2\|\mathbf{x}\|_{\mathbf{K}}^2. \tag{16}$$

See the proof in Appendix C. In particular, we show

$$\sigma_A^2 \leq \mathbb{E}[\|\mathbf{I}_{nd}(\xi_a) - \mathbf{R}\|^2] \cdot \bar{\rho}_{\max} \cdot \|\mathbf{A}\|^2, \tag{17}$$

i.e., σ_A^2 decreases if the fraction of edges selected in each realization of the random subgraphs $\mathcal{G}(\xi_a)$ increases. Finally, the subsequent discussions use the shorthand notations:

$$\bar{\mathbf{x}}^t := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^t, \quad \sum_{i=1}^n \|\mathbf{x}_i^t - \bar{\mathbf{x}}^t\|^2 = \|\mathbf{x}^t\|_{\mathbf{K}}^2. \tag{18}$$

Convergence of `FSPDA-SA`. We show that (proof in Appendix D):

Theorem 3.5. *Under Assumptions 3.1, 3.2, 3.3. Suppose that the step sizes satisfy the conditions defined in (52). Then, for any $T \geq 1$ with the random stopping iteration $\mathsf{T} \sim \mathrm{Unif}\{0, ..., T-1\}$, the iterates generated by `FSPDA-SA` satisfy*

$$\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^{\mathsf{T}})\|^2 + \frac{\mathtt{a}\gamma\rho_{\min}}{\alpha}\sum_{i=1}^n \|\mathbf{x}_i^{\mathsf{T}} - \bar{\mathbf{x}}^{\mathsf{T}}\|^2\right] \leq \frac{F_0 - f_\star}{\alpha T/8} + 8\alpha\mathbb{C}_\sigma\frac{\bar{\sigma}^2}{n}, \tag{19}$$

for any $\mathtt{a} > 0$ and F_0, $\mathbb{C}_\sigma$ are defined in (50), (56).

Setting $\mathtt{a} = \min\{1/n^2, 1/(\sum_{i=1}^n \|\mathbf{x}_i^0 - \bar{\mathbf{x}}^0\|^2)\}$, $\gamma = \gamma_\infty$, $\eta = \eta_\infty$, $\beta = 1$, $\alpha = \sqrt{n/(T\bar{\sigma}^2)}$ (assuming $\bar{\sigma} > 0$) for $\eta_\infty, \gamma_\infty$ defined in (52), we have

$$\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^{\mathsf{T}})\|^2 + \frac{\mathtt{a}\gamma_\infty\rho_{\min}\bar{\sigma}\sqrt{T}}{\sqrt{n}}\sum_{i=1}^n \|\mathbf{x}_i^{\mathsf{T}} - \bar{\mathbf{x}}^{\mathsf{T}}\|^2\right] = \mathcal{O}\left(\frac{L\bar{\sigma}}{\sqrt{nT}}\right), \tag{20}$$

which is nearly the same *asymptotic convergence rate* as centralized SGD that takes n stochastic gradient samples uniformly from each agent, i.e., linear speedup [Lian et al., 2017]. When $\sigma_A \gg 1$, taking $\gamma = \gamma_\infty = \rho_{\min}/(2\rho_{\max}\sigma_A^2)$ leads to $\mathbb{E}\left[\sum_{i=1}^n \|\mathbf{x}_i^{\mathsf{T}} - \bar{\mathbf{x}}^{\mathsf{T}}\|^2\right] = \mathcal{O}(\sigma_A^2\rho_{\max}/(\mathtt{a}T\rho_{\min}^2))$ under the step size choices in (20). Notice that for $T \gg 1$, the effect of random topology only degrades the convergence of consensus error, keeping the transient rate in (20) unaffected. If the gradients are deterministic ($\bar{\sigma} = 0$), setting $\mathtt{a} = (L^2\eta_\infty\rho_{\min})^{1/3}$, $\alpha = \alpha_\infty$ will yield a better convergence rate as $\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^{\mathsf{T}})\|^2\right] = \mathcal{O}(\sigma_A^4\sqrt{n}/T)$. Without a transient phase, the error due to random graph and coordinate sparsification is persistent through σ_A^4 in the above convergence rate. Lastly, we discuss the transient time and effects of random topology in Appendix B.

The convergence can be further accelerated if (1) satisfies the Polyak-Lojasiewicz (PL) condition:

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

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

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

$$\mathbb{E}_t[F_{t+1} - f_\star] \leq (1-\delta)(F_t - f_\star) + \mathbb{C}_\sigma \alpha^2 \bar{\sigma}^2 / n \tag{21}$$

for $F_t, \mathbb{C}_\sigma$ defined in (50), (56), and $\delta = \min\{\alpha\mu/4,\ \gamma\rho_{\min}/16,\ \eta\beta/(3\rho_{\min}), \eta/12\}$.

The proof can be found in Appendix D.6. By setting $\alpha = c\ln(T)/(n^2T)$ in (21), with a carefully chosen c and a sufficiently large T such that $\alpha \leq \alpha_\infty$, we have $\mathbb{E}[F(\bar{\mathbf{x}}^T) - f_\star + \|\mathbf{x}^T\|^2_{\mathbf{K}}] = \mathcal{O}\left(\bar{\sigma}^2 \ln(T)/(\mu n T)\right)$. When $\bar{\sigma}^2 = 0$, setting $\alpha = \alpha_\infty$ in (21) ensures a linear convergence rate of $\mathbb{E}[F(\bar{\mathbf{x}}^T) - f_\star + \|\mathbf{x}^T\|^2_{\mathbf{K}}] = \mathcal{O}((1-\delta)^T)$, 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 transmission round per iteration.

Convergence of `FSPDA-STORM`. To exploit the variance reduction with control variates, we need:

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

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

The above assumption, which implies the smoothness in Assumption 3.1, is also known as the mean-square smoothness condition. We observe the convergence guarantee for `FSPDA-STORM` (proof in Appendix E).

Theorem 3.9. *Under Assumptions 3.1, 3.2, 3.3, 3.8. Suppose that the step sizes satisfy the conditions in (192) - (222). Then, for any $T \geq 1$ with the random stopping iteration $\mathsf{T} \sim \mathrm{Unif}\{0, ..., T-1\}$, the iterates generated by `FSPDA-STORM` satisfy*

$$\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^{\mathsf{T}})\|^2 + \frac{\mathtt{a}\gamma\rho_{\min}}{4\alpha}\sum_{i=1}^n \|\mathbf{x}_i^{\mathsf{T}} - \bar{\mathbf{x}}^{\mathsf{T}}\|^2\right] \leq \frac{F_0 - f_\star}{T\alpha/4} + \frac{(\mathtt{e}\cdot 2a_x^2 + \mathtt{f}\cdot 4a_x^2 n)\bar{\sigma}^2}{\alpha/4}, \tag{23}$$

where the constants F_0, $\mathtt{a}, \mathtt{e}, \mathtt{f}$ are defined in (119).

By the hyperparameter choices in Theorem E.8 and initializing such that $\|\mathbf{v}^0\|^2_{\mathbf{K}} = \mathcal{O}(n^{-1}T^{-2/3})$, $\|\overline{\mathbf{m}}_x^0 - (1/n)\mathbf{1}_\otimes^\top \nabla \mathbf{f}(\mathbf{x}^0)\|^2 = \mathcal{O}(T^{-1/3})$ and $\|\mathbf{m}_x^0 - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^0, \boldsymbol{\lambda}^0)\|^2 = \mathcal{O}(T^{-1/3})$, we have

$$\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^{\mathsf{T}})\|^2 + \frac{L^2\rho_{\min}}{4n}\sum_{i=1}^n \|\mathbf{x}_i^{\mathsf{T}} - \bar{\mathbf{x}}^{\mathsf{T}}\|^2\right] = \mathcal{O}\left(\frac{(L^2(L_s^2 + L^2))^{1/3}\bar{\sigma}^{2/3}}{T^{2/3}}\right). \tag{24}$$

The displayed dependence on L, L_s in (24) assumes $L \geq 1$, and the dependence for the other case ($L < 1$) can be inferred from Appendix E.8. The convergence rate of `FSPDA-STORM` matches the lower bound T-dependence [Arjevani et al., 2023] for non-convex functions under the same smoothness assumption. By the same choice of step sizes, the consensus error converges at the rate of $\mathbb{E}\left[\sum_{i=1}^n \|\mathbf{x}_i^{\mathsf{T}} - \bar{\mathbf{x}}^{\mathsf{T}}\|^2\right] = \mathcal{O}(n\bar{\sigma}^{2/3}\rho_{\min}^{-1}T^{-2/3})$. We remark that in (24), 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` surpasses that of other decentralized algorithms on time varying topologies.

Insights from the Analysis. Unlike the analysis performed in [Nedic et al., 2017, Liu et al., 2024] and other prior works on arbitrary time varying graphs, Theorems 3.5 and 3.9 do not assume strong convexity with the objective function in (1). As a result, the mean field updates in `FSPDA` no longer constitute a contractive map that drive the iterates to converge to a unique point. As a remedy, we constructed a new Lyapunov function to control the relevant quantities in the algorithm.

Below we shed some insights from the analysis of `FSPDA-SA` that explains why the algorithms converge under laxer conditions. From (8a), the following relationship holds for $\bar{\mathbf{x}}^t$:

$$\bar{\mathbf{x}}^{t+1} = \bar{\mathbf{x}}^t - \frac{\alpha}{n}\sum_{i=1}^n \nabla f_i(\mathbf{x}_i^t; \xi_i^t). \tag{25}$$

Thus the evolution of $\{\bar{\mathbf{x}}^t\}_{t\geq 0}$ is analogus to centralized SGD applied on (1) except that the local gradients are evaluated on $\{\mathbf{x}_i^t\}_{i=1}^n$. However, it is still not straightforward to analyze the convergence of FSPDA-SA as the update of $\mathbf{x}^t$ involves the dual variable $\boldsymbol{\lambda}^t$ which lacks an intuitive interpretation.

The fixed point(s) of (8) exhibits additional structures whose control can help us in establishing the convergence of FSPDA-SA. Suppose that for some $t_\star$, the fixed point conditions $\mathbb{E}[\boldsymbol{\lambda}^{t_\star+1} \mid \xi^{:t_\star}] = \boldsymbol{\lambda}^{t_\star}, \mathbb{E}[\mathbf{x}^{t_\star+1} \mid \xi^{:t_\star}] = \mathbf{x}^{t_\star}$ hold. Since $\mathbf{R} \neq \mathbf{0}$ is a diagonal matrix,

$$\mathbb{E}[\boldsymbol{\lambda}^{t_\star+1} \mid \xi^{:t_\star}] = \boldsymbol{\lambda}^{t_\star} \iff \mathbf{R}\mathbf{A}\mathbf{x}^{t_\star} = \mathbf{0} \iff \mathbf{A}\mathbf{x}^{t_\star} = \mathbf{0}. \tag{26}$$

On the other hand, the primal update yields

$$\mathbb{E}[\mathbf{x}^{t_\star+1} \mid \xi^{:t_\star}] = \mathbf{x}^{t_\star} - \alpha\nabla\mathbf{f}(\mathbf{x}^{t_\star}) - \eta\mathbf{A}^\top\boldsymbol{\lambda}^{t_\star}. \tag{27}$$

Since $\mathbf{x}_1^{t_\star} = \cdots = \mathbf{x}_n^{t_\star}$ at the fixed point (due to (26)), the consensus condition yields

$$\begin{aligned}
&\mathbb{E}[\mathbf{x}^{t_\star+1} \mid \xi^{:t_\star}] - \mathbf{x}^{t_\star} = (\mathbf{1}\otimes\mathbf{I}_d)(\mathbb{E}[\bar{\mathbf{x}}^{t_\star+1} \mid \xi^{:t_\star}] - \bar{\mathbf{x}}^{t_\star}) \\
\iff &\alpha\nabla\mathbf{f}(\mathbf{x}^{t_\star}) + \eta\mathbf{A}^\top\boldsymbol{\lambda}^{t_\star} = \frac{\alpha}{n}(\mathbf{1}\mathbf{1}^\top\otimes\mathbf{I}_d)\nabla\mathbf{f}(\mathbf{x}^{t_\star}) \\
\iff &\eta\mathbf{A}^\top\boldsymbol{\lambda}^{t_\star} = \alpha\left(\frac{1}{n}\mathbf{1}\mathbf{1}^\top - \mathbf{I}_n\right)\otimes\mathbf{I}_d\ \nabla\mathbf{f}((\mathbf{1}\otimes\mathbf{I})\bar{\mathbf{x}}^{t_\star}).
\end{aligned} \tag{28}$$

From (28), we see that $\widehat{\boldsymbol{\lambda}}^t$ shall converge to the difference between global and local gradient. Inspired by the above and to facilitate our analysis, define

$$\mathbf{v}^t := \mathbf{A}^\top\boldsymbol{\lambda}^t + \frac{\alpha}{\eta}\nabla\mathbf{f}((\mathbf{1}\otimes\mathbf{I})\bar{\mathbf{x}}^t), \tag{29}$$

for any $t \geq 0$. We see that $\|\mathbf{v}^t\|_{\mathbf{K}}^2$ measures the violation of (28) in tracking the average deterministic gradient using the stochastic dual variables in FSPDA-SA. The latter will be instrumental in analyzing the consensus error bound, as revealed in Lemma D.2.

4 Numerical Experiments

This section reports the numerical experiments on practical performance of FSPDA. For the time varying topology, we take an extreme setting where for each realization $\mathcal{G}(\xi_a)$, only one edge will be selected uniformly at random from $\mathcal{G}$. We evaluate the performance with the worst-agent metric, i.e., we present the training loss as $\max_{i\in[n]} F(\mathbf{x}_i^t)$, and the stationarity/gradient-norm measure as $\max_{i\in[n]} \|\nabla F(\mathbf{x}_i^t)\|^2$. This captures the worst-case of the solutions produced by the algorithms. Unless otherwise specified, all algorithms are initialized with $\mathbf{x}_i^0 = \bar{\mathbf{x}}^0$, and for FSPDA we initialize $\widehat{\boldsymbol{\lambda}}^0 = \mathbf{m}_{x,i}^0 = \mathbf{m}_{\lambda,i}^0 = \mathbf{0}$, and the stochastic gradients are estimated with a batch size of 256. In the interest of space, omitted details and hyperparameters of the experiments can be found in Appendix G. The underlying static topology $\mathcal{G}$ is drawn as an Erdos-Renyi (ER) graph with connectivity $p = 0.5$. NVIDIA RTX 3090 GPUs are used to perform the experiments.

The hyperparameters of FSPDA are chosen by tuning over $\alpha, \eta \in \{10^{-k} : k \in \mathbb{N}\}$, $a_x, a_\lambda \in \{0.1, 0.01\}$, while the remaining parameters use the default values $\gamma = 0.5$, $\beta = 1$ for FSPDA-SA and $\beta = 0.1$ for FSPDA-STORM.

MNIST Experiments. The first set of experiments considers a moderate-scale setting of training a one hidden layer feed-forward neural network with 100 hidden neurons (total number of parameters $d = 79{,}510$) on the MNIST dataset with $m = 60,000$ samples of 784-dimensional features. The dataset is split among a $n = 10$ node network.

We compare FSPDA-SA, FSPDA-STORM with benchmark algorithms utilizing different types of time varying topology. Among them, DSGD [Koloskova et al., 2020], Swarm-SGD [Nadiradze et al., 2021] use the general time varying topology setting as FSPDA where each edge of $\mathcal{G}(\xi_a)$ is active uniformly at random, in addition to random sparsification used FSPDA-SA and adaptive quantized used in Swarm-SGD; CHOCO-SGD [Koloskova et al., 2019b] takes $\mathcal{G}(\xi_a)$ as an broadcasting subgraph where one agent selects all its neighbors; Decen-Scaffnew [Mishchenko et al., 2022], LED [Alghunaim, 2024], and K-GT [Liu et al., 2024] utilize local updates where $\mathcal{G}(\xi_a)$ is either taken as an empty topology, or as the static topology $\mathcal{G}$. We configure these algorithms such that they have the same

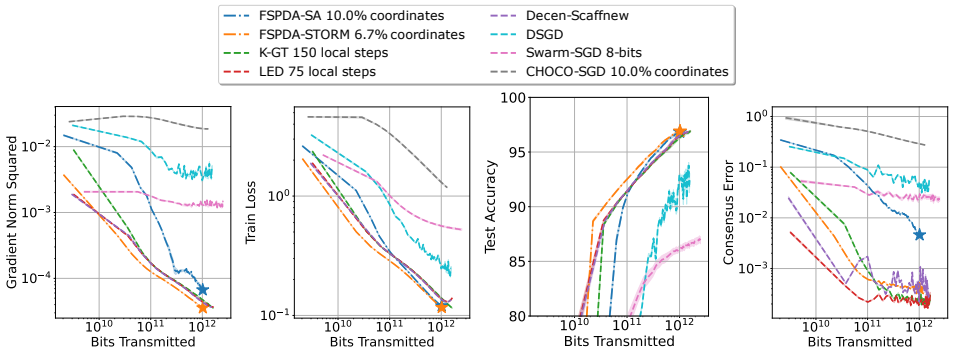

Figure 1: Feed-forward neural network classification training on MNIST using 10^6 iterations. The shaded area illustrates the variance of each metric by 1 standard deviation of 5 random seeded runs.

communication cost (in terms of bits transmitted over network) *on average*. For instance, the local update algorithms (`Decen-Scaffnew`, `LED`, `K-GT`) only communicate once using $\mathcal{G}$ every $\mathcal{O}\left(\frac{|\mathcal{E}|d}{k}\right)$ iterations to match the communication cost of k-coordinate sparse one-edge random graph in `FSPDA`.

The local objective function held by each agent is the cross-entropy classification loss on a local dataset with $m_i = 6,000$ samples, plus a regularization loss $\frac{\lambda}{2}\|\mathbf{x}_i\|^2$ with $\lambda = 10^{-4}$, where $\mathbf{x}_i$ are the weight parameters of the feed-forward neural network classifier. We split the training set into $n = 10$ disjoint sets such that each set contains only one class label and assign each set to one agent as its local dataset. Note that as we do not shuffle the data samples across local datasets, the local objective function held by different agents will become highly heterogeneous.

Fig. 1 compares the squared gradient norm, training loss, consensus error of the benchmarked algorithms. We first note that both `FSPDA` algorithms have significantly outperformed `DSGD`, `Swarm-SGD` on the general time varying topology as well as `CHOCO-SGD`. Meanwhile, the performance of `FSPDA` is comparable to the local update algorithms `Decen-Scaffnew`, `LED`, `K-GT`. Notice that the latter 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 ablation studies over data heterogeneity, sparsity levels, graph topologies, gradient noise and dual momentum in Appendix F.

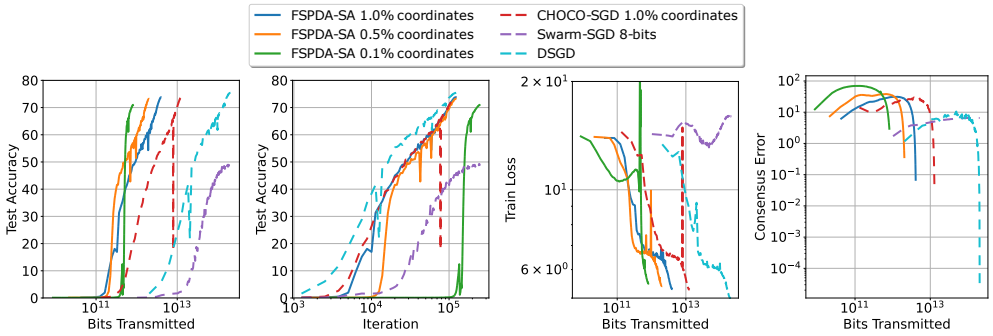

Figure 2: Resnet-50 classification training on Imagenet.

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 1,000 classes, re-scaled and cropped to 256×256 image dimensions) split among $n = 8$ nodes. We consider cross-entropy classification loss plus the same L2 norm regularization loss as in the previous setup. The performance metrics are measured at the network average iterate $\bar{\mathbf{x}}^t$. 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 α, η of `FSPDA-SA` are scheduled simultaneously such that α_t/η_t remains constant, as illustrated in Appendix G. We draw a batch of 128 samples to estimate the stochastic gradient.

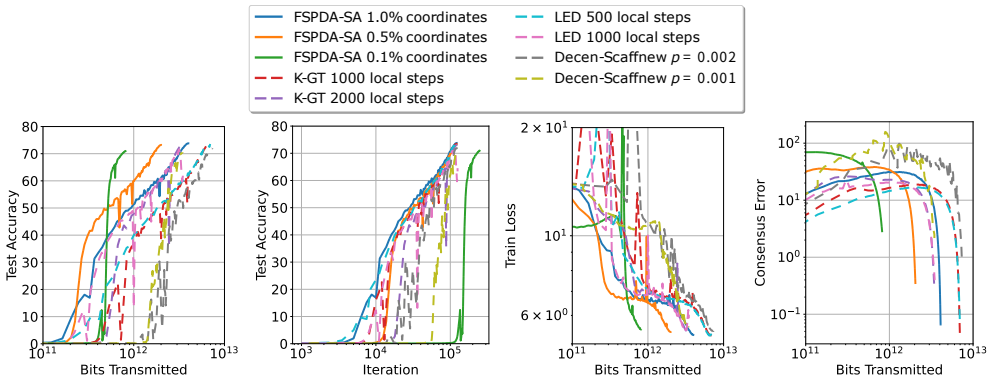

Figure 3: Resnet-50 classification training on Imagenet.

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 and 3, which compare the test accuracy and training loss against iteration number and bits transmitted. When compared to the existing algorithms, `FSPDA-SA` achieves almost the same accuracy using one-edge random graphs with at least 10x reduction (against `CHOCO-SGD`) and around 1.5x reduction (against static graph algorithms `K-GT`, `LED` and `Decen-Scaffnew`) 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 a comparable level of accuracy.

5 Conclusion

This paper proposed the `FSPDA` framework for decentralized optimization over arbitrary time varying random networks. We utilize a new stochastic equality reformulation and apply SA to search for a saddle point for its augmented Lagrangian function. 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 bounded heterogeneity and the type of time varying topologies. The proposed framework may open up new directions to design fast decentralized optimization over time varying networks.

Acknowledgments and Disclosure of Funding

This research was supported in part by project #MMT-p5-23 of the Shun Hing Institute of Advanced Engineering, The Chinese University of Hong Kong.

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.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE transactions on information theory*, 52(6):2508–2530, 2006.

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.

Yiyue Chen, Abolfazl Hashemi, and Haris Vikalo. Accelerated distributed stochastic non-convex optimization over time-varying directed networks. *IEEE Transactions on Automatic Control*, 2024.

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.

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

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

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

Liyuan Liang, Yilong Song, and Kun Yuan. Decentralized optimization over time-varying row-stochastic digraphs. *arXiv preprint arXiv:2512.24483*, 2025.

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

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

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

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

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

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

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

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

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

Edward Duc Hien Nguyen, Xin Jiang, Bicheng Ying, and César A Uribe. On graphs with finite-time consensus and their use in gradient tracking. *SIAM Journal on Optimization*, 35(2):872–898, 2025.

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

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

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

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

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

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

Chung-Yiu Yau and Hoi-To Wai. Fully stochastic distributed convex optimization on time-varying graph with compression. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 145–150. IEEE, 2023.

Xinlei Yi, Shengjun Zhang, Tao Yang, Tianyou Chai, and Karl H Johansson. Linear convergence of first-and zeroth-order primal–dual algorithms for distributed nonconvex optimization. *IEEE Transactions on Automatic Control*, 67(8):4194–4201, 2021.

Yijie Zhou and Shi Pu. Accelerating decentralized optimization via overlapping local steps. *arXiv preprint arXiv:2601.01493*, 2026.

A Asynchronous Implementation of FSPDA

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

The thread `communication_thread()` aims at preparing the message $\sum_{j\in\mathcal{N}_i(\xi_a^t)} \mathbf{C}_{ij}(\xi_a^t)(\mathbf{x}_j^t - \mathbf{x}_i^t)$ needed at (8a), (8b) using a gossip-like step. At each run by agent i, the agent checks if it is 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 agents (i, j) are connected, they will exchange the sparsified decision variable $\{(\mathbf{x}_{i,k}^{t_i}, \mathbf{x}_{j,k}^{t_j})\}_{k\in\mathcal{I}_{ij}(\xi_a^t)}$ and make preparation for the computation thread. The neighbor j whose communicated with agent i will be added to the buffer $\mathcal{B}_i$. Notice that the protocol is designed such that the asynchronous implementation of FSPDA-SA aligns with (8a), (8b). For instance, our asynchronous gradient model with $b_i(\xi^t) = 0$ allows agent i to skip gradient computation and the sparse extended graph model (6) allows agent i to skip communication, as implemented in [L2, Alg. 2] which is performed after the *idle* state of agent i to compensate for the missed iterations led by neighbor $j \in \mathcal{B}_i$ using the locally stored $\widehat{\boldsymbol{\lambda}}_i^{t_i}$.

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

Overall, by running the two parallel threads persistently at each agent, the decentralized system effectively implements FSPDA-SA in (8a), (8b) as an asynchronous algorithm. The same asynchronous implementation 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

local variable: buffer $\mathcal{B}_i$; counters g_i, t_i.
for $(i,j) \in \mathcal{E}(\xi_a^t)$ **do**
  if $\mathcal{B}_i = \emptyset$ and $\mathcal{B}_j = \emptyset$ **then**
    Agents i, j exchanges t_i, t_j.
    if $t_i < t_j$ **then**
      Interrupt [L3, Alg. 3] of *computation_thread()* and run [L3, Alg. 3] to consume (any) SG buffer.
      ➥ *Local non-SG step:* evaluate

$$\mathbf{x}_i^{t_j} = \mathbf{x}_i^{t_i} - (t_j - t_i)\eta \widehat{\boldsymbol{\lambda}}_i^{t_i}, \tag{30}$$

      and set $\widehat{\boldsymbol{\lambda}}_i^{t_j} = \widehat{\boldsymbol{\lambda}}_i^{t_i}$, $t_i \leftarrow t_j$.
    end if
    `//* begin streaming index-value pairs *//`
    for $k \in \mathcal{I}_{ij}(\xi_a^t)$ **do**
      Agent j receives $(k, \mathbf{x}_{i,k}^{t_i})$ from agent i.
      Agent i receives $\mathbf{x}_{j,k}^{t_j}$ from agent j.
    end for
    `//* end streaming upon time-out or interruption *//`
    if gossip is successful **then**
      Update $\mathcal{B}_i \leftarrow \{j\}$, $\mathcal{B}_j \leftarrow \{i\}$.
    end if
  end if
end for

Algorithm 3 *computation_thread()* of Agent i

local variable: buffer $\mathcal{B}_i$; counters g_i, t_i.
`//* 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 *//`
if communication buffer $\mathcal{B}_i = \emptyset$ **then**
  ➥ *Local SG step:* with $\hat{c}_i = g_i/(t_i+1)$, evaluate

$$\mathbf{x}_i^{t_i+1} = \mathbf{x}_i^{t_i} - \eta \widehat{\boldsymbol{\lambda}}_i^{t_i} - \alpha \hat{c}_i \nabla f_i(\mathbf{x}_i^{t_i}; \xi_i^{t_i}), \tag{31}$$

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

$$t_i' = \max\{t_i,\ \max_{j \in \mathcal{B}_i} t_j\}, \quad d_i = 1 + t_i' - t_i, \quad \mathbf{C}_{ij}(\xi^{t_i'}) = \mathrm{BinDiag}(\mathcal{I}_{ij}(\xi^{t_i'})), \tag{32}$$

$$\hat{c}_i = \begin{cases} g_i/(t_i'+1) & \text{if } \nabla f_i(\mathbf{x}_i^{t_i}; \xi_i^{t_i'}) \text{ is ready,} \\ 0 & \text{otherwise.} \end{cases} \tag{33}$$

  ➥ *Gossip with SG step*: evaluate

$$\mathbf{x}_i^{t_i'+1} = \mathbf{x}_i^{t_i} - \gamma \textstyle\sum_{j \in \mathcal{B}_i} \mathbf{C}_{ij}(\xi^{t_i'})(\mathbf{x}_i^{t_i} - \mathbf{x}_j^{t_j}) - d_i \eta \widehat{\boldsymbol{\lambda}}_i^{t_i} - \alpha \hat{c}_i \nabla f_i(\mathbf{x}_i^{t_i}; \xi_i^{t_i'}), \tag{34}$$

$$\widehat{\boldsymbol{\lambda}}_i^{t_i'+1} = \widehat{\boldsymbol{\lambda}}_i^{t_i} + \beta \textstyle\sum_{j \in \mathcal{B}_i} \mathbf{C}_{ij}(\xi^{t_i'})(\mathbf{x}_i^{t} - \mathbf{x}_j^{t}), \tag{35}$$

  and set $t_i \leftarrow t_i' + 1$, $\mathcal{B}_i \leftarrow \emptyset$.
end if

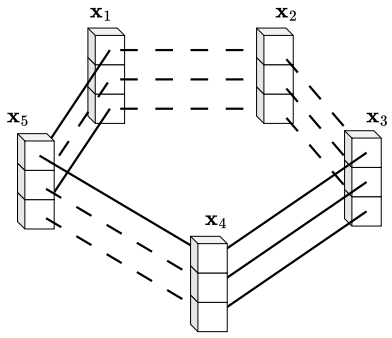

Figure 4: Illustration of a (time varying) random graph $\mathcal{G}(\xi)$ for primal variable of dimension $d = 3$ on a ring network of $n = 5$ nodes. Solid lines represent active edges while dashed lines represent disconnected edges. In this example, node 2 is considered as idle in an asynchronous environment. $\mathbf{C}_{15}(\xi)$ is a diagonal matrix such that $\mathrm{diag}(\mathbf{C}_{15}(\xi)) = (1, 0, 1)$.

B Detailed Convergence Rate Analysis for FSPDA-SA

Under the parameter choices of $\mathtt{a} = \min\{1/n^2, 1/(\sum_{i=1}^n \|\mathbf{x}_i^0 - \bar{\mathbf{x}}^0\|^2)\}$, $\gamma = \gamma_\infty$, $\eta = \eta_\infty$, $\beta = 1$, $\alpha = \sqrt{n/(T\bar{\sigma}^2)}$, we list several remarks on the convergence rate analysis of FSPDA-SA as follows.

- In addition to the crude bound in (20), 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 (20), independent of the network topology [Pu et al., 2021]. Using the upper bound of $\mathbb{C}_\sigma$ in (56) and consider $\rho_{\max} \geq 2$, when $\rho_{\max} < 2\sigma_A^2$, it can be shown that FSPDA-SA has a transient time of

$$T_{\mathrm{trans}} = \Omega\left(\max\left\{\frac{n}{\bar{\sigma}^2}, \frac{\bar{\rho}_{\max}^4 \rho_{\max}^6 \sigma_A^4}{\bar{\sigma}^2 n \rho_{\min}^8}, T_{\mathrm{feas}}\right\}\right), \tag{36}$$

  with a feasible time T_{feas} to satisfy the condition $\alpha \leq \alpha_\infty(\eta_\infty)$ such that

$$T_{\mathrm{feas}} = \frac{n}{\bar{\sigma}^2 \alpha_\infty(\eta_\infty)^2} = \Omega\left(\frac{n^6 \sigma_A^4 \rho_{\max}^2}{\bar{\sigma}^2 \rho_{\min}^4}\right), \tag{37}$$

  where we have hidden the dependence of L, $F(\bar{\mathbf{x}}^0) - f_\star$, $\|\mathbf{x}^0\|_{\mathbf{K}}^2$, $\|\mathbf{v}^0\|_{\mathbf{K}}^2$ in the $\Omega(\cdot)$ notation. When $\sigma_A^2 \leq \rho_{\max}/2$, e.g., where the graphs are deterministic with $\sigma_A = 0$, it can be shown similarly that

$$T_{\mathrm{trans}} = \Omega\left(\max\left\{\frac{n}{\bar{\sigma}^2}, \frac{\bar{\rho}_{\max}^4 \rho_{\max}^8}{\bar{\sigma}^2 n \rho_{\min}^8}, T_{\mathrm{feas}}\right\}\right). \tag{38}$$

- Theorem 3.5 also guarantees that $\mathbb{E}[\|\mathbf{x}^{\mathsf{T}}\|_{\mathbf{K}}^2] \to 0$ and $\mathbb{E}[\|\mathbf{v}^{\mathsf{T}}\|_{\mathbf{K}}^2] \to 0$ as $T \to \infty$. 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 (29)). 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 $\mathbb{E}\left[\|\mathbf{x}^{\mathsf{T}}\|_{\mathbf{K}}^2\right] = \mathcal{O}(n^2\sigma_A^2\rho_{\max}/(T\rho_{\min}^2))$, keeping the rate in (20) 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 σ. 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 Lemma 3.4

We observe that

$$\mathbb{E}[\|\mathbf{A}^\top \mathbf{A}(\xi_a)\mathbf{x} - \mathbf{A}^\top \mathbf{R}\mathbf{A}\mathbf{x}\|^2] \tag{39}$$
$$= \mathbb{E}[\|\mathbf{A}^\top \mathbf{I}_{nd}(\xi_a)\mathbf{A}\mathbf{x} - \mathbf{A}^\top \mathbf{R}\mathbf{A}\mathbf{x}\|^2] \tag{40}$$
$$= \mathbb{E}[\|(\mathbf{A}^\top \mathbf{I}_{nd}(\xi_a) - \mathbf{A}^\top \mathbf{R})(\mathbf{A}\mathbf{x})\|^2] \tag{41}$$
$$\leq \mathbb{E}[\|\mathbf{A}^\top \mathbf{I}_{nd}(\xi_a) - \mathbf{A}^\top \mathbf{R}\|^2] \cdot \|\mathbf{A}\mathbf{x}\|^2 \tag{42}$$
$$\leq \mathbb{E}[\|\mathbf{I}_{nd}(\xi_a) - \mathbf{R}\|^2] \cdot \|\mathbf{A}\|^2 \cdot \bar{\rho}_{\max} \cdot \|\mathbf{x}\|^2_{\mathbf{K}} \tag{43}$$

where $\mathbf{R} = \mathbb{E}[\mathbf{I}_{nd}(\xi_a)]$ and the last inequality is due to Assumption 3.2.

D 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:

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

$$\mathbb{E}_t\left[F(\bar{\mathbf{x}}^{t+1})\right] \leq F(\bar{\mathbf{x}}^t) - \frac{\alpha}{4}\left\|\nabla F(\bar{\mathbf{x}}^t)\right\|^2 + \frac{3\alpha L^2}{4n}\|\mathbf{x}^t\|^2_{\mathbf{K}} + \frac{\alpha^2 L}{2n}\bar{\sigma}^2. \tag{44}$$

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

$$(1/T)\textstyle\sum_{t=1}^T \mathbb{E}[\|\nabla F(\bar{\mathbf{x}}^t)\|^2] = \mathcal{O}(1/\sqrt{T}). \tag{45}$$

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

$$\gamma \leq \min\left\{\frac{\rho_{\min}}{\rho^2_{\max}}, \frac{\rho_{\min}}{4\sigma_A^2}\right\},\ \alpha \leq 1,\ \eta \leq 1. \tag{46}$$

Lemma D.2. *Under Assumptions 3.1, 3.2, 3.3, Lemma 3.4 and the step size condition (46). The consensus error follows the recursive inequality*

$$\mathbb{E}\left[\|\mathbf{x}^{t+1}\|^2_{\mathbf{K}}\right] \leq \left[1 - \frac{\gamma}{2}\rho_{\min} + \alpha(1+3L^2)\right]\mathbb{E}\left[\left\|\mathbf{x}^t\right\|^2_{\mathbf{K}}\right] + 2\eta^2\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\mathbf{K}}\right] \tag{47}$$
$$- 2\eta\mathbb{E}\left[\left\langle \mathbf{x}^t \mid \mathbf{v}^t\right\rangle_{\mathbf{K}-\gamma\mathbf{A}^\top\mathbf{R}\mathbf{A}}\right] + 2\alpha^2 n\bar{\sigma}^2. \tag{48}$$

See Appendix D.2 for the proof.

The above lemma shows an intricate structure for the consensus error as the latter also depends on the violation of (28), i.e., $\|\mathbf{v}^t\|^2_{\mathbf{K}}$, and the *weighted* inner product between $\mathbf{x}^t, \mathbf{v}^t$. Naturally, we can further control the above terms:

Lemma D.3. *Under Assumption 3.1, 3.3. Let $\alpha \leq 1$, then for any constant $\mathsf{c} > 0$, the dual error satisfies*

$$\mathbb{E}\left[\left\|\mathbf{v}^{t+1}\right\|^2_{\mathbf{Q}+\mathsf{c}\mathbf{K}}\right] \leq \mathbb{E}\left[\left\|\mathbf{v}^t\right\|^2_{\mathbf{Q}+\mathsf{c}\mathbf{K}}\right] + 2\alpha(\rho_{\min}^{-1} + \mathsf{c})\mathbb{E}\left[\left\|\mathbf{v}^t\right\|^2_{\mathbf{K}}\right] + 2\beta\mathbb{E}\left[\left\langle \mathbf{v}^t \mid \mathbf{x}^t\right\rangle_{\mathbf{K}+\mathsf{c}\mathbf{A}^\top\mathbf{R}\mathbf{A}}\right]$$
$$+ \left(2\beta^2\bar{\rho}^2_{\max} + \frac{10\alpha^3 L^4}{\eta^2}\right)(\rho_{\min}^{-1} + \mathsf{c})\mathbb{E}\left[\|\mathbf{x}^t\|^2_{\mathbf{K}}\right] + \frac{5\alpha^3}{\eta^2}L^2(\rho_{\min}^{-1} + \mathsf{c})\left\{\bar{\sigma}^2 + 2n\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^t)\|^2\right]\right\}. \tag{49}$$

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

Lemma D.4. *Under Assumption 3.1, 3.3, and the step size condition (46), then*

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

See Appendix D.4 for the proof.

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

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

$$F_t = \mathbb{E}\left[F(\bar{\mathbf{x}}^t) + \mathtt{a}\|\mathbf{x}^t\|_{\mathbf{K}}^2 + \mathtt{b}\|\mathbf{v}^t\|_{\mathbf{Q}+\mathtt{c}\mathbf{K}}^2 + \mathtt{d}\left\langle \mathbf{x}^t \mid \mathbf{v}^t\right\rangle_{\mathbf{K}}\right]. \tag{50}$$

Then, by the following choice of hyperparameters

$$\mathtt{b} = \mathtt{a}\cdot\frac{\eta}{\beta}, \quad \mathtt{c} = \frac{(\eta\beta+\gamma)\mathtt{d}-2\eta\gamma\mathtt{a}}{2\beta\mathtt{b}}, \quad \mathtt{d} = \delta_1\eta\mathtt{a}, \tag{51}$$

$$\eta \leq \eta_\infty, \quad \alpha \leq \alpha_\infty(\eta), \quad \gamma = \gamma_\infty, \quad \beta = 1, \quad \delta_1 \geq 8, \tag{52}$$

where

$$\gamma_\infty := \frac{\rho_{\min}}{\rho_{\max}^2}\min\left\{1, \frac{\rho_{\max}}{2\sigma_A^2}, \frac{\rho_{\max}^2}{4\sigma_A^2}\right\}, \ \eta_\infty := \min\left\{1, \min\left\{1, \frac{\rho_{\min}^2}{64\delta_1^2\bar{\rho}_{\max}^2\rho_{\max}^2}\right\}\gamma_\infty\right\}, \tag{53}$$

$$\alpha_\infty(\eta) := \frac{\gamma_\infty\rho_{\min}}{80\delta_1\sqrt{n}}\min\left\{\frac{\eta}{\gamma_\infty\rho_{\min}}, \frac{1}{L^2}, \frac{\mathtt{a}}{L^2}, \eta\rho_{\min}, \sqrt{\frac{\eta\rho_{\min}}{L^2\mathtt{a}}}\right\}, \tag{54}$$

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*

$$F_{t+1} \leq F_t - \frac{\alpha}{8}\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^t)\|^2\right] - \frac{\mathtt{a}\gamma\rho_{\min}}{8}\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right] - \mathtt{a}\eta^2\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathbf{K}}^2\right] + \mathbb{C}_\sigma\alpha^2\bar{\sigma}^2/n \tag{55}$$

such that

$$\mathbb{C}_\sigma \leq \frac{L}{2} + 2\mathtt{a}n^2 + \mathtt{a}n\alpha L^2\left(\frac{5}{\eta\beta\rho_{\min}} + \frac{5\delta_1\gamma}{\eta\beta} + 5\delta_1\right) + \mathtt{a}\delta_1 n^2\alpha. \tag{56}$$

See Appendix D.5 for the proof.

By summing up (55) from $t = 0$ to $t = T-1$ and rearranging terms, we obtain the bounds that lead to the theorem. □

D.1 Proof of Lemma D.1

By Assumption 3.1, it is straightforward to derive that

$$\mathbb{E}_t\left[F(\bar{\mathbf{x}}^{t+1})\right] \leq F(\bar{\mathbf{x}}^t) - \frac{\alpha}{n}\left\langle \nabla F(\bar{\mathbf{x}}^t) \mid \mathbf{1}_\otimes^\top\nabla\mathbf{f}(\mathbf{x}^t)\right\rangle + \frac{\alpha^2 L}{2n^2}\mathbb{E}_t\left[\left\|\mathbf{1}_\otimes^\top\nabla\mathbf{f}(\mathbf{x}^t;\xi^t)\right\|^2\right], \tag{57}$$

where we have used the shorthand notation $\mathbf{1}_\otimes^\top := \mathbf{1}^\top\otimes\mathbf{I}$. The second term of (57) can be bounded as

$$\frac{\alpha}{n}\left\langle \nabla F(\bar{\mathbf{x}}^t) \mid \mathbf{1}_\otimes^\top\nabla\mathbf{f}(\mathbf{x}^t)\right\rangle \geq \frac{\alpha}{2}\left\|\nabla F(\bar{\mathbf{x}}^t)\right\|^2 - \frac{\alpha L^2}{2n}\left\|\mathbf{x}^t\right\|_{\mathbf{K}}^2. \tag{58}$$

The third term of (57) can be bounded as

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

Notice that due to the independence of gradient noise, we have $\mathbb{E}_t\left[\left\|\mathbf{1}_{\otimes}^{\top}\left(\nabla\mathbf{f}(\mathbf{x}^t;\xi^t)-\nabla\mathbf{f}(\mathbf{x}^t)\right)\right\|^2\right]\leq n\bar{\sigma}^2$. Furthermore,

$$\left\|\mathbf{1}_{\otimes}^{\top}\nabla\mathbf{f}(\mathbf{x}^t)\right\|^2\leq 2nL^2\|\mathbf{x}^t\|_{\mathbf{K}}^2+2n^2\|\nabla F(\bar{\mathbf{x}}^t)\|^2. \tag{59}$$

Substituting the above into (57) and setting the step size $\alpha\leq 1/(4L)$ concludes the proof of the lemma. □

D.2 Proof of Lemma D.2

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:

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

We can simplify the primal update as

$$\begin{aligned}\mathbf{x}^{t+1}&=(\mathbf{I}-\gamma\mathbf{L})\,\mathbf{x}^t-\eta\mathbf{A}^{\top}\boldsymbol{\lambda}^t-\alpha\nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^t)+\mathbf{e}_s^t+\mathbf{e}_g^t\\ &=(\mathbf{I}-\gamma\mathbf{L})\,\mathbf{x}^t-\eta\mathbf{v}^t+\mathbf{e}_s^t+\mathbf{e}_g^t.\end{aligned} \tag{60}$$

By (60), the consensus error can be measured by

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2\right]=\mathbb{E}\left[\left\|(\mathbf{I}-\gamma\mathbf{L})\,\mathbf{x}^t-\eta\mathbf{v}^t+\mathbf{e}_s^t+\mathbf{e}_g^t\right\|_{\mathbf{K}}^2\right]\\ &\overset{(i)}{=}\mathbb{E}\left[\left\|(\mathbf{I}-\gamma\mathbf{L})\,\mathbf{x}^t-\eta\mathbf{v}^t+\mathbf{e}_g^t\right\|_{\mathbf{K}}^2\right]+\mathbb{E}\left[\|\mathbf{e}_s^t\|_{\mathbf{K}}^2\right]\\ &=\mathbb{E}\left[\left\|(\mathbf{I}-\gamma\mathbf{L})\,\mathbf{x}^t\right\|_{\mathbf{K}}^2\right]+\mathbb{E}\left[\|\eta\mathbf{v}^t-\mathbf{e}_g^t\|_{\mathbf{K}}^2\right]-2\mathbb{E}\left[\left\langle(\mathbf{I}-\gamma\mathbf{L})\,\mathbf{x}^t\;\middle|\;\eta\mathbf{v}^t-\mathbf{e}_g^t\right\rangle_{\mathbf{K}}\right]+\mathbb{E}\left[\|\mathbf{e}_s^t\|_{\mathbf{K}}^2\right],\end{aligned} \tag{61}$$

where (i) uses the conditional unbiasedness $\mathbb{E}_t[\mathbf{e}_s^t|\mathbf{x}^t,\boldsymbol{\lambda}^t]=\mathbf{0}$. Let $\gamma\leq\rho_{\min}/\rho_{\max}^2$, by Assumption 3.2, the first term of (61) can be bounded by

$$\mathbb{E}\left[\left\|(\mathbf{I}-\gamma\mathbf{L})\,\mathbf{x}^t\right\|_{\mathbf{K}}^2\right]\leq\left(1-\gamma\rho_{\min}\right)\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right]. \tag{62}$$

The second term of (61) can be bounded by

$$\mathbb{E}\left[\|\eta\mathbf{v}^t-\mathbf{e}_g^t\|_{\mathbf{K}}^2\right]\leq 2\eta^2\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathbf{K}}^2\right]+2\mathbb{E}\left[\|\mathbf{e}_g^t\|_{\mathbf{K}}^2\right], \tag{63}$$

and the third term of (61) can be bounded by

$$\begin{aligned}&-2\mathbb{E}\left[\left\langle(\mathbf{I}-\gamma\mathbf{L})\,\mathbf{x}^t\;\middle|\;\eta\mathbf{v}^t-\mathbf{e}_g^t\right\rangle_{\mathbf{K}}\right]\\ &\leq-2\eta\mathbb{E}\left[\left\langle\mathbf{x}^t\;\middle|\;\mathbf{v}^t\right\rangle_{(\mathbf{I}-\gamma\mathbf{L})\mathbf{K}}\right]+\alpha\mathbb{E}\left[\left\|(\mathbf{I}-\gamma\mathbf{L})\,\mathbf{x}^t\right\|_{\mathbf{K}}^2\right]+\frac{1}{\alpha}\mathbb{E}\left[\|\mathbf{e}_g^t\|_{\mathbf{K}}^2\right].\end{aligned} \tag{64}$$

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

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

Now we tackle the error term $\|\mathbf{e}_s^t\|^2$,

$$\mathbb{E}\left[\|\mathbf{e}_s^t\|_{\mathbf{K}}^2\right]\leq 2\alpha^2\textstyle\sum_{i=1}^n\sigma_i^2+2\gamma^2\sigma_A^2\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right], \tag{65}$$

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

$$\mathbb{E}\left[\|\mathbf{e}_g^t\|_{\mathbf{K}}^2\right]\leq\alpha^2L^2\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right]. \tag{66}$$

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

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

Using the step size condition (46) to simplify the first term completes the proof. □

D.3 Proof of Lemma D.3

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:

$$\begin{aligned}\mathbf{v}^{t+1} &= \mathbf{A}^\top \boldsymbol{\lambda}^{t+1} + \frac{\alpha}{\eta} \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^{t+1}) \\ &= \mathbf{v}^t + \underbrace{\beta \mathbf{A}^\top \mathbf{A}(\xi^t)\mathbf{x}^t + \frac{\alpha}{\eta}\nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^{t+1}) - \frac{\alpha}{\eta}\nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^t)}_{=:\Delta \mathbf{v}^t}.\end{aligned} \tag{67}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{v}^{t+1}\|^2_{\mathbf{Q}+\mathsf{c}\mathbf{K}}\right] \le \mathbb{E}\left[\|\mathbf{v}^t\|^2_{\mathbf{Q}+\mathsf{c}\mathbf{K}}\right] + 2\mathbb{E}\left[\langle \mathbf{v}^t \mid \Delta \mathbf{v}^t\rangle_{\mathbf{Q}+\mathsf{c}\mathbf{K}}\right] + 2\beta^2 \mathbb{E}\left[\|\mathbf{A}^\top \mathbf{A}(\xi^t)\mathbf{x}^t\|^2_{\mathbf{Q}+\mathsf{c}\mathbf{K}}\right] \\ + \frac{2\alpha^2}{\eta^2}\mathbb{E}\left[\left\|\nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^t)\right\|^2_{\mathbf{Q}+\mathsf{c}\mathbf{K}}\right].\end{aligned} \tag{68}$$

The second term of (68) can be simplified as

$$\begin{aligned}\mathbb{E}\left[\langle \mathbf{v}^t \mid \Delta \mathbf{v}^t\rangle_{\mathbf{Q}+\mathsf{c}\mathbf{K}}\right] \le \beta \mathbb{E}\left[\langle \mathbf{v}^t \mid \mathbf{x}^t\rangle_{\mathbf{K}+\mathsf{c}\mathbf{L}}\right] + \alpha \mathbb{E}\left[\|\mathbf{v}^t\|^2_{\mathbf{Q}+\mathsf{c}\mathbf{K}}\right] \\ + \frac{\alpha}{4\eta^2}\mathbb{E}\left[\left\|\nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^t)\right\|^2_{\mathbf{Q}+\mathsf{c}\mathbf{K}}\right],\end{aligned} \tag{69}$$

where we used $\mathbb{E}[(\mathbf{Q}+\mathsf{c}\mathbf{K})\mathbf{A}^\top \mathbf{A}(\xi^t)] = \mathbf{K} + \mathsf{c}\mathbf{L}$ and note that $\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\mathbf{Q}+\mathsf{c}\mathbf{K}}\right] \le (\rho_{\min}^{-1} + \mathsf{c})\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\mathbf{K}}\right]$ due to Assumption 3.2. The third term of (68) can be simplified as

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

The fourth term of (68) can be simplified using Lemma D.6 as

$$\begin{aligned}&\mathbb{E}\left[\left\|\nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^t)\right\|^2_{\mathbf{Q}+\mathsf{c}\mathbf{K}}\right] \\ &\le \frac{4\alpha^2 n L^2}{(\rho_{\min}^{-1} + \mathsf{c})^{-1}}\left\{\frac{1}{2n^2}\sum_{i=1}^n \sigma_i^2 + \mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^t)\|^2\right] + \frac{L^2}{n}\mathbb{E}\left[\|\mathbf{x}^t\|^2_{\mathbf{K}}\right]\right\}.\end{aligned} \tag{70}$$

Combining the above inequalities and using the condition $\alpha \le 1$ to simplify constants yields the lemma. □

D.4 Proof of Lemma D.4

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

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

then by the recursions in (60) and (67),

$$\begin{aligned}&\mathbb{E}\left[\langle \mathbf{x}^{t+1} \mid \mathbf{v}^{t+1}\rangle_{\mathbf{K}}\right] \\ &= \mathbb{E}\left[\langle \mathbf{x}^t \mid \mathbf{v}^t\rangle_{\mathbf{K}-(\gamma+\eta\beta)\mathbf{L}}\right] + \mathbb{E}\left[\|\mathbf{x}^t\|^2_{\beta(\mathbf{I}-\gamma\mathbf{L})\mathbf{L}}\right] - \eta \mathbb{E}\left[\|\mathbf{v}^t\|^2_{\mathbf{K}}\right] \\ &\quad + \frac{\alpha}{\eta}\mathbb{E}\left[\langle \mathbf{x}^t \mid \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^t)\rangle_{(\mathbf{I}-\gamma\mathbf{L})\mathbf{K}}\right] - \alpha \mathbb{E}\left[\langle \mathbf{v}^t \mid \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^t)\rangle_{\mathbf{K}}\right] \\ &\quad + \mathbb{E}\left[\langle \mathbf{e}_g^t \mid \mathbf{v}^t + \Delta \mathbf{v}^t\rangle_{\mathbf{K}}\right] + \frac{\alpha}{\eta}\mathbb{E}\left[\langle \mathbf{e}_s^t \mid \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^{t+1}) - \nabla \mathbf{f}(\mathbf{1}_\otimes \bar{\mathbf{x}}^t)\rangle_{\mathbf{K}}\right],\end{aligned} \tag{71}$$

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

$$\begin{aligned}&\mathbb{E}\left[\left\langle \mathbf{x}^t \mid \nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^{t+1}) - \nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^t)\right\rangle_{(\mathbf{I}-\gamma\mathbf{L})\mathbf{K}}\right]\\ &\leq \frac{1}{2}(1-\gamma\rho_{\min})\mathbb{E}\left[\|\mathbf{x}^t\|^2_{\mathbf{K}}\right] + \frac{1}{2}\mathbb{E}\left[\|\nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^{t+1}) - \nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^t)\|^2\right].\end{aligned} \tag{72}$$

Moreover,

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

Next, it is derived in (66) that $\mathbb{E}\left[\|\mathbf{e}_g^t\|^2_{\mathbf{K}}\right] \leq \alpha^2L^2\mathbb{E}\left[\|\mathbf{x}^t\|^2_{\mathbf{K}}\right]$. Thus,

$$\begin{aligned}&\mathbb{E}\left[\left\langle \mathbf{e}_g^t \mid \mathbf{v}^t + \Delta\mathbf{v}^t\right\rangle_{\mathbf{K}}\right]\\ &\leq \frac{1}{2}(\frac{3}{\eta})\mathbb{E}\left[\|\mathbf{e}_g^t\|^2_{\mathbf{K}}\right] + \frac{1}{2}(\frac{\eta}{3})\mathbb{E}\left[\|\mathbf{v}^t + \Delta\mathbf{v}^t\|^2_{\mathbf{K}}\right]\\ &\leq \frac{3}{2\eta}\mathbb{E}\left[\|\mathbf{e}_g^t\|^2_{\mathbf{K}}\right] + \frac{\eta}{2}\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\mathbf{K}}\right] + \frac{\eta\beta^2}{2}\mathbb{E}\left[\|\mathbf{L}\mathbf{x}^t\|^2_{\mathbf{K}}\right] + \frac{\eta}{2}\cdot\frac{\alpha^2}{\eta^2}\mathbb{E}\left[\|\nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^{t+1}) - \nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^t)\|^2_{\mathbf{K}}\right]\\ &\leq \left(\frac{3\alpha^2L^2}{2\eta} + \frac{\eta\beta^2\rho_{\max}^2}{2}\right)\mathbb{E}\left[\|\mathbf{x}^t\|^2_{\mathbf{K}}\right] + \frac{\eta}{2}\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\mathbf{K}}\right] + \frac{\eta}{2}\cdot\frac{\alpha^2}{\eta^2}\mathbb{E}\left[\|\nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^{t+1}) - \nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^t)\|^2_{\mathbf{K}}\right].\end{aligned} \tag{73}$$

Similarly, it is derived in (65) that $\mathbb{E}\left[\|\mathbf{e}_s^t\|^2_{\mathbf{K}}\right] \leq \alpha^2\sum_{i=1}^n\sigma_i^2 + \gamma^2\sigma_A^2\rho_{\max}\mathbb{E}\left[\|\mathbf{x}^t\|^2_{\mathbf{K}}\right]$. Thus,

$$\begin{aligned}&\mathbb{E}\left[\left\langle \mathbf{e}_s^t \mid \nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^{t+1}) - \nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^t)\right\rangle_{\mathbf{K}}\right]\\ &\leq \frac{1}{2}\mathbb{E}\left[\|\mathbf{e}_s^t\|^2_{\mathbf{K}}\right] + \frac{1}{2}\mathbb{E}\left[\|\nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^{t+1}) - \nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^t)\|^2_{\mathbf{K}}\right]\\ &\leq \alpha^2\sum_{i=1}^n\sigma_i^2 + \gamma^2\sigma_A^2\rho_{\max}\mathbb{E}\left[\|\mathbf{x}^t\|^2_{\mathbf{K}}\right] + \frac{1}{2}\mathbb{E}\left[\|\nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^{t+1}) - \nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^t)\|^2_{\mathbf{K}}\right].\end{aligned} \tag{74}$$

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

$$\begin{aligned}\mathbb{E}\left[\left\langle \mathbf{x}^{t+1} \mid \mathbf{v}^{t+1}\right\rangle_{\mathbf{K}}\right] \leq\ &\mathbb{E}\left[\left\langle \mathbf{x}^t \mid \mathbf{v}^t\right\rangle_{\mathbf{K}-(\gamma+\eta\beta)\mathbf{L}}\right] + \frac{\alpha-\eta}{2}\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\mathbf{K}}\right] + \frac{\alpha^3}{\eta}\sum_{i=1}^n\sigma_i^2\\ &+ \left\{\beta\rho_{\max} + \frac{\alpha}{2\eta} + \frac{3\alpha^2L^2}{2\eta} + \frac{\eta\beta^2\rho_{\max}^2}{2} + \frac{\alpha\gamma^2\sigma_A^2\rho_{\max}}{\eta}\right\}\mathbb{E}\left[\|\mathbf{x}^t\|^2_{\mathbf{K}}\right]\\ &+ (\frac{2\alpha}{\eta} + \frac{\alpha^2}{2\eta})\mathbb{E}\left[\|\nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^{t+1}) - \nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^t)\|^2\right].\end{aligned} \tag{75}$$

The proof is concluded by applying Lemma D.6 to bound the last term. □

D.5 Proof of Theorem D.5

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

$$F_{t+1} \leq F_t - \mathbb{C}_{\nabla F}\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^t)\|^2\right] + \mathbb{C}_\sigma\alpha^2\bar{\sigma}^2/n + \mathbb{E}\left[\mathbb{C}_{\mathbf{x}}\|\mathbf{x}^t\|^2_{\mathbf{K}} + \mathbb{C}_{\mathbf{v}}\|\mathbf{v}^t\|^2_{\mathbf{K}} + \left\langle \mathbf{x}^t \mid \mathbf{v}^t\right\rangle_{\mathbf{C}_{xv}}\right], \tag{76}$$

where

$$\mathbb{C}_{\nabla F} := \frac{\alpha}{4} - \mathtt{b}\frac{10\alpha^3L^2n}{\eta^2}(\rho_{\min}^{-1} + \mathtt{c}) - \mathtt{d}\frac{10\alpha^3L^2n}{\eta} \tag{77}$$

$$\mathbb{C}_\sigma := \frac{L}{2} + 2\mathtt{a}n^2 + \mathtt{b}\frac{5n\alpha L^2}{\eta^2}(\rho_{\min}^{-1} + \mathtt{c}) + \mathtt{d}\left(\frac{5n\alpha L^2}{\eta} + \frac{\alpha n^2}{\eta}\right) \tag{78}$$

$$\mathbb{C}_{\mathbf{x}} := \frac{3\alpha L^2}{4n} + \mathtt{a}\left(-\frac{\gamma}{2}\rho_{\min} + \alpha(1+3L^2)\right) + \mathtt{b}\left(2\beta^2\bar{\rho}_{\max}^2 + \frac{10\alpha^3L^4}{\eta^2}\right)(\rho_{\min}^{-1} + \mathtt{c})$$

$$+ \mathtt{d}\left(\beta\rho_{\max} + \frac{\alpha}{2\eta} + \frac{3\alpha^2 L^2}{2\eta} + \frac{\alpha\gamma^2\sigma_A^2\rho_{\max}}{\eta}\right) + \mathtt{d}\left(\frac{10\alpha^3 L^4}{\eta} + \frac{\eta\beta^2\rho_{\max}^2}{2}\right) \tag{79}$$

$$\mathbb{C}_{\mathbf{v}} := \mathtt{a}(2\eta^2) + \mathtt{b}(2\alpha(\rho_{\min}^{-1} + \mathtt{c})) + \mathtt{d}\frac{\alpha - \eta}{2} \tag{80}$$

$$\mathbf{C}_{xv} := (2\gamma\eta\mathtt{a} + 2\beta\mathtt{bc} - \mathtt{d}(\gamma + \eta\beta))\,\mathbf{A}^\top\mathbf{R}\mathbf{A} + (-2\eta\mathtt{a} + 2\beta\mathtt{b})\,\mathbf{K}. \tag{81}$$

The equality $\mathbf{C}_{xv} = \mathbf{0}$ is ensured by the parameter choices in (51). We then observe that for any $\delta_0 > 0$, it holds

$$F_t \geq F(\bar{\mathbf{x}}^t) + \left(\mathtt{a} - \frac{\mathtt{d}}{2\delta_0}\right)\|\mathbf{x}^t\|_{\mathbf{K}}^2 + \|\mathbf{v}^t\|^2_{\mathtt{b}\mathbf{Q} + (\mathtt{bc} - \frac{\mathtt{d}\delta_0}{2})\mathbf{K}} \tag{82}$$

$$\geq \frac{1}{n}\sum_{i=1}^n f_i(\bar{\mathbf{x}}^t) + \|\mathbf{v}^t\|^2_{(\mathtt{b}\rho_{\max}^{-1} + \mathtt{bc} - \frac{\mathtt{d}^2}{4\mathtt{a}})\mathbf{K}} \tag{83}$$

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

$$\mathtt{b}\rho_{\max}^{-1} + \mathtt{bc} - \frac{\mathtt{d}^2}{4\mathtt{a}} = (\frac{\eta\mathtt{a}}{\beta\rho_{\max}} - \frac{\mathtt{d}^2}{4\mathtt{a}}) + (\frac{\gamma\mathtt{d}}{2\beta} - \frac{\eta\gamma\mathtt{a}}{\beta}) + \frac{\eta\mathtt{d}}{2} \geq 0, \tag{84}$$

which is due to the step size choices

$$\mathtt{d} = \delta_1\eta\mathtt{a}, \quad \delta_1 \geq 8, \quad \eta\beta \leq \frac{4}{\delta_1^2\rho_{\max}}. \tag{85}$$

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

$$\begin{cases} \frac{2\alpha\eta}{\beta\rho_{\min}} \leq \frac{\eta^2}{3} & \Leftrightarrow \alpha \leq \frac{\eta\beta\rho_{\min}}{6}, \\ \frac{2\delta_1\alpha\eta\gamma}{\beta} \leq \frac{\eta^2}{3} & \Leftrightarrow \alpha \leq \frac{\eta\beta}{6\delta_1\gamma}, \\ \frac{\delta_1\alpha\eta}{2} \leq \frac{\eta^2}{3} & \Leftrightarrow \alpha \leq \frac{2\eta}{3\delta_1}. \end{cases} \tag{86}$$

We obtain that

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

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 have

$$\mathbb{C}_{\mathbf{x}} \leq \frac{3\alpha L^2}{4n} + \Big(-\frac{\gamma}{2}\rho_{\min} + \alpha(1 + 3L^2 + 4\delta_1) + (2\eta\beta\bar{\rho}_{\max}^2 + \frac{5\alpha}{2\eta\beta})(\rho_{\min}^{-1} + \delta_1\gamma) + \frac{3}{2}\delta_1\eta\beta\rho_{\max}\Big)\mathtt{a}.$$

With the step size conditions:

$$\begin{cases} \alpha(1 + 3L^2 + 4\delta_1) \leq \frac{\gamma}{16}\rho_{\min} & \Leftrightarrow \alpha \leq \frac{\gamma}{16(1+3L^2+4\delta_1)}\rho_{\min}, \\ 2\eta\beta\bar{\rho}_{\max}^2\rho_{\min}^{-1} \leq \frac{\gamma}{32}\rho_{\min} & \Leftrightarrow \eta\beta \leq \frac{\gamma}{64}\cdot\frac{\rho_{\min}^2}{\bar{\rho}_{\max}^2}, \\ 2\delta_1\gamma\eta\beta\bar{\rho}_{\max}^2\rho_{\min}^{-1} \leq \frac{\gamma}{32}\rho_{\min} & \Leftrightarrow \eta\beta \leq \frac{1}{64\delta_1}\cdot\frac{\rho_{\min}^2}{\bar{\rho}_{\max}^2}, \\ \frac{5\alpha}{2\eta\beta\rho_{\min}} \leq \frac{\gamma}{32}\rho_{\min} & \Leftrightarrow \alpha \leq \frac{\gamma\eta\beta}{80}\rho_{\min}^2, \\ \frac{5\delta_1\gamma\alpha}{2\eta\beta} \leq \frac{\gamma}{32}\rho_{\min} & \Leftrightarrow \alpha \leq \frac{\eta\beta}{80\delta_1}\rho_{\min}, \\ \frac{3}{2}\delta_1\eta\beta\rho_{\max} \leq \frac{\gamma}{16}\rho_{\min} & \Leftrightarrow \eta\beta \leq \frac{\gamma}{24\delta_1}\cdot\frac{\rho_{\min}}{\rho_{\max}}, \end{cases} \tag{87}$$

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

$$\mathbb{C}_{\mathbf{x}} \leq 3\alpha L^2/(4n) - \mathtt{a}\gamma\rho_{\min}/4 \leq -\mathtt{a}\gamma\rho_{\min}/8. \tag{88}$$

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

$$\begin{cases} \mathtt{a}\frac{10\alpha^3 L^2 n}{\eta\beta\rho_{\min}} \le \frac{\alpha}{24} & \Leftrightarrow \alpha \le \sqrt{\frac{\eta\beta\rho_{\min}}{240L^2 n\mathtt{a}}}, \\ \mathtt{a}\frac{10\delta_1\alpha^3\gamma L^2 n}{\eta\beta} \le \frac{\alpha}{24} & \Leftrightarrow \alpha \le \sqrt{\frac{\eta\beta}{240\delta_1\gamma L^2 n\mathtt{a}}}, \\ \mathtt{a}\cdot 10\delta_1\alpha^3 L^2 n \le \frac{\alpha}{24} & \Leftrightarrow \alpha \le \sqrt{\frac{1}{240\delta_1 L^2 n\mathtt{a}}}, \end{cases} \tag{89}$$

we have

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

Finally, the condition $\eta\beta \le \gamma$ guarantees:

$$\mathbb{C}_\sigma \le \frac{L}{2} + 2\mathtt{a}n^2 + \mathtt{a}n\alpha L^2\left(\frac{5}{\eta\beta\rho_{\min}} + \frac{5\delta_1\gamma}{\eta\beta} + 5\delta_1\right) + \mathtt{a}\delta_1 n^2\alpha. \tag{90}$$

This shows the desired properties regarding $\mathbb{C}_{\mathbf{x}}, \mathbb{C}_{\nabla F}, \mathbb{C}_\sigma, \mathbb{C}_{\mathbf{v}}$.

It remains to verify that the stepsize condition (52) indeed implies the restrictions above. Without loss of generality, we assume $\rho_{\max}, \bar\rho_{\max}, L \ge 1$. Since $\rho_{\max} \ge \rho_{\min}$, the definition of γ_∞ gives

$$\gamma_\infty \le \frac{\rho_{\min}}{\rho_{\max}^2}, \qquad \gamma_\infty \le \frac{\rho_{\min}}{2\sigma_A^2}, \qquad \gamma_\infty\rho_{\min} \le 1, \quad \gamma_\infty \le \frac{1}{\sigma_A\sqrt{\rho_{\max}}}.$$

Therefore all the γ-conditions used above hold. Next, by the definition of η_∞, and using $\rho_{\min} \le \rho_{\max}$, we have

$$\eta \le 1, \qquad \eta\beta = \eta \le \gamma, \qquad \eta\beta \le \rho_{\max}^{-1},$$

and

$$\eta\beta \le \frac{4}{\delta_1^2\rho_{\max}}, \qquad \eta\beta \le \frac{\gamma\rho_{\min}^2}{64\bar\rho_{\max}^2}, \qquad \eta\beta \le \frac{\gamma\rho_{\min}^2}{64\delta_1\bar\rho_{\max}^2}, \qquad \eta\beta \le \frac{\gamma\rho_{\min}}{24\delta_1\rho_{\max}}.$$

Thus the lower-boundedness condition in (85) and the η-dependent conditions in (87) are satisfied.

We now check the α-conditions. Observe that as $L \ge 1$, $\delta_1 \ge 8$, $n \ge 1$, and $\gamma\rho_{\min} \le 1$, we get

$$\alpha \le \frac{\gamma\rho_{\min}}{80\delta_1\sqrt{n}}\frac{1}{L^2}, \quad \alpha \le 1, \quad \alpha \le \frac{1}{4L}, \quad \alpha \le \frac{1}{2L^2}, \quad \alpha \le \frac{\gamma\rho_{\min}}{16(1+3L^2+4\delta_1)}.$$

The term $\mathtt{a}/L^2$ in α_∞ gives

$$\alpha \le \frac{\mathtt{a}\gamma n\rho_{\min}}{6L^2}.$$

The term $\eta\rho_{\min}$ in $\alpha_\infty(\eta)$ gives

$$\alpha \le \frac{\eta\rho_{\min}}{6}, \qquad \alpha \le \frac{\eta}{6\delta_1\gamma}, \qquad \alpha \le \frac{\gamma\eta\rho_{\min}^2}{80}, \qquad \alpha \le \frac{\eta\rho_{\min}}{80\delta_1}.$$

The term $\eta/(\gamma\rho_{\min})$ in $\alpha_\infty(\eta)$ gives

$$\alpha \le \frac{\eta}{80\delta_1\sqrt{n}} \le \frac{2\eta}{3\delta_1}.$$

Finally, the square-root term in $\alpha_\infty(\eta)$ gives

$$\alpha \le \sqrt{\frac{\eta\rho_{\min}}{240L^2 n\mathtt{a}}}, \qquad \alpha \le \sqrt{\frac{\eta}{240\delta_1\gamma L^2 n\mathtt{a}}}, \qquad \alpha \le \sqrt{\frac{1}{240\delta_1 L^2 n\mathtt{a}}}.$$

Consequently, all auxiliary conditions in (85), (86), (87), and (89) are implied by (52). This verifies that the stepsize conditions are jointly satisfiable and completes the proof of Theorem D.5. □

D.6 Proof of Corollary 3.7

The proof can be seen as a direct extension of (55) with Assumption 3.6. For instance, from (55) we get

$$\begin{aligned} F_{t+1} - f_\star &\leq F_t - f_\star - \frac{\alpha}{8}\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^t)\|^2\right] + \mathbb{C}_\sigma \alpha^2 \bar{\sigma}^2 - \frac{\mathtt{a}\gamma\rho_{\min}}{8}\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right] - \mathtt{a}\eta^2\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathbf{K}}^2\right] \\ &= (1-\delta)(F_t - f_\star) + \delta(\mathbb{E}\left[F(\bar{\mathbf{x}}^t)\right] - f_\star) - \frac{\alpha}{8}\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^t)\|^2\right] + \mathbb{C}_\sigma\alpha^2\bar{\sigma}^2 \\ &\quad + (\delta\mathtt{a} - \frac{\mathtt{a}\gamma\rho_{\min}}{8})\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right] + \delta\mathtt{b}\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathbf{Q}}^2\right] + (\delta\mathtt{bc} - \mathtt{a}\eta^2)\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathbf{K}}^2\right] \\ &\quad + \delta\mathtt{d}\mathbb{E}\left[\left\langle \mathbf{x}^t \mid \mathbf{v}^t\right\rangle_{\mathbf{K}}\right], \end{aligned} \tag{91}$$

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 us

$$\begin{aligned} F_{t+1} - f_\star &\leq (1-\delta)(F_t - f_\star) + (\frac{\delta}{2\mu} - \frac{\alpha}{8})\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^t)\|^2\right] + \mathbb{C}_\sigma\alpha^2\bar{\sigma}^2 \\ &\quad + \left[\delta(\mathtt{a} + \frac{\mathtt{d}}{2}) - \frac{\mathtt{a}\gamma\rho_{\min}}{8}\right]\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right] + \left[\delta(\frac{\mathtt{b}}{\rho_{\min}} + \mathtt{bc} + \frac{\mathtt{d}}{2}) - \mathtt{a}\eta^2\right]\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathbf{K}}^2\right]. \end{aligned} \tag{92}$$

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 coefficients of excessive terms to be non-positive, thus give rise to (21). □

D.7 Auxiliary Lemma

Lemma D.6. *Under Assumption 3.1 and 3.3,*

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

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

Proof of Lemma D.6. By the Lipschitz gradient assumption on each local objective function f_i,

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

The latter can be further expanded as

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

Lastly, we note that

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

This completes the proof. □

E Proof of Theorem 3.9

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

Notations. We introduce following shorthand notations to facilitate the analysis:

$$\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda};\xi) := \nabla\mathbf{f}(\mathbf{x};\xi) + \frac{\eta}{\alpha}\mathbf{A}^\top\boldsymbol{\lambda} + \frac{\gamma}{\alpha}\mathbf{A}^\top\mathbf{A}(\xi)\mathbf{x}, \tag{93}$$

$$\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda}) := \nabla\mathbf{f}(\mathbf{x}) + \frac{\eta}{\alpha}\mathbf{A}^\top\boldsymbol{\lambda} + \frac{\gamma}{\alpha}\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}, \tag{94}$$

$$\overline{\mathbf{m}}_x^t := \frac{1}{n}\mathbf{1}_\otimes^\top\mathbf{m}_x^t. \tag{95}$$

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

$$\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha\mathbf{m}_x^t, \tag{96}$$

$$\boldsymbol{\lambda}^{t+1} = \boldsymbol{\lambda}^t + \beta\mathbf{m}_\lambda^t, \tag{97}$$

$$\mathbf{m}_x^{t+1} = \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1};\xi^{t+1}) + (1-a_x)(\mathbf{m}_x^t - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t;\xi^{t+1})), \tag{98}$$

$$\mathbf{m}_\lambda^{t+1} = \mathbf{A}(\xi^{t+1})\mathbf{x}^{t+1} + (1-a_\lambda)(\mathbf{m}_\lambda^t - \mathbf{A}(\xi^{t+1})\mathbf{x}^t). \tag{99}$$

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

$$\mathbb{E}_t\left[F(\bar{\mathbf{x}}^{t+1})\right] \le F(\bar{\mathbf{x}}^t) - \frac{\alpha}{4}\left\|\overline{\mathbf{m}}_x^t\right\|^2 - \frac{\alpha}{2}\left\|\nabla F(\bar{\mathbf{x}}^t)\right\|^2 + \frac{\alpha L^2}{n}\|\mathbf{x}^t\|_{\mathbf{K}}^2 + \alpha\left\|\overline{\mathbf{m}}_x^t - \frac{1}{n}\mathbf{1}_\otimes^\top\nabla\mathbf{f}(\mathbf{x}^t)\right\|^2. \tag{100}$$

See Appendix E.1 for the proof.

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

$$\overline{\mathbf{m}}_x^{t+1} = \frac{1}{n}\mathbf{1}_\otimes^\top\nabla\mathbf{f}(\mathbf{x}^{t+1};\xi^{t+1}) + (1-a_x)(\overline{\mathbf{m}}_x^t - \frac{1}{n}\mathbf{1}_\otimes^\top\nabla\mathbf{f}(\mathbf{x}^t;\xi^{t+1})), \tag{101}$$

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

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

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

$$\mathbb{E}\left[\left\|\overline{\mathbf{m}}_x^{t+1} - \frac{1}{n}\mathbf{1}_\otimes^\top\nabla\mathbf{f}(\mathbf{x}^{t+1})\right\|^2\right] \le (1-a_x)^2\mathbb{E}\left[\left\|\overline{\mathbf{m}}_x^t - \frac{1}{n}\mathbf{1}_\otimes^\top\nabla\mathbf{f}(\mathbf{x}^t)\right\|^2\right] \tag{102}$$

$$+ \frac{8(L_{\mathsf{s}}^2+L^2)\alpha^2}{n}\mathbb{E}\left[\|\mathbf{m}_x^t - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\|^2\right] + 32(L_{\mathsf{s}}^2+L^2)\alpha^2\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^t)\|^2\right] \tag{103}$$

$$+ \frac{32(L_{\mathsf{s}}^2+L^2)\eta^2}{n}\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathbf{K}}^2\right] + \frac{32(L_{\mathsf{s}}^2+L^2)(\alpha^2L^2+\gamma^2\rho_{\max}^2)}{n}\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right] + 2a_x^2\bar{\sigma}^2.$$

See Appendix E.2 for the proof.

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

$$\mathbb{E}\left[\left\|\mathbf{m}_x^{t+1} - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1})\right\|^2\right] \le (1-\frac{a_x}{4})\mathbb{E}\left[\left\|\mathbf{m}_x^t - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\right\|^2\right] \tag{104}$$

$$+ 4a_x^2n\bar{\sigma}^2 + 64\alpha^2nG\cdot\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^t)\|^2\right] + \left(\frac{16a_x^2\eta^2\gamma^2\sigma_A^2}{\alpha^2} + 64\eta^2G\right)\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathbf{K}}^2\right] \tag{105}$$

$$+ \left(\frac{20a_x^2\gamma^2\sigma_A^2}{\alpha^2} + 64(\alpha^2L^2+\gamma^2\rho_{\max}^2)G\right)\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right], \tag{106}$$

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

See Appendix E.3 for the proof.

Lemma E.4. *Under Assumption 3.1, 3.2 and the step size condition* $\alpha \leq \gamma\rho_{\min}\sqrt{1-\gamma\rho_{\min}}/(\sqrt{32}L)$,

$$\mathbb{E}\left[\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2\right] \leq (1-\frac{\gamma\rho_{\min}}{8})\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right] + \frac{\eta^2}{(1-\gamma\rho_{\min})(1-\gamma\rho_{\min}/2)}\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathbf{K}}^2\right] \tag{107}$$

$$- \frac{2\eta(1-\gamma\rho_{\min}/4)}{(1-\gamma\rho_{\min})}\left\langle \mathbf{x}^t \mid \mathbf{v}^t\right\rangle_{\mathbf{K}-\gamma\mathbf{A}^\top\mathbf{R}\mathbf{A}} + \frac{2\alpha^2}{\gamma\rho_{\min}}\mathbb{E}\left[\|\mathbf{m}_x^t - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\|^2\right]. \tag{108}$$

See Appendix E.4 for the proof.

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

$$\mathbb{E}\left[\left\langle \mathbf{v}^{t+1} \mid \mathbf{x}^{t+1}\right\rangle_{\mathbf{K}}\right] \leq \mathbb{E}\left[\left\langle \mathbf{v}^t \mid \mathbf{x}^t\right\rangle_{\mathbf{K}-(\gamma+\eta\beta)\mathbf{A}^\top\mathbf{R}\mathbf{A}}\right] - \frac{\eta}{2}\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathbf{K}}^2\right] \tag{109}$$

$$+ \left(\frac{3\alpha^2L^2}{\eta} + 2\beta\rho_{\max}^2 - \gamma\beta\rho_{\min}^2 + \frac{5\beta}{2} + \frac{5\alpha^2}{2\eta}\right)\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right] \tag{110}$$

$$+ \left((\frac{3}{\eta}+\frac{1}{2})\alpha^2 + 2\alpha^2\beta + \frac{2\alpha^4}{\eta}\right)\mathbb{E}\left[\|\mathbf{m}_x^t - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\|^2\right] \tag{111}$$

$$+ \frac{\beta}{2}\mathbb{E}\left[\|\mathbf{A}^\top\mathbf{m}_\lambda^t - \mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t\|^2\right] + \frac{\alpha^2L^2n}{2\eta}\mathbb{E}\left[\|\overline{\mathbf{m}}_x^t\|^2\right]. \tag{112}$$

See Appendix E.5 for the proof.

Lemma E.6. *Under Assumption 3.1, 3.2 and the step size conditions* $\alpha \leq \eta, \beta \leq 1$, *for any constant* $\mathsf{b}, \mathsf{c} > 0$,

$$\mathbb{E}\left[\|\mathbf{v}^{t+1}\|_{\mathsf{b}\mathbf{Q}+\mathsf{c}\mathbf{K}}^2\right] \leq (1+\beta+\frac{\alpha}{\eta})\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathsf{b}\mathbf{Q}+\mathsf{c}\mathbf{K}}^2\right] + 2\beta\mathbb{E}\left[\left\langle \mathbf{v}^t \mid \mathbf{x}^t\right\rangle_{\mathsf{b}\mathbf{K}+\mathsf{c}\mathbf{A}^\top\mathbf{R}\mathbf{A}}\right] \tag{113}$$

$$+ 3\beta^2\rho_{\max}^2(\mathsf{b}\rho_{\min}^{-1}+\mathsf{c})\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right] + 4\beta(\mathsf{b}\rho_{\min}^{-1}+\mathsf{c})\mathbb{E}\left[\|\mathbf{A}^\top\mathbf{m}_\lambda^t - \mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t\|^2\right] \tag{114}$$

$$+ \frac{4n\alpha^3L^2}{\eta}(\mathsf{b}\rho_{\min}^{-1}+\mathsf{c})\mathbb{E}\left[\|\overline{\mathbf{m}}_x^t\|^2\right]. \tag{115}$$

See Appendix E.6 for the proof.

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

$$\mathbb{E}\left[\|\mathbf{A}^\top\mathbf{m}_\lambda^{t+1} - \mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^{t+1}\|^2\right] \leq (1-a_\lambda)^2\mathbb{E}\left[\|\mathbf{A}^\top\mathbf{m}_\lambda^t - \mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t\|^2\right] \tag{116}$$

$$+ \left(10a_\lambda^2\sigma_A^2 + 32(\rho_{\max}^2+\bar{\rho}_{\max}^2)(\alpha^2L^2+\gamma^2\rho_{\max}^2)\right)\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right] + 16\eta^2\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathbf{K}}^2\right] \tag{117}$$

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

See Appendix E.7 for the proof.[2] Below we observe the main convergence theorem.

[2]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 E.7 and applying $\mathbf{m}_\lambda^t = \mathbf{A}^\top\mathbf{A}(\xi^t)\mathbf{x}^t$ together with Lemma 3.4.

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

$$F_t = \mathbb{E}\Big[F(\bar{\mathbf{x}}^t) + \mathtt{a}\|\mathbf{x}^t\|^2_{\mathbf{K}} + \|\mathbf{v}^t\|^2_{\mathtt{b}\mathbf{Q}+\mathtt{c}\mathbf{K}} + \mathtt{d}\left\langle \mathbf{x}^t \;\middle|\; \mathbf{v}^t\right\rangle_{\mathbf{K}} + \mathtt{e}\|\overline{\mathbf{m}}^t_x - \frac{1}{n}\mathbf{1}^\top_\otimes \nabla \mathbf{f}(\mathbf{x}^t)\|^2 \tag{118}$$

$$+ \mathtt{f}\|\mathbf{m}^t_x - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t, \boldsymbol{\lambda}^t)\|^2 + \mathtt{g}\|\mathbf{A}^\top \mathbf{m}^t_\lambda - \mathbf{A}^\top \mathbf{R}\mathbf{A}\mathbf{x}^t\|^2\Big] \tag{119}$$

and the constants $R := \max\{1, L^2\}$, $S := \max\{1, L_s^2 + L^2\}$, $H := \min\{R^{-2}, (SR)^{-1/2}\}$. *Then, by the following choice of hyperparameters*

$$\mathtt{a} = n^{-1}, \quad \mathtt{b} = \mathtt{a}\cdot \frac{\eta(1-\gamma\rho_{\min}/4)}{\beta(1-\gamma\rho_{\min})} = \Theta(n^{-1}R^{-4}\alpha^{-2}), \tag{120}$$

$$\mathtt{c} = \frac{\mathtt{d}}{2}\cdot\left(\frac{\gamma}{\beta}+\eta\right) - \mathtt{a}\cdot\frac{\eta\gamma}{\beta}\cdot\frac{1-\gamma\rho_{\min}/4}{1-\gamma\rho_{\min}} = \Theta(n^{-1}R^{-4}\alpha^{-2}), \tag{121}$$

$$\mathtt{d} = \Theta(n^{-1}R^{-2}\alpha^{-1}), \quad \mathtt{e} = \Theta(S^{-1}R^{-1}\alpha^{-1}), \tag{122}$$

$$\mathtt{f} = \Theta(n^{-1}S^{-1}R^{-2}\alpha^{-1}), \quad \mathtt{g} = \Theta(n^{-1}R^{-3}\alpha^{-1}), \tag{123}$$

$$\alpha = \min\{\Theta(S^{-1/3}R^{-1/3}\bar{\sigma}^{-2/3}T^{-1/3}), \Theta(H)\}, \quad \eta = \Theta(R^{-1}), \tag{124}$$

$$\gamma = \Theta(R\alpha), \quad \beta = \Theta(\frac{\mathtt{a}}{\mathtt{d}}\cdot\gamma) = \Theta(R^3\alpha^2), \tag{125}$$

$$a_x = \Theta(SR\alpha^2), \quad a_\lambda = \Theta(R^2\alpha), \tag{126}$$

the potential function follows the inequality

$$F_{t+1} \le F_t - \frac{\alpha}{4}\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^t)\|^2\right] - \frac{\alpha}{8}\mathbb{E}\left[\|\overline{\mathbf{m}}^t_x\|^2\right] + (\mathtt{e}\cdot 2a_x^2 + \mathtt{f}\cdot 4a_x^2 n)\bar{\sigma}^2 \tag{127}$$

$$- \mathtt{a}\cdot\frac{\gamma\rho_{\min}}{16}\mathbb{E}\left[\|\mathbf{x}^t\|^2_{\mathbf{K}}\right] - \mathtt{d}\cdot\frac{\eta}{4}\mathbb{E}\left[\|\mathbf{v}^t\|^2_{\mathbf{K}}\right] - \mathtt{e}\cdot\frac{a_x}{2}\mathbb{E}\left[\|\overline{\mathbf{m}}^t_x - \frac{1}{n}\mathbf{1}^\top_\otimes\nabla\mathbf{f}(\mathbf{x}^t)\|^2\right] \tag{128}$$

$$- \mathtt{f}\cdot\frac{a_x}{8}\mathbb{E}\left[\|\mathbf{m}^t_x - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\|^2\right] - \mathtt{g}\cdot\frac{a_\lambda}{2}\mathbb{E}\left[\|\mathbf{A}^\top\mathbf{m}^t_\lambda - \mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t\|^2\right]. \tag{129}$$

See Appendix E.8 for the proof. We remark that the hyperparameter and stepsize conditions in the theorem can be relaxed in practice.

Summing up (129) from $t = 0$ to $t = T-1$ give us the convergence bound for the network average iterate and consensus error as

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^t)\|^2 + \frac{\mathtt{a}\gamma\rho_{\min}}{4\alpha}\|\mathbf{x}^t\|^2_{\mathbf{K}}\right] &\le \frac{F_0 - F_T}{T\alpha/4} + \frac{(\mathtt{e}\cdot 2a_x^2 + \mathtt{f}\cdot 4a_x^2 n)\bar{\sigma}^2}{\alpha/4} \\ &= \mathcal{O}\left(\frac{F_0 - F_T + \bar{\sigma}^{2/3}}{T^{2/3}}\right).\end{aligned} \tag{130}$$

E.1 Proof of Lemma E.1

By Assumption 3.1,

$$\mathbb{E}_t[F(\bar{\mathbf{x}}^{t+1})] \tag{131}$$

$$\le F(\bar{\mathbf{x}}^t) - \alpha\left\langle \nabla F(\bar{\mathbf{x}}^t) \;\middle|\; \overline{\mathbf{m}}^t_x\right\rangle + \frac{\alpha^2 L}{2}\|\overline{\mathbf{m}}^t_x\|^2 \tag{132}$$

$$= F(\bar{\mathbf{x}}^t) - \left(\frac{\alpha}{2} + \frac{\alpha^2 L}{2}\right)\|\overline{\mathbf{m}}^t_x\|^2 - \frac{\alpha}{2}\|\nabla F(\bar{\mathbf{x}}^t)\|^2 + \frac{\alpha}{2}\|\overline{\mathbf{m}}^t_x - \nabla F(\bar{\mathbf{x}}^t)\|^2 \tag{133}$$

$$\overset{(i)}{\le} F(\bar{\mathbf{x}}^t) - \frac{\alpha}{4}\|\overline{\mathbf{m}}^t_x\|^2 - \frac{\alpha}{2}\|\nabla F(\bar{\mathbf{x}}^t)\|^2 + \alpha\left\|\overline{\mathbf{m}}^t_x - \frac{1}{n}\mathbf{1}^\top_\otimes\nabla\mathbf{f}(\mathbf{x}^t)\right\|^2 + \alpha\left\|\frac{1}{n}\mathbf{1}^\top_\otimes\nabla\mathbf{f}(\mathbf{x}^t) - \nabla F(\bar{\mathbf{x}}^t)\right\|^2,$$

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

E.2 Proof of Lemma E.2

By the network average momentum update in (101),

$$\mathbb{E}\left[\left\|\overline{\mathbf{m}}_x^{t+1} - \frac{1}{n}\mathbf{1}_{\otimes}^{\top}\nabla\mathbf{f}(\mathbf{x}^{t+1})\right\|^2\right] \tag{134}$$

$$= \mathbb{E}\big[\big\|(1-a_x)(\overline{\mathbf{m}}_x^{t} - \frac{1}{n}\mathbf{1}_{\otimes}^{\top}\nabla\mathbf{f}(\mathbf{x}^{t})) + a_x\frac{1}{n}\mathbf{1}_{\otimes}^{\top}(\nabla\mathbf{f}(\mathbf{x}^{t+1};\xi^{t+1}) - \nabla\mathbf{f}(\mathbf{x}^{t+1})) \tag{135}$$

$$+ (1-a_x)\frac{1}{n}\mathbf{1}_{\otimes}^{\top}(\nabla\mathbf{f}(\mathbf{x}^{t}) - \nabla\mathbf{f}(\mathbf{x}^{t};\xi^{t+1}) - (\nabla\mathbf{f}(\mathbf{x}^{t+1}) - \nabla\mathbf{f}(\mathbf{x}^{t+1};\xi^{t+1}))\big\|^2\big] \tag{136}$$

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

$$+ 2(1-a_x)^2\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{1}_{\otimes}^{\top}(\nabla\mathbf{f}(\mathbf{x}^{t}) - \nabla\mathbf{f}(\mathbf{x}^{t+1})) - (\nabla\mathbf{f}(\mathbf{x}^{t};\xi^{t+1}) - \nabla\mathbf{f}(\mathbf{x}^{t+1};\xi^{t+1}))\right\|^2\right]. \tag{137}$$

Now observe that

$$2a_x^2\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{1}_{\otimes}^{\top}(\nabla\mathbf{f}(\mathbf{x}^{t+1};\xi^{t+1}) - \nabla\mathbf{f}(\mathbf{x}^{t+1}))\right\|^2\right] \leq 2a_x^2\bar{\sigma}^2, \tag{138}$$

and

$$2(1-a_x)^2\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{1}_{\otimes}^{\top}(\nabla\mathbf{f}(\mathbf{x}^{t}) - \nabla\mathbf{f}(\mathbf{x}^{t+1})) - (\nabla\mathbf{f}(\mathbf{x}^{t};\xi^{t+1}) - \nabla\mathbf{f}(\mathbf{x}^{t+1};\xi^{t+1}))\right\|^2\right] \tag{139}$$

$$\leq 4(1-a_x)^2\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{1}_{\otimes}^{\top}(\nabla\mathbf{f}(\mathbf{x}^{t}) - \nabla\mathbf{f}(\mathbf{x}^{t+1}))\right\|^2\right] \tag{140}$$

$$+ 4(1-a_x)^2\mathbb{E}\left[\left\|\frac{1}{n}\mathbf{1}_{\otimes}^{\top}(\nabla\mathbf{f}(\mathbf{x}^{t};\xi^{t+1}) - \nabla\mathbf{f}(\mathbf{x}^{t+1};\xi^{t+1}))\right\|^2\right] \tag{141}$$

$$\leq \frac{4(1-a_x)^2}{n}(L^2 + L_{\mathsf{s}}^2)\mathbb{E}\left[\|\mathbf{x}^{t+1} - \mathbf{x}^{t}\|^2\right]. \tag{142}$$

Finally, applying Lemma E.10 and utilizing the step size condition $0 < a_x \leq 1$ to simplify the coefficients will complete the proof. □

E.3 Proof of Lemma E.3

$$\mathbb{E}\left[\left\|\mathbf{m}_x^{t+1} - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1})\right\|^2\right] \tag{143}$$

$$= \mathbb{E}\big[\big\|(1-a_x)\left(\mathbf{m}_x^{t} - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t},\boldsymbol{\lambda}^{t})\right) + a_x\left(\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1};\xi^{t+1}) - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1})\right) \tag{144}$$

$$+ (1-a_x)\left(\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t},\boldsymbol{\lambda}^{t}) - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t},\boldsymbol{\lambda}^{t},\xi^{t+1})\right) \tag{145}$$

$$+ (1-a_x)\left(\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1};\xi^{t+1}) - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1})\right)\|^2\big] \tag{146}$$

$$\leq (1-a_x)^2\mathbb{E}\left[\|\mathbf{m}_x^{t} - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t},\boldsymbol{\lambda}^{t})\|^2\right] \tag{147}$$

$$+ 2a_x^2\mathbb{E}\left[\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1};\xi^{t+1}) - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1})\|^2\right] \tag{148}$$

$$+ 2(1-a_x)^2\mathbb{E}\big[\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1};\xi^{t+1}) - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t},\boldsymbol{\lambda}^{t},\xi^{t+1}) \tag{149}$$

$$- (\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1}) - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t},\boldsymbol{\lambda}^{t}))\|^2\big]. \tag{150}$$

Now observe that

$$2a_x^2\mathbb{E}\left[\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1};\xi^{t+1}) - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^{t+1},\boldsymbol{\lambda}^{t+1})\|^2\right] \tag{151}$$

$$= 2a_x^2\mathbb{E}\left[\left\|\nabla\mathbf{f}(\mathbf{x}^{t+1};\xi^{t+1}) - \nabla\mathbf{f}(\mathbf{x}^{t+1}) + \frac{\gamma}{\alpha}\mathbf{A}^{\top}\mathbf{A}(\xi^{t+1})\mathbf{x}^{t+1} - \frac{\gamma}{\alpha}\mathbf{A}^{\top}\mathbf{R}\mathbf{A}\mathbf{x}^{t+1}\right\|^2\right] \tag{152}$$

$$\leq 4a_x^2 n\bar{\sigma}^2 + \frac{4a_x^2\gamma^2\sigma_A^2}{\alpha^2}\mathbb{E}\left[\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2\right] \tag{153}$$

$$\leq 4a_x^2 n\bar{\sigma}^2 + \frac{20a_x^2\gamma^2\sigma_A^2}{\alpha^2}\mathbb{E}\left[\|\mathbf{x}^{t}\|_{\mathbf{K}}^2\right] + \frac{16a_x^2\eta^2\gamma^2\sigma_A^2}{\alpha^2}\mathbb{E}\left[\|\mathbf{v}^{t}\|_{\mathbf{K}}^2\right] \tag{154}$$

$$+ 16a_x^2\gamma^2\sigma_A^2\mathbb{E}\left[\|\mathbf{m}_x^t - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\|^2\right], \tag{155}$$

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

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

$$= \mathbb{E}\big[\|\nabla\mathbf{f}(\mathbf{x}^{t+1};\xi^{t+1}) - \nabla\mathbf{f}(\mathbf{x}^t;\xi^{t+1}) - (\nabla\mathbf{f}(\mathbf{x}^{t+1}) - \nabla\mathbf{f}(\mathbf{x}^t)) \tag{156}$$

$$+ \frac{\gamma}{\alpha}(\mathbf{A}^\top\mathbf{A}(\xi^{t+1})(\mathbf{x}^{t+1}-\mathbf{x}^t) - \mathbf{A}^\top\mathbf{R}\mathbf{A}(\mathbf{x}^{t+1}-\mathbf{x}^t))\|^2\big] \tag{157}$$

$$\leq 4\left(L_{\mathsf{s}}^2 + L^2 + \frac{\gamma^2}{\alpha^2}\bar{\rho}_{\max}^2 + \frac{\gamma^2}{\alpha^2}\rho_{\max}^2\right)\mathbb{E}\left[\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2\right]. \tag{158}$$

Finally, applying Lemma E.10 and utilizing the step size conditions $\alpha \leq \sqrt{a_x/(64(L^2+L_{\mathsf{s}}^2))}$, $\gamma \leq \sqrt{a_x/(64(\rho_{\max}^2+\bar{\rho}_{\max}^2))}$, $0 < a_x \leq 1$, $\gamma \leq 1/(8\sqrt{a_x}\sigma_A^2)$ to simplify the coefficients will complete the proof. □

E.4 Proof of Lemma E.4

$$\mathbb{E}\left[\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2\right] = \mathbb{E}\left[\|\mathbf{x}^t - \alpha\mathbf{m}_x^t\|_{\mathbf{K}}^2\right] \tag{159}$$

$$= \mathbb{E}\left[\|\mathbf{x}^t - \alpha\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t) - \alpha(\mathbf{m}_x^t - \nabla_x\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t))\|_{\mathbf{K}}^2\right] \tag{160}$$

$$\leq (1+z_1)\mathbb{E}\left[\|\mathbf{x}^t - \alpha\nabla\mathbf{f}(\mathbf{x}^t) - \eta\mathbf{A}^\top\boldsymbol{\lambda}^t - \gamma\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t\|_{\mathbf{K}}^2\right] \tag{161}$$

$$+ (1+z_1^{-1})\alpha^2\mathbb{E}\left[\|\mathbf{m}_x^t - \nabla_x\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\|^2\right] \quad \text{for any } z_1 > 0. \tag{162}$$

Now observe that

$$\mathbb{E}\left[\|\mathbf{x}^t - \alpha\nabla\mathbf{f}(\mathbf{x}^t) - \eta\mathbf{A}^\top\boldsymbol{\lambda}^t - \gamma\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t\|_{\mathbf{K}}^2\right] \tag{163}$$

$$= \mathbb{E}\left[\|(\mathbf{I}-\gamma\mathbf{A}^\top\mathbf{R}\mathbf{A})\mathbf{x}^t - \eta\mathbf{v}^t - \alpha(\nabla\mathbf{f}(\mathbf{x}^t) - \nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^t))\|_{\mathbf{K}}^2\right] \tag{164}$$

$$\leq (1+z_2)\mathbb{E}\left[\|(\mathbf{I}-\gamma\mathbf{A}^\top\mathbf{R}\mathbf{A})\mathbf{x}^t - \eta\mathbf{v}^t\|_{\mathbf{K}}^2\right] \tag{165}$$

$$+ (1+z_2^{-1})\alpha^2\mathbb{E}\left[\|\nabla\mathbf{f}(\mathbf{x}^t) - \nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^t)\|^2\right] \quad \text{for any } z_2 > 0, \tag{166}$$

$$\leq (1+z_2)(1-\gamma\rho_{\min})\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right] + (1+z_2)\eta^2\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathbf{K}}^2\right] \tag{167}$$

$$- 2(1+z_2)\eta\left\langle\mathbf{x}^t \,\middle|\, \mathbf{v}^t\right\rangle_{\mathbf{K}-\gamma\mathbf{A}^\top\mathbf{R}\mathbf{A}} + (1+z_2^{-1})\alpha^2L^2\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right]. \tag{168}$$

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 $z_2 = \frac{\gamma\rho_{\min}/4}{1-\gamma\rho_{\min}/2}$ so that $(1+z_1)(1+z_2)(1-\gamma\rho_{\min}) = 1-\gamma\rho_{\min}/4$. Finally, observing that

$$(1+z_1)(1+z_2^{-1})\alpha^2L^2 \leq \frac{4\alpha^2L^2}{\gamma\rho_{\min}(1-\gamma\rho_{\min})} \leq \frac{\gamma\rho_{\min}}{8} \tag{169}$$

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

E.5 Proof of Lemma E.5

By the update rules, we have

$$\mathbf{v}^{t+1} = \mathbf{v}^t + \beta\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t + \beta\mathbf{A}^\top(\mathbf{m}_\lambda^t - \mathbf{R}\mathbf{A}\mathbf{x}^t) + \frac{\alpha}{\eta}(\nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^{t+1}) - \nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^t)), \tag{170}$$

$$\mathbf{x}^{t+1} = (1-\gamma\mathbf{A}^\top\mathbf{R}\mathbf{A})\mathbf{x}^t - \eta\mathbf{v}^t - \alpha(\nabla\mathbf{f}(\mathbf{x}^t) - \nabla\mathbf{f}(\mathbf{1}_{\otimes}\bar{\mathbf{x}}^t)) - \alpha(\mathbf{m}_x^t - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)). \tag{171}$$

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

$$\mathbb{E}\left[\left\langle\mathbf{v}^{t+1} \,\middle|\, \mathbf{x}^{t+1}\right\rangle_{\mathbf{K}}\right] \leq \mathbb{E}\left[\left\langle\mathbf{v}^t \,\middle|\, \mathbf{x}^t\right\rangle_{\mathbf{K}-(\gamma+\eta\beta)\mathbf{A}^\top\mathbf{R}\mathbf{A}}\right] + (-1 + \frac{z_1}{2} + \frac{z_2}{2})\eta\cdot\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathbf{K}}^2\right] \tag{172}$$

$$+\left(\frac{\alpha^2L^2}{2z_1\eta}+\beta\rho_{\max}^2+\frac{\beta^2}{2}\rho_{\max}^2+\frac{\alpha^2L^2}{2}-\gamma\beta\rho_{\min}^2+\frac{\beta^2}{2}\rho_{\max}^2\right)\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right] \tag{173}$$

$$+\left(\frac{\alpha^2}{2z_2\eta}+\frac{\alpha^2}{2}\right)\mathbb{E}\left[\|\mathbf{m}_x^t-\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\|^2\right]+\frac{\beta}{2}\mathbb{E}\left[\|\mathbf{A}^\top\mathbf{m}_\lambda^t-\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t\|_{\mathbf{K}}^2\right] \tag{174}$$

$$+\left(\frac{\beta}{2}+\frac{\alpha^2}{2\eta}\right)\mathbb{E}\left[\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2\right]+\frac{L^2n}{2\eta}\mathbb{E}\left[\|\bar{\mathbf{x}}^{t+1}-\bar{\mathbf{x}}^t\|^2\right]. \tag{175}$$

Note that by (171), we have

$$\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2\le 4(1-\gamma\rho_{\min})^2\|\mathbf{x}^t\|_{\mathbf{K}}^2+4\eta^2\|\mathbf{v}^t\|_{\mathbf{K}}^2+4\alpha^2L^2\|\mathbf{x}^t\|_{\mathbf{K}}^2+4\alpha^2\|\mathbf{m}_x^t-\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\|^2. \tag{176}$$

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

E.6 Proof of Lemma E.6

By the dual update rule, we have

$$\mathbb{E}\left[\|\mathbf{v}^{t+1}\|_{\mathtt{b}\mathbf{Q}+\mathtt{c}\mathbf{K}}^2\right] \tag{177}$$

$$=\mathbb{E}\left[\|\mathbf{v}^t+\beta\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t+\beta\mathbf{A}^\top(\mathbf{m}_\lambda^t-\mathbf{R}\mathbf{A}\mathbf{x}^t)+\frac{\alpha}{\eta}(\nabla\mathbf{f}(\mathbf{1}_\otimes\bar{\mathbf{x}}^{t+1})-\nabla\mathbf{f}(\mathbf{1}_\otimes\bar{\mathbf{x}}^t))\|_{\mathtt{b}\mathbf{Q}+\mathtt{c}\mathbf{K}}^2\right]$$

$$\le\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathtt{b}\mathbf{Q}+\mathtt{c}\mathbf{K}}^2\right]+\mathbb{E}\left[\|\mathbf{v}^{t+1}-\mathbf{v}^t\|_{\mathtt{b}\mathbf{Q}+\mathtt{c}\mathbf{K}}^2\right] \tag{178}$$

$$+2\mathbb{E}\left[\left\langle\mathbf{v}^t\,\middle|\,\beta\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t+\beta\mathbf{A}^\top(\mathbf{m}_\lambda^t-\mathbf{R}\mathbf{A}\mathbf{x}^t)+\frac{\alpha}{\eta}(\nabla\mathbf{f}(\mathbf{1}_\otimes\bar{\mathbf{x}}^{t+1})-\nabla\mathbf{f}(\mathbf{1}_\otimes\bar{\mathbf{x}}^t))\right\rangle_{\mathtt{b}\mathbf{Q}+\mathtt{c}\mathbf{K}}\right]$$

$$\le(1+\beta+\frac{\alpha}{\eta})\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathtt{b}\mathbf{Q}+\mathtt{c}\mathbf{K}}^2\right]+2\beta\mathbb{E}\left[\left\langle\mathbf{v}^t\,\middle|\,\mathbf{x}^t\right\rangle_{\mathtt{b}\mathbf{K}+\mathtt{c}\mathbf{A}^\top\mathbf{R}\mathbf{A}}\right] \tag{179}$$

$$+(\frac{\alpha}{\eta}+\frac{3\alpha^2}{\eta^2})\mathbb{E}\left[\|\nabla\mathbf{f}(\mathbf{1}_\otimes\bar{\mathbf{x}}^{t+1})-\nabla\mathbf{f}(\mathbf{1}_\otimes\bar{\mathbf{x}}^t)\|_{\mathtt{b}\mathbf{Q}+\mathtt{c}\mathbf{K}}^2\right]+3\beta^2\rho_{\max}^2\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathtt{b}\mathbf{Q}+\mathtt{c}\mathbf{K}}^2\right] \tag{180}$$

$$+(\beta+3\beta^2)\mathbb{E}\left[\|\mathbf{A}^\top\mathbf{m}_\lambda^t-\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t\|_{\mathtt{b}\mathbf{Q}+\mathtt{c}\mathbf{K}}^2\right] \tag{181}$$

$$\le(1+\beta+\frac{\alpha}{\eta})\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathtt{b}\mathbf{Q}+\mathtt{c}\mathbf{K}}^2\right]+2\beta\mathbb{E}\left[\left\langle\mathbf{v}^t\,\middle|\,\mathbf{x}^t\right\rangle_{\mathtt{b}\mathbf{K}+\mathtt{c}\mathbf{A}^\top\mathbf{R}\mathbf{A}}\right] \tag{182}$$

$$+\frac{4n\alpha L^2}{\eta}(\mathtt{b}\rho_{\min}^{-1}+\mathtt{c})\mathbb{E}\left[\|\bar{\mathbf{x}}^{t+1}-\bar{\mathbf{x}}^t\|^2\right]+3\beta^2\rho_{\max}^2(\mathtt{b}\rho_{\min}^{-1}+\mathtt{c})\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right] \tag{183}$$

$$+4\beta(\mathtt{b}\rho_{\min}^{-1}+\mathtt{c})\mathbb{E}\left[\|\mathbf{A}^\top\mathbf{m}_\lambda^t-\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t\|^2\right]. \tag{184}$$

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

E.7 Proof of Lemma E.7

$$\mathbb{E}\left[\|\mathbf{A}^\top\mathbf{m}_\lambda^{t+1}-\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^{t+1}\|^2\right] \tag{185}$$

$$=\mathbb{E}\big[\|(1-a_\lambda)(\mathbf{A}^\top\mathbf{m}_\lambda^t-\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t)+a_\lambda(\mathbf{A}^\top\mathbf{A}(\xi^{t+1})\mathbf{x}^{t+1}-\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^{t+1}) \tag{186}$$

$$+(1-a_\lambda)(\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t-\mathbf{A}^\top\mathbf{A}(\xi^{t+1})\mathbf{x}^t-(\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^{t+1}-\mathbf{A}^\top\mathbf{A}(\xi^{t+1})\mathbf{x}^{t+1}))\|^2\big]$$

$$\le(1-a_\lambda)^2\mathbb{E}\left[\|\mathbf{A}^\top\mathbf{m}_\lambda^t-\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t\|^2\right]+2a_\lambda^2\mathbb{E}\left[\|\mathbf{A}^\top\mathbf{A}(\xi^{t+1})\mathbf{x}^{t+1}-\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^{t+1}\|^2\right] \tag{187}$$

$$+2(1-a_\lambda)^2\mathbb{E}\left[\|\mathbf{A}^\top\mathbf{R}\mathbf{A}(\mathbf{x}^t-\mathbf{x}^{t+1})-\mathbf{A}^\top\mathbf{A}(\xi^{t+1})(\mathbf{x}^t-\mathbf{x}^{t+1})\|^2\right] \tag{188}$$

$$\le(1-a_\lambda)^2\mathbb{E}\left[\|\mathbf{A}^\top\mathbf{m}_\lambda^t-\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t\|^2\right]+2a_\lambda^2\sigma_A^2\mathbb{E}\left[\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2\right] \tag{189}$$

$$+4(1-a_\lambda)^2(\rho_{\max}^2+\bar{\rho}_{\max}^2)\mathbb{E}\left[\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2\right]. \tag{190}$$

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

$$\|\mathbf{x}^{t+1}\|_{\mathbf{K}}^2\le 4(1-\gamma\rho_{\min})^2\|\mathbf{x}^t\|_{\mathbf{K}}^2+4\eta^2\|\mathbf{v}^t\|_{\mathbf{K}}^2+4\alpha^2L^2\|\mathbf{x}^t\|_{\mathbf{K}}^2+4\alpha^2\|\mathbf{m}_x^t-\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\|^2. \tag{191}$$

Finally, applying Lemma E.10 and the step size condition $a_\lambda \leq \sigma_A^{-1}$ to simplify the coefficients will complete the proof. □

E.8 Proof of Theorem E.8

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

$$\alpha \leq \min\left\{\frac{1}{2L}, \sqrt{\frac{a_x}{64(L^2+L_{\mathsf{s}}^2)}}, \frac{\gamma\rho_{\min}\sqrt{1-\gamma\rho_{\min}}}{\sqrt{32}L}, \frac{1}{\sqrt{12}}, \frac{1}{2L}, \eta\right\}, \tag{192}$$

$$\gamma \leq \min\left\{\sqrt{\frac{a_x}{64(\rho_{\max}^2+\bar{\rho}_{\max}^2)}}, \frac{1}{8\sqrt{a_x}\sigma_A^2}\right\}, \tag{193}$$

$$\eta\beta \leq \frac{1}{12}, \quad \beta \leq 1, \quad 0 < a_x \leq 1, \quad a_\lambda \leq \frac{1}{\sigma_A}, \tag{194}$$

we obtain

$$F_{t+1} \leq F_t + \mathbb{C}_{\nabla F}\mathbb{E}\left[\|\nabla F(\bar{\mathbf{x}}^t)\|^2\right] + \mathbb{C}_\sigma\bar{\sigma}^2 + \mathbb{C}_{\mathbf{x}}\mathbb{E}\left[\|\mathbf{x}^t\|_{\mathbf{K}}^2\right] + \mathbb{C}_{\mathbf{v}}\mathbb{E}\left[\|\mathbf{v}^t\|_{\mathbf{K}}^2\right] \tag{195}$$

$$+ \left\langle \mathbf{x}^t \;\middle|\; \mathbf{v}^t \right\rangle_{\mathbf{C}_{xv}} + \mathbb{C}_{\Delta\overline{\mathbf{m}}_x}\mathbb{E}\left[\left\|\overline{\mathbf{m}}_x^t - \frac{1}{n}\mathbf{1}_\otimes^\top\nabla\mathbf{f}(\mathbf{x}^t)\right\|^2\right] \tag{196}$$

$$+ \mathbb{C}_{\Delta\mathbf{m}_x}\mathbb{E}\left[\|\mathbf{m}_x^t - \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\|^2\right] + \mathbb{C}_{\Delta\mathbf{m}_\lambda}\mathbb{E}\left[\|\mathbf{A}^\top\mathbf{m}_\lambda^t - \mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t\|^2\right] \tag{197}$$

$$+ \mathbb{C}_{\overline{\mathbf{m}}_x}\mathbb{E}\left[\|\overline{\mathbf{m}}_x^t\|^2\right], \tag{198}$$

where

$$\mathbb{C}_{\nabla F} = -\frac{\alpha}{2} + \mathsf{e}\cdot 32(L_{\mathsf{s}}^2+L^2)\alpha^2 + \mathsf{f}\cdot 64\alpha^2 nG + \mathsf{g}\cdot 32\alpha^2 n(\rho_{\max}^2+\bar{\rho}_{\max}^2) \tag{199}$$

$$\leq -\frac{\alpha}{4} \quad \text{when} \quad \begin{cases} \alpha \leq \min\left\{\mathsf{e}^{-1}\cdot\frac{1}{512(L_{\mathsf{s}}^2+L^2)}, \quad \mathsf{g}^{-1}\cdot\frac{1}{512n(\rho_{\max}^2+\bar{\rho}_{\max}^2)}\right\}, \\ \mathsf{f}\cdot 64\alpha^2 nG \leq \frac{\alpha}{8} \Leftarrow \begin{cases} \alpha \leq \mathsf{f}^{-1}\cdot\frac{1}{1024n(L_{\mathsf{s}}^2+L^2)}, \\ \gamma^2 \leq \mathsf{f}^{-1}\cdot\frac{\alpha}{1024n(\rho_{\max}^2+\bar{\rho}_{\max}^2)}, \end{cases} \end{cases} \tag{200}$$

$$\mathbb{C}_\sigma = \mathsf{e}\cdot 2a_x^2 + \mathsf{f}\cdot 4a_x^2 n, \tag{201}$$

$$\mathbb{C}_{\mathbf{x}} = -\mathsf{a}\cdot\frac{\gamma\rho_{\min}}{8} + \frac{\alpha L^2}{n} + 3\beta^2\rho_{\max}^2(\mathsf{b}\rho_{\min}^{-1} + \mathsf{c}) \tag{202}$$

$$+ \mathsf{d}\cdot\left(\frac{3\alpha^2L^2}{\eta} + 2\beta\rho_{\max}^2 - \gamma\beta\rho_{\min}^2 + \frac{5\beta}{2} + \frac{5\alpha^2}{2\eta}\right) \tag{203}$$

$$+ \mathsf{e}\cdot 32(L_{\mathsf{s}}^2+L^2)(\alpha^2L^2+\gamma^2\rho_{\max}^2)/n \tag{204}$$

$$+ \mathsf{f}\cdot\left(\frac{20a_x^2\gamma^2\sigma_A^2}{\alpha^2} + 64(\alpha^2L^2+\gamma^2\rho_{\max}^2)G\right) \tag{205}$$

$$+ \mathsf{g}\cdot\left(10a_\lambda^2\sigma_A^2 + 32(\rho_{\max}^2+\bar{\rho}_{\max}^2)(\alpha^2L^2+\gamma^2\rho_{\max}^2)\right) \tag{206}$$

$$\leq -\mathtt{a}\cdot\frac{\gamma\rho_{\min}}{16} \quad \text{when} \begin{cases} \alpha \leq \mathtt{a}\cdot\frac{\gamma\rho_{\min}n}{256L^2}, \\ \alpha^2 \leq \min\left\{\frac{\mathtt{a}}{\mathtt{d}}\cdot\frac{\eta\gamma\rho_{\min}}{768L^2},\quad \frac{\mathtt{a}}{\mathtt{d}}\cdot\frac{\eta\gamma\rho_{\min}}{640}\right\}, \\ \beta^2 \leq \min\left\{\frac{\mathtt{a}}{\mathtt{b}}\cdot\frac{\gamma\rho_{\min}^2}{1536\rho_{\max}^2},\quad \frac{\mathtt{a}}{\mathtt{c}}\cdot\frac{\gamma\rho_{\min}}{1536\rho_{\max}^2}\right\}, \\ \alpha^3 \leq \frac{\mathtt{a}}{\mathtt{d}}\cdot\frac{\eta\gamma\rho_{\min}}{768L^4}, \\ \beta \leq \min\left\{\frac{\mathtt{a}}{\mathtt{d}}\cdot\frac{\gamma\rho_{\min}}{512\rho_{\max}^2},\quad \frac{\mathtt{a}}{\mathtt{d}}\cdot\frac{\gamma\rho_{\min}}{640}\right\}, \\ \alpha^2 \leq \min\left\{\frac{\mathtt{a}}{\mathtt{e}}\cdot\frac{\gamma\rho_{\min}n}{8192L^2(L_{\mathtt{s}}^2+L^2)},\quad \frac{\mathtt{a}}{\mathtt{g}}\cdot\frac{\gamma\rho_{\min}}{8192L^2(\bar{\rho}_{\max}^2+\rho_{\max}^2)}\right\}, \\ \gamma \leq \min\left\{\frac{\mathtt{a}}{\mathtt{e}}\cdot\frac{\rho_{\min}n}{8192\rho_{\max}^2(L_{\mathtt{s}}^2+L^2)},\quad \frac{\mathtt{a}}{\mathtt{g}}\cdot\frac{\rho_{\min}}{8192\rho_{\max}^2(\bar{\rho}_{\max}^2+\rho_{\max}^2)}\right\}, \\ a_x^2 \leq \frac{\mathtt{a}}{\mathtt{f}}\cdot\frac{\alpha^2\rho_{\min}}{5120\gamma\sigma_A^2}, \\ \mathtt{f}\cdot 64\alpha^2L^2G \leq \mathtt{a}\cdot\frac{\gamma\rho_{\min}}{256} \Leftarrow \begin{cases}\alpha^2 \leq \frac{\mathtt{a}}{\mathtt{f}}\cdot\frac{\gamma\rho_{\min}}{32768L^2(L_{\mathtt{s}}^2+L^2)}, \\ \gamma \leq \frac{\mathtt{a}}{\mathtt{f}}\cdot\frac{\rho_{\min}}{32768L^2(\bar{\rho}_{\max}^2+\rho_{\max}^2)},\end{cases} \\ \mathtt{f}\cdot 64\gamma^2\rho_{\max}^2G \leq \mathtt{a}\cdot\frac{\gamma\rho_{\min}}{256} \Leftarrow \begin{cases}\gamma \leq \frac{\mathtt{a}}{\mathtt{f}}\cdot\frac{\rho_{\min}}{32768\rho_{\max}^2(L_{\mathtt{s}}^2+L^2)}, \\ \gamma^3 \leq \frac{\mathtt{a}}{\mathtt{f}}\cdot\frac{\alpha^2\rho_{\min}}{32768\rho_{\max}^2(\bar{\rho}_{\max}^2+\rho_{\max}^2)},\end{cases} \\ a_\lambda^2 \leq \frac{\mathtt{a}}{\mathtt{g}}\cdot\frac{\gamma\rho_{\min}}{2560\sigma_A^2}, \end{cases} \tag{207}$$

$$\mathbb{C}_{\mathbf{v}} = -\mathtt{d}\cdot\frac{\eta}{2} + \mathtt{a}\cdot\frac{\eta^2}{(1-\gamma\rho_{\min})(1-\gamma\rho_{\min}/2)} + (\beta+\frac{\alpha}{\eta})(\mathtt{b}\rho_{\min}^{-1}+\mathtt{c}) \tag{208}$$

$$+\mathtt{e}\cdot\frac{32(L_{\mathtt{s}}^2+L^2)\eta^2}{n} + \mathtt{f}\cdot\left(\frac{16a_x^2\eta^2\gamma^2\sigma_A^2}{\alpha^2}+64\eta^2G\right) + \mathtt{g}\cdot 16\eta^2 \tag{209}$$

$$\leq -\mathtt{d}\cdot\frac{\eta}{4} \quad \text{when} \begin{cases} \eta \leq \frac{\mathtt{d}}{\mathtt{a}}\cdot\frac{(1-\gamma\rho_{\min})(1-\gamma\rho_{\min}/2)}{28}, \\ \beta+\frac{\alpha}{\eta} \leq \min\left\{\frac{\mathtt{d}}{\mathtt{b}}\cdot\frac{\eta\rho_{\min}}{28},\quad \frac{\mathtt{d}}{\mathtt{c}}\cdot\frac{\eta}{28}\right\}, \\ \eta \leq \min\left\{\frac{\mathtt{d}}{\mathtt{e}}\cdot\frac{n}{896(L_{\mathtt{s}}^2+L^2)},\quad \frac{\mathtt{d}}{\mathtt{g}}\cdot(1/448)\right\}, \\ a_x^2\eta\gamma^2 \leq \frac{\mathtt{d}}{\mathtt{f}}\cdot\frac{\alpha^2}{448\sigma_A^2}, \\ \mathtt{f}\cdot 64\eta^2G \leq \mathtt{d}\cdot\frac{\eta}{28} \Leftarrow \begin{cases}\eta \leq \frac{\mathtt{d}}{\mathtt{f}}\cdot(3584(L_{\mathtt{s}}^2+L^2))^{-1}, \\ \eta\gamma^2 \leq \frac{\mathtt{d}}{\mathtt{f}}\cdot\frac{\alpha^2}{3584(\bar{\rho}_{\max}^2+\rho_{\max}^2)},\end{cases} \end{cases} \tag{210}$$

$$\mathbf{C}_{xv} = \left(-\mathtt{a}\cdot\frac{2\eta(1-\gamma\rho_{\min}/4)}{1-\gamma\rho_{\min}} + \mathtt{b}\cdot 2\beta\right)\mathbf{K} \tag{211}$$

$$+\left(\mathtt{a}\cdot\frac{2\eta\gamma(1-\gamma\rho_{\min}/4)}{1-\gamma\rho_{\min}} + \mathtt{c}\cdot 2\beta - \mathtt{d}\cdot(\gamma+\eta\beta)\right)\mathbf{A}^\top\mathbf{R}\mathbf{A} \tag{212}$$

$$= \mathbf{0} \quad \text{when} \begin{cases} \mathtt{b} = \frac{\eta(1-\gamma\rho_{\min}/4)}{\beta(1-\gamma\rho_{\min})}\mathtt{a}, \\ \mathtt{c} = \frac{\mathtt{d}(\gamma+\eta\beta)-\mathtt{a}\cdot 2\eta\gamma(1-\gamma\rho_{\min}/4)(1-\gamma\rho_{\min})^{-1}}{2\beta}, \end{cases} \tag{213}$$

$$\mathbb{C}_{\Delta\overline{\mathbf{m}}_x} = -\mathtt{e}\cdot a_x + \alpha \tag{214}$$

$$\leq -\mathtt{e}\cdot\frac{a_x}{2} \quad \text{when } \alpha \leq \mathtt{e}\cdot\frac{a_x}{2}, \tag{215}$$

$$\mathbb{C}_{\Delta\mathbf{m}_x} = -\mathtt{f}\cdot\frac{a_x}{4} + \mathtt{a}\cdot\frac{2\alpha^2}{\gamma\rho_{\min}} + \mathtt{d}\cdot((\frac{3}{\eta}+\frac{1}{2})\alpha^2 + 2\alpha^2\beta + \frac{2\alpha^4}{\eta}) \tag{216}$$

$$+\mathtt{e}\cdot\frac{8(L_{\mathtt{s}}^2+L^2)\alpha^2}{n} + \mathtt{g}\cdot 8\alpha^2(1+\rho_{\max}^2+\bar{\rho}_{\max}^2) \tag{217}$$

$$\leq -\mathtt{f}\cdot\frac{a_x}{8} \quad \text{when} \begin{cases} \alpha^2 \leq \min\left\{\frac{\mathtt{f}}{\mathtt{a}}\cdot\frac{a_x\gamma\rho_{\min}}{64},\quad \frac{\mathtt{f}}{\mathtt{d}}\cdot\frac{a_x\eta}{384},\quad \frac{\mathtt{f}}{\mathtt{d}}\cdot\frac{a_x}{64}\right\}, \\ \alpha^2\beta \leq \frac{\mathtt{f}}{\mathtt{d}}\cdot\frac{a_x}{256}, \\ \alpha^4 \leq \frac{\mathtt{f}}{\mathtt{d}}\cdot\frac{a_x\eta}{256}, \\ \alpha^2 \leq \min\left\{\frac{\mathtt{f}}{\mathtt{e}}\cdot\frac{a_x n}{256(L_s^2+L^2)},\quad \frac{\mathtt{f}}{\mathtt{g}}\cdot\frac{a_x}{256(1+\rho_{\max}^2+\bar{\rho}_{\max}^2)}\right\}, \end{cases} \tag{218}$$

$$\mathbb{C}_{\Delta\mathbf{m}_\lambda} = -\mathtt{g}\cdot a_\lambda + 4\beta(\mathtt{b}\rho_{\min}^{-1}+\mathtt{c}) + \mathtt{d}\cdot\frac{\beta}{2} \tag{219}$$

$$\leq -\mathtt{g}\cdot\frac{a_\lambda}{2} \quad \text{when } \beta \leq \min\left\{\frac{\mathtt{g}}{\mathtt{b}}\cdot\frac{a_\lambda\rho_{\min}}{24},\quad \frac{\mathtt{g}}{\mathtt{c}}\cdot\frac{a_\lambda}{24},\quad \frac{\mathtt{g}}{\mathtt{d}}\cdot\frac{a_\lambda}{3}\right\}, \tag{220}$$

$$\mathbb{C}_{\overline{\mathbf{m}}_x} = -\frac{\alpha}{4} + \frac{4n\alpha^3L^2}{\eta}(\mathtt{b}\rho_{\min}^{-1}+\mathtt{c}) + \mathtt{d}\cdot\frac{\alpha^2L^2n}{2\eta} \tag{221}$$

$$\leq -\frac{\alpha}{8} \quad \text{when} \begin{cases} \alpha^2 \leq \min\left\{\mathtt{b}^{-1}\cdot\frac{\eta\rho_{\min}}{96nL^2},\quad \mathtt{c}^{-1}\cdot\frac{\eta}{96nL^2}\right\}, \\ \alpha \leq \mathtt{d}^{-1}\frac{\eta}{12nL^2}. \end{cases} \tag{222}$$

Remark E.9 (**On the hidden constants**). [3] Let

$$R := \max\{1, L^2\}, \qquad S := \max\{1, L_s^2 + L^2\}, \qquad H := \min\{R^{-2}, (SR)^{-1/2}\}.$$

Thus, $S \geq R \geq 1$. There exists a sufficiently small constant $\varepsilon \in (0,1)$, depending only on the network constants and absolute numerical constants, but independent of $n, L, L_s, T, \bar{\sigma}$, such that the following construction satisfies the conditions in (192)–(222). Here, independence of n is understood with the network constants held fixed.

For any

$$0 < h \leq \varepsilon^{20}H,$$

choose

$$\mathtt{a} = \frac{1}{n}, \qquad \alpha = h, \qquad \eta = \frac{\varepsilon^2}{R}, \qquad \gamma = \varepsilon^{-4}Rh,$$
$$\beta = \varepsilon^{-9}R^3h^2, \qquad a_x = \varepsilon^{-13}SRh^2, \qquad a_\lambda = \varepsilon^{-4}R^2h,$$
$$\mathtt{d} = \frac{\varepsilon^6}{nR^2h}, \qquad \mathtt{e} = \frac{\varepsilon^{12}}{SRh}, \qquad \mathtt{f} = \frac{\varepsilon^{13}}{nSR^2h}, \qquad \mathtt{g} = \frac{\varepsilon^5}{nR^3h}.$$

Define $\mathtt{b}, \mathtt{c}$ by the exact identities

$$\mathtt{b} = \mathtt{a}\frac{\eta(1-\gamma\rho_{\min}/4)}{\beta(1-\gamma\rho_{\min})},$$
$$\mathtt{c} = \frac{\mathtt{d}}{2}\left(\frac{\gamma}{\beta}+\eta\right) - \mathtt{b}\gamma.$$

In particular,

$$\beta = \varepsilon\frac{\mathtt{a}\gamma}{\mathtt{d}}.$$

We first verify positivity and the cancellation of the mixed term. Since

$$\gamma\rho_{\min} = \varepsilon^{-4}Rh\rho_{\min} \leq \varepsilon^{16}\rho_{\min},$$

we have $\gamma\rho_{\min} \leq 1/2$ for sufficiently small ε. Consequently, with

$$\kappa := \frac{1-\gamma\rho_{\min}/4}{1-\gamma\rho_{\min}},$$

we have $1 \leq \kappa \leq 2$ and

$$\mathtt{b} = \frac{\kappa\varepsilon^{11}}{nR^4h^2},$$

[3] The following discussions/constructions on the stepsize conditions have been conducted with the aid of GPT-6.

$$\mathtt{c} = \frac{\varepsilon^{11}}{2nR^4h^2}\left(1 + \varepsilon^{-3}Rh - 2\kappa\varepsilon^{-4}Rh\right).$$

Using $Rh \le \varepsilon^{20}$ and decreasing ε if necessary, we obtain

$$\frac{\varepsilon^{11}}{4nR^4h^2} \le \mathtt{c} \le \frac{\varepsilon^{11}}{nR^4h^2}.$$

Thus $\mathtt{b}, \mathtt{c} > 0$, and

$$\mathtt{b}, \mathtt{c} = \Theta(n^{-1}R^{-4}h^{-2}),$$

with constants independent of n, L, L_s, h. Moreover,

$$\frac{\mathtt{d}^2}{4\mathtt{a}} = \frac{\varepsilon^{12}}{4nR^4h^2} \le \mathtt{c}.$$

Hence the quadratic part of the potential is nonnegative:

$$\mathtt{a}\|x\|_K^2 + \|v\|^2_{\mathtt{b}Q+\mathtt{c}K} + \mathtt{d}\langle x, v\rangle_K \ge 0,$$

and therefore $F_t \ge f_\star$.

The definitions of $\mathtt{b}, \mathtt{c}$ also yield

$$2\beta\mathtt{b} = 2\mathtt{a}\eta\kappa,$$
$$2\beta\mathtt{c} = \mathtt{d}(\gamma + \eta\beta) - 2\mathtt{a}\eta\gamma\kappa.$$

These identities give the exact cancellation $C_{xv} = 0$ in (211)–(213).

We next verify the inequality restrictions. The bound on h implies

$$h \le \varepsilon^{20}, \qquad Rh \le \varepsilon^{20}, \qquad R^2h \le \varepsilon^{20},$$
$$R^2h^2 \le \varepsilon^{40}, \qquad SRh^2 \le \varepsilon^{40}, \qquad SR^3h^4 \le \varepsilon^{80}.$$

For each scalar restriction, move all positive factors to obtain an equivalent inequality with a positive right-hand side, and consider the ratio of its left-hand side to its right-hand side. For the square-root restrictions in (192)–(193), use the equivalent squared inequalities. Splitting each minimum into its individual restrictions, direct substitution gives the following bounds:

Restrictions	Upper bound on each ratio
(192)–(194)	$C\varepsilon^5$
(200), explicit sufficient restrictions	$C\varepsilon^5$
(207), explicit sufficient restrictions	$C\varepsilon$
(210), explicit sufficient restrictions	$C\varepsilon$
(215)	2ε
(218)	$C\varepsilon^4$
(220)	$C\varepsilon$
(222)	$C\varepsilon^4$

Here and below, C depends only on network and numerical constants and may change between displays.

For example, the identities

$$\mathtt{e}a_x = \varepsilon^{-1}h, \qquad \mathtt{f}a_x = \frac{h}{nR}$$

give

$$\frac{2\alpha}{\mathtt{e}a_x} = 2\varepsilon, \qquad \frac{64\mathtt{a}\alpha^2}{\mathtt{f}a_x\gamma\rho_{\min}} = \frac{64\varepsilon^4}{\rho_{\min}}, \qquad \frac{384\mathtt{d}\alpha^2}{\mathtt{f}a_x\eta} = 384\varepsilon^4.$$

The restrictions involving both $\mathtt{b}, \mathtt{c}$ and $\beta + \alpha/\eta$ satisfy

$$\frac{\mathtt{b}(\beta + \alpha/\eta)}{\mathtt{d}\eta} + \frac{\mathtt{c}(\beta + \alpha/\eta)}{\mathtt{d}\eta} \le C\left(\varepsilon + \varepsilon^{-6}R^2h\right) \le C\left(\varepsilon + \varepsilon^{14}\right).$$

Furthermore,

$$\frac{\mathtt{f}a_x^2\gamma}{\mathtt{a}\alpha^2} = \varepsilon^{-17}SRh^2 \le \varepsilon^{23}, \qquad \frac{\mathtt{f}a_x^2\eta\gamma^2}{\mathtt{d}\alpha^2} = \varepsilon^{-25}SR^3h^4 \le \varepsilon^{55}.$$

Finally,

$$\frac{\mathtt{d}nL^2\alpha}{\eta} = \varepsilon^4 \frac{L^2}{R} \le \varepsilon^4.$$

To check the intermediate conditions involving G, write

$$S_0 := L_s^2 + L^2 \le S, \qquad \mathcal{R}_A := \rho_{\max}^2 + \bar{\rho}_{\max}^2.$$

Then

$$G = S_0 + \varepsilon^{-8}\mathcal{R}_A R^2,$$

and consequently

$$\begin{aligned}
\mathtt{f}\alpha n G &\le \frac{\varepsilon^{13}}{R^2} + \frac{\varepsilon^5 \mathcal{R}_A}{S}, \\
\frac{\mathtt{f}\alpha^2 L^2 G}{\mathtt{a}\gamma} &\le \frac{\varepsilon^{17}}{R^2} + \frac{\varepsilon^9 \mathcal{R}_A}{S}, \\
\frac{\mathtt{f}\gamma G}{\mathtt{a}} &\le \frac{\varepsilon^9}{R} + \varepsilon \mathcal{R}_A \frac{R}{S}, \\
\frac{\mathtt{f}\eta G}{\mathtt{d}} &\le \frac{\varepsilon^9}{R} + \varepsilon \mathcal{R}_A \frac{R}{S}.
\end{aligned}$$

Since $S \ge R \ge 1$, these quantities tend to zero uniformly in n, L, L_s, h as ε decreases. Thus the intermediate conditions involving G in (200), (207), and (210) also hold.

All restrictions can therefore be enforced simultaneously by a single sufficiently small ε independent of $n, L, L_s, T, \bar{\sigma}$. The resulting coefficient bounds are

$$\begin{aligned}
&C_{\nabla F} \le -\frac{\alpha}{4}, \qquad C_x \le -\frac{\mathtt{a}\gamma\rho_{\min}}{16}, \qquad C_v \le -\frac{\mathtt{d}\eta}{4}, \\
&C_{\Delta \bar{m}_x} \le -\frac{\mathtt{e}a_x}{2}, \qquad C_{\Delta m_x} \le -\frac{\mathtt{f}a_x}{8}, \qquad C_{\Delta m_\lambda} \le -\frac{\mathtt{g}a_\lambda}{2}, \qquad C_{\bar{m}_x} \le -\frac{\alpha}{8}.
\end{aligned}$$

The noise coefficient has the explicit form

$$C_\sigma = 2\mathtt{e}a_x^2 + 4n\mathtt{f}a_x^2 = 2\varepsilon^{-14} SRh^3 + 4\varepsilon^{-13} Sh^3.$$

Therefore, if the initialization satisfies $F_0 - f_\star \le \Delta$ for a positive upper bound Δ independent of n, L, L_s, T, summing the potential decrease gives

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(\bar{x}^t)\|^2 + \frac{\varepsilon^{-4} R\rho_{\min}}{4n}\|x^t\|_K^2\right] \\
&\qquad \le \frac{4\Delta}{Th} + \left(8\varepsilon^{-14} SR + 16\varepsilon^{-13} S\right) h^2 \bar{\sigma}^2.
\end{aligned}$$

For $\bar{\sigma} > 0$, choose

$$h = \varepsilon^{20} \min\left\{H, \left(\frac{\Delta}{SR\bar{\sigma}^2 T}\right)^{1/3}\right\}.$$

For $\bar{\sigma} = 0$, choose $h = \varepsilon^{20} H$.

These choices yield

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[\|\nabla F(\bar{x}^t)\|^2 + \frac{\varepsilon^{-4} R\rho_{\min}}{4n}\|x^t\|_K^2\right] \\
&\qquad = O\left(\frac{\Delta H^{-1}}{T} + \frac{\Delta^{2/3}(SR)^{1/3}\bar{\sigma}^{2/3}}{T^{2/3}}\right).
\end{aligned}$$

For $L \ge 1$, the leading stochastic smoothness factor is

$$(SR)^{1/3} = \left[L^2(L_s^2 + L^2)\right]^{1/3}.$$

The uniform initialization bound on Δ is an additional requirement; it does not follow from the step size choices alone.

E.9 Auxiliary Lemma

Lemma E.10. *Under Assumption 3.1, 3.2,*

$$\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2 \le 2\alpha^2\|\mathbf{m}_x^t-\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\|^2+8\alpha^2 n\|\nabla F(\bar{\mathbf{x}}^t)\|^2 \tag{223}$$

$$+8\eta^2\|\mathbf{v}^t\|_{\mathbf{K}}^2+(8\alpha^2L^2+8\gamma^2\rho_{\max}^2)\|\mathbf{x}^t\|_{\mathbf{K}}^2. \tag{224}$$

Proof of Lemma E.10.

$$\|\mathbf{x}^{t+1}-\mathbf{x}^t\|^2=\alpha^2\|\mathbf{m}_x^t\|^2\le 2\alpha^2\|\mathbf{m}_x^t-\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\|^2+2\alpha^2\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\|^2. \tag{225}$$

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

$$\|\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^t,\boldsymbol{\lambda}^t)\|^2=\left\|\nabla\mathbf{f}(\mathbf{x}^t)+\frac{\eta}{\alpha}\mathbf{A}^\top\boldsymbol{\lambda}^t+\frac{\gamma}{\alpha}\mathbf{A}^\top\mathbf{R}\mathbf{A}\mathbf{x}^t\right\|^2 \tag{226}$$

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

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

$$+\frac{4\gamma^2\rho_{\max}^2}{\alpha^2}\|\mathbf{x}^t\|_{\mathbf{K}}^2 \tag{227}$$

$$\le 4n\|\nabla F(\bar{\mathbf{x}}^t)\|^2+\frac{4\eta^2}{\alpha^2}\|\mathbf{v}^t\|_{\mathbf{K}}^2+(4L^2+\frac{4\gamma^2\rho_{\max}^2}{\alpha^2})\|\mathbf{x}^t\|_{\mathbf{K}}^2. \tag{228}$$

Combining the above inequalities completes the proof. □

F Ablation Study

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

F.1 Data Heterogeneity

To study the effect of heterogeneity of data distribution across agents, we experiment with two types of data splitting for MNIST: (i) the dataset is split into $n=10$ disjoint sets according to the class 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 5 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 5 by Table 5 in Appendix G.

F.2 Random Graph Sparsity

Next, we study the effects of random graph sparsity by the experiments shown in Figure 6. As the random graph sparsity decreases, the random graph variance σ_A^2 in Lemma 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 6, we tuned FSPDA-SA to use the step sizes $\alpha=10^{-4},\eta=10^{-6},\gamma=0.5,\beta=1$ for all configurations.

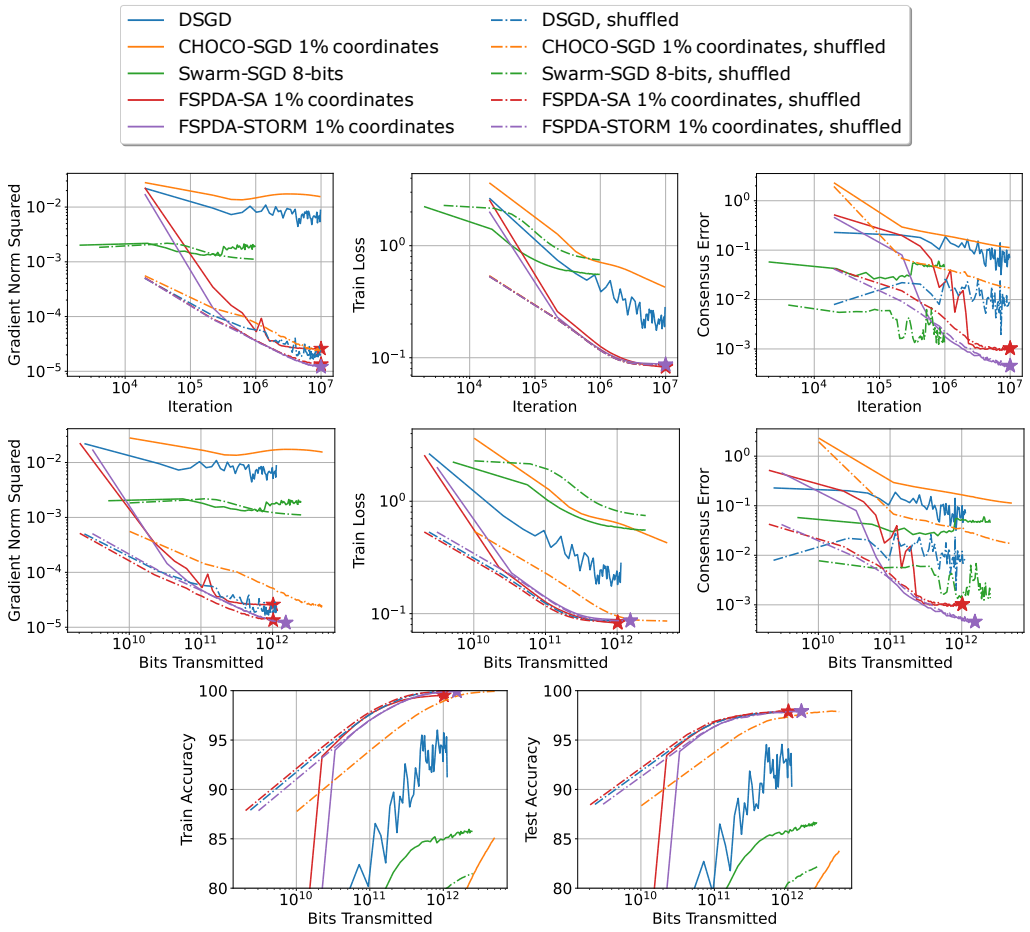

Figure 5: Feed-forward neural network classification training on MNIST with two levels of data heterogeneity.

F.3 Graph Topology

We then study the effects of network topology $\mathcal{G}$ in `FSPDA-SA` by drawing one-edge random graphs from a complete graph, an ER graph with probability $p = 0.5$ and a ring graph in Figure 7. Note that the communication cost per iteration is the same across different topologies due to the use of one-edge random graph. The result indicates the transient effect of topology where it only slow down the convergence of consensus error while converging to the same level of stationarity. For Figure 7, we tuned `FSPDA-SA` to use the step sizes $\alpha = 10^{-4}, \eta = 10^{-6}, \gamma = 0.5, \beta = 1$ for all configurations.

Furthermore, we provide an additional set of Imagenet experiments on a static graph $\mathcal{G}$ of complete graph in Figure 8. We can see the performance of `FSPDA-SA` is insensitive to the graph topology $\mathcal{G}$.

F.4 Deterministic Gradient

For the case when the gradient estimate is exact, i.e., $\bar{\sigma}^2 = 0$, we compare the performance of `FSPDA-SA` against deterministic gradient algorithm `DIGing` [Nedic et al., 2017] in Figure 9. Despite `FSPDA-SA` only performs model parameter gossip while `DIGing` performs an extra step of gradient tracker gossip, `FSPDA-SA` shows comparative performance and reduced the communication cost by half.

Also, in the absence of stochastic gradient noise, notice that the effect of parameter sparsification immediately transfer to a slower optimization convergence. This is in line with our theorem by observing the convergence rate in Theorem 3.5, where σ_A^4 remains dominant in the convergence bound. We list the hyperparameters used in Figure 9 by Table 7 in Appendix G.

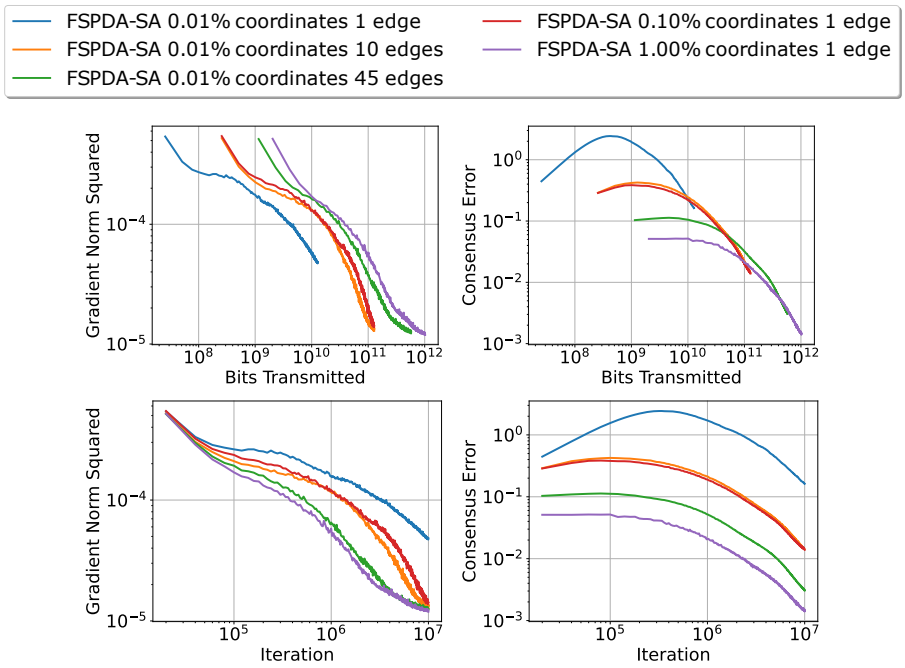

Figure 6: Feed-forward neural network classification training on shuffled MNIST. Random graph of k edges in expectation ($k \in \{1, 10, 45\}$) is drawn from a complete topology per iteration.

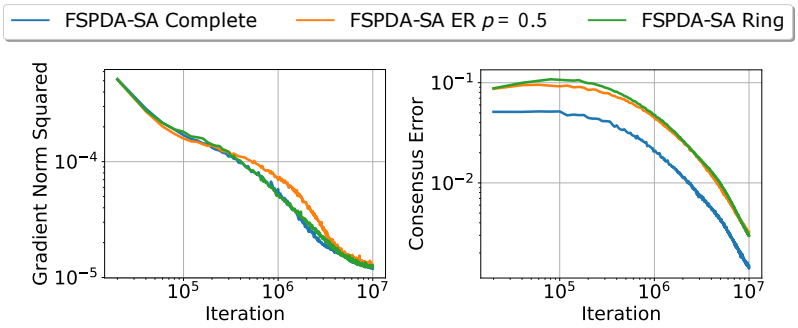

Figure 7: Feed-forward neural network classification training on shuffled MNIST with different graph topology $\mathcal{G}$. Only one edge is activated per iteration, exchanging 1% coordinates of model parameters.

F.5 Dual Momentum

To investigate the benefits of dual momentum in `FSPDA-STORM`, we construct a case where both the primal and dual stochastic gradients carry large variance error. In Figure 10, the local objective function gradient is estimated by batch size 1, and one-edge random graphs with 0.01% coordinate sparsification is adopted for communication. We can observe that applying dual momentum ($a_\lambda = 0.01$) outperforms not applying dual momentum ($a_\lambda = 1$) in terms of consensus error convergence. We list the hyperparameters used in Figure 10 by Table 8 in Appendix G.

F.6 Graph Size

Now we investigate the effect of graph size on `FSPDA-SA`. With an increased number of workers, the training dataset is split into n chunks correspondingly, while keeping the local batch size as 256 across different sizes of graphs to illustrate the effect of parallelization. We also maintain the same amount of network transmission per communication link. From the result of Figure 11, we can see a

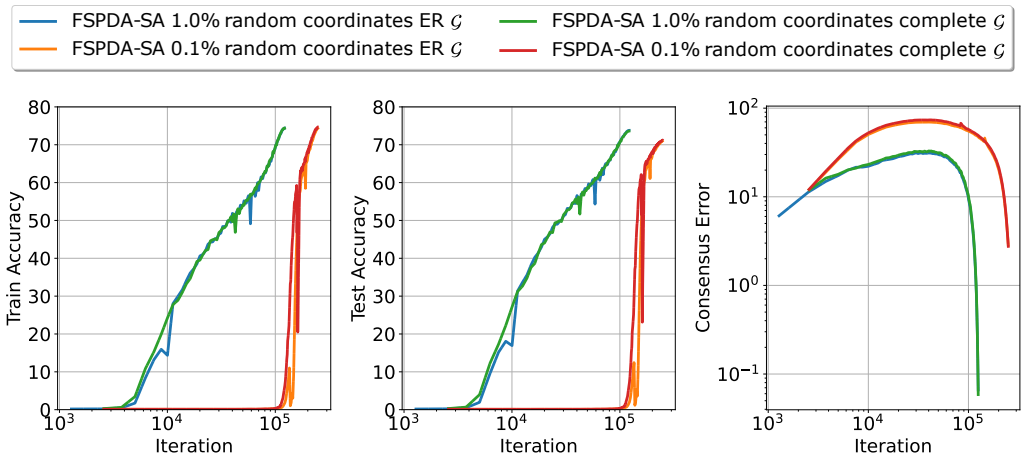

Figure 8: Resnet-50 classification training on Imagenet with different graph topology $\mathcal{G}$. Only one edge is activated per iteration.

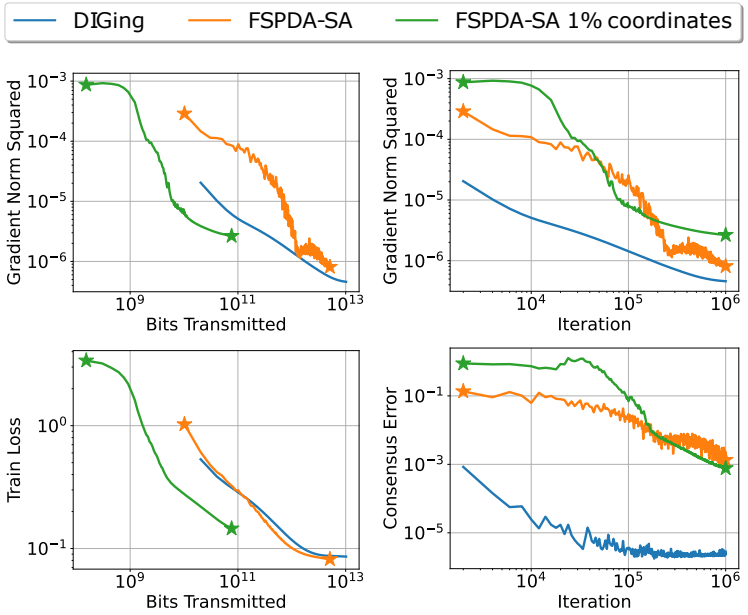

Figure 9: Feed-forward neural network classification training on class separated MNIST using exact local gradient, i.e., $\sigma_i = 0 \; \forall i$. One-edge random graph is drawn from a complete topology per iteration.

more stable convergence in larger graphs due to the reduced variance in local gradient estimation. The adopted hyperparameter values are $\alpha = 10^{-4}$, $\eta = 10^{-4}$, $\gamma = 0.5$ for all runs.

G Experiment Hyperparameters

This section lists the set of hyperparameters used in the experiments.

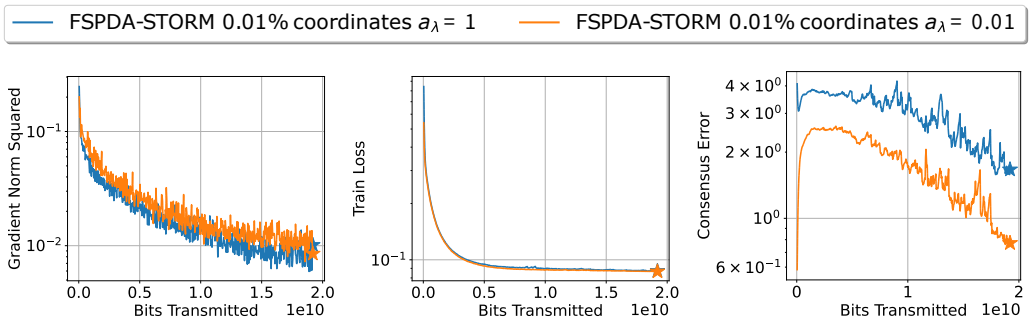

Figure 10: Feed-forward neural network classification training on shuffled MNIST using 10^7 iterations. One-edge random graph is drawn from a complete topology per iteration.

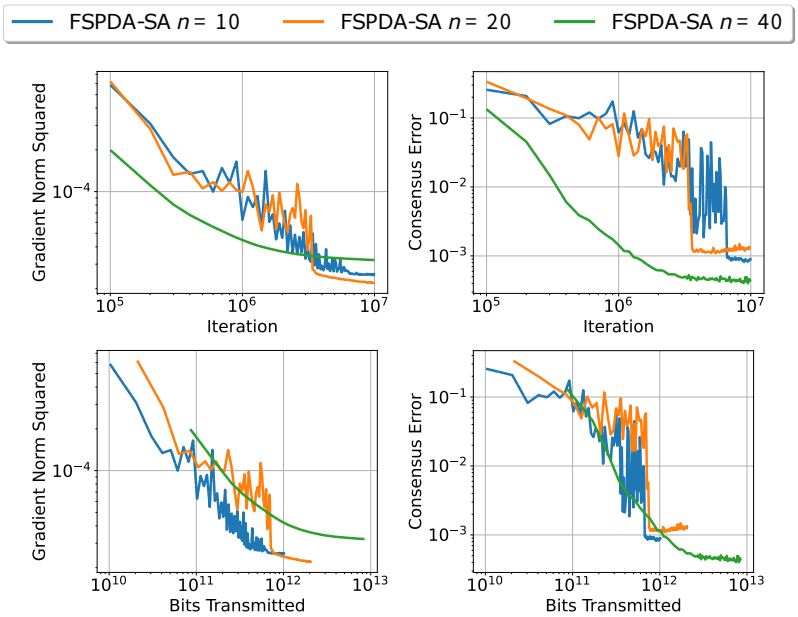

Figure 11: Feed-forward neural network classification training on class separated MNIST with various graph size n. The underlying static graph $\mathcal{G}$ is an Erdos-Renyi graph with connectivity of $p = 0.5$, and agents exchange 1% coordinates of model parameters.

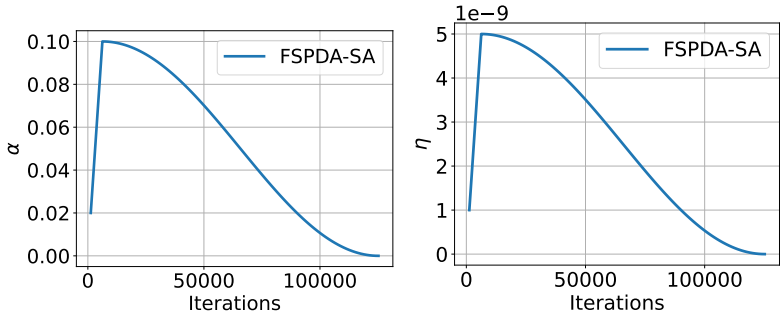

Figure 12: Illustration of step size cosine scheduling used in Figure 2, 3, 8 for `FSPDA-SA` with 1% sparse coordinates.

FSPDA	α	η	γ	β	a_x	a_λ
-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}
	Opt. S.S.	**Local Steps**	η_s	-	-	-
K-GT	10^{-4}	150	1	-	-	-
	Opt. S.S.	**Local Steps**	-	-	-	-
LED	10^{-4}	75	-	-	-	-
	Opt. S.S.	τ	p	-	-	-
Decen-Scaffnew	10^{-4}	130	0.013	-	-	-
	Opt. S.S.	**Edge Prob.**	-	-	-	-
DSGD	10^{-4}	0.013	-	-	-	-
	Opt. S.S.	**Quant. Side Length**	-	-	-	-
Swarm-SGD	5×10^{-5}	10^{-4}	-	-	-	-
	Opt. S.S.	**Consensus S.S.**	**Active Prob.**	-	-	-
CHOCO-SGD	10^{-4}	10^{-3}	0.03	-	-	-

Table 2: Hyperparameter values used in Figure 1.

FSPDA-SA	$\max_t \alpha_t$	$\max_t \eta_t$	γ	β
1% coordinates	0.1	10^{-9}	0.5	1
0.5% coordinates	0.1	10^{-9}	0.5	1
0.1% coordinates	0.05	5×10^{-10}	0.5	1
CHOCO-SGD	**Max. Opt. S.S.**	**Consensus S.S.**	**Active Prob.**	-
1% coordinates	0.1	0.005	0.1	-
Swarm-SGD	**Max. Opt. S.S.**	**Quant. Side Length**	-	-
8-bits quantization	0.005	3×10^{-5}	-	-
DSGD	**Max. Opt. S.S.**	-	-	-
one-edge	0.1	-	-	-

Table 3: Hyperparameter values used in Figure 2.

FSPDA-SA	$\max_t \alpha_t$	$\max_t \eta_t$	γ	β
1% coordinates	0.1	10^{-9}	0.5	1
0.5% coordinates	0.1	10^{-9}	0.5	1
0.1% coordinates	0.05	5×10^{-10}	0.5	1
K-GT	**Max. Opt. S.S.**	η_s	-	-
1000 local steps	0.1	1	-	-
2000 local steps	0.1	1	-	-
LED	**Max. Opt. S.S.**	-	-	-
500 local steps	0.1	-	-	-
1000 local steps	0.1	-	-	-
Decen-Scaffnew	**Max. Opt. S.S.**	**Consensus S.S.**	-	-
$p = 0.002$	0.1	0.1	-	-
$p = 0.001$	0.1	0.1	-	-

Table 4: Hyperparameter values used in Figure 3.

DSGD	**Opt. S.S.**	**Edge Prob.**	-	-	-	-
(i) hetero.	10^{-4}	5×10^{-4}	-	-	-	-
(ii) homo.	10^{-4}	5×10^{-4}	-	-	-	-
CHOCO-SGD	**Opt. S.S.**	**Consensus S.S.**	**Active Prob.**	-	-	-
(i) hetero.	10^{-4}	10^{-3}	0.1	-	-	-
(ii) homo.	10^{-4}	10^{-3}	0.1	-	-	-
Swarm-SGD	**Opt. S.S.**	**Quant. Side Length**	-	-	-	-
(i) hetero.	5×10^{-5}	10^{-4}	-	-	-	-
(ii) homo.	5×10^{-5}	10^{-4}	-	-	-	-
FSPDA-SA	α	η	γ	β	-	-
(i) hetero.	10^{-4}	10^{-4}	0.5	1	-	-
(ii) homo.	10^{-4}	10^{-5}	0.5	1	-	-
FSPDA-STORM	α	η	γ	β	a_x	a_λ
(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 5: Hyperparameter values used in Figure 5.

FSPDA-SA	$\max_t \alpha_t$	$\max_t \eta_t$	γ	β
1% coordinates, ER $\mathcal{G}$	0.1	10^{-9}	0.5	1
0.1% coordinates, ER $\mathcal{G}$	0.05	5×10^{-10}	0.5	1
1% coordinates, complete $\mathcal{G}$	0.1	10^{-9}	0.5	1
0.1% coordinates, complete $\mathcal{G}$	0.05	5×10^{-10}	0.5	1

Table 6: Hyperparameter values used in Figure 8.

FSPDA-SA	α	η	γ	β
no sparse	10^{-3}	5×10^{-6}	0.5	1
1% coordinates	10^{-4}	5×10^{-4}	0.5	1
DIGing	**Opt. S.S.**	-	-	-
no sparse	10^{-3}	-	-	-

Table 7: Hyperparameter values used in Figure 9.

	α	η	γ	β	a_x	a_λ
FSPDA-STORM	10^{-3}	5×10^{-6}	0.5	1	10^{-3}	1
FSPDA-STORM	10^{-3}	5×10^{-6}	0.5	1	10^{-3}	10^{-2}

Table 8: Hyperparameter values used in Figure 10.

	Time	Peak Memory	Machine
Figure 1	27 hours	339 MB	Intel Xeon Gold 6148 CPU
Figure 5	197 hours	372 MB	Intel Xeon Gold 6148 CPU
Figure 6	201 hours	360 MB	Intel Xeon Gold 6148 CPU
Figure 7	72 hours	1146 MB	Intel Xeon Gold 6148 CPU
Figure 8	38 hours	(GPU) 112 GB (CPU) 6650 MB	8× NVIDIA V100 GPU
Figure 9	354 hours	1066 MB	Intel Xeon Gold 6148 CPU
Figure 10	46 hours	350 MB	Intel Xeon Gold 6148 CPU

Table 9: Statistics of experiment compute time (per algorithm run) and compute instance.

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: The main claims in the abstract and introduction match the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] 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 [N/A] 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: We have discussed the limitation of the current analysis for `FSPDA-STORM`.

   Guidelines:

   - The answer [N/A] 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: All the proofs and assumptions are provided in the paper.

Guidelines:

- The answer [N/A] 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: The experiment results are reproducible.

   Guidelines:

   - The answer [N/A] 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: As explained in the above.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.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://neurips.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, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See details in the paper.

   Guidelines:

   - The answer [N/A] 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: [Yes]

   Justification: See Figure 1.

   Guidelines:

   - The answer [N/A] 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: As written in the paper.

   Guidelines:

   - The answer [N/A] 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: [N/A]

   Guidelines:

   - The answer [N/A] 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: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] 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., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] 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` 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: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] 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: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] 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**

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

Answer: [Yes]

Justification: See the paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.